Social-Media-Daten:

Chancen und Herausforderungen der Nutzung von Social-Media-Daten im Kontext wissenschaftlicher Forschung

Dissertation

zur Erlangung des akademischen Grades
doctor rerum politicarum
(Dr. rer. pol.)

vorgelegt dem
Rat der Wirtschaftswissenschaftlichen Fakultät
der Friedrich-Schiller-Universität Jena

am 9. Juli 2022

von: Diplom Kaufmann Sven Gehrke
geboren am: 1.4.1972 in Potsdam

Danksagung

Neben den normalen Herausforderungen einer Dissertation, kam mit der COVID-19 Pandemie noch zusätzliche Widrigkeiten hinzu. Ich möchte mich bei allen bedanken, die mir über die Jahre Kraft und Unterstützung gegeben haben.

Mein besonderer Dank gilt zunächst meinem Doktorvater Prof. Dr. Johannes Ruhland, für seine Anregungen, positive Kritik und zahlreichen Diskussionen.

Ich möchte mich auch ganz herzlich bei meinen lieben Kollegen Birgit, Sandra, Marek und Thomas für ihre permanente Unterstützung und die vielen Diskussionen die wir während der letzten 6 Jahre hatten bedanken.

Nicht vergessen möchte ich meine Familie - insbesondere Anja - und meinen Freunden die mich ertragen haben, mich unterstützten und wenn es notwendig wurde wieder motivierten.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

ABM	Agent Based Modeling (Agenten basierte Modellierung)
AIC	Akaike Information Criterion (dt. Akaike Informationskriterium)
BIC	Baysian Information Criterion (dt. Bayessches Informationskriterium)
COBIT	Control Objectives for Information and Related Technologies (dt. Kontrollziele für Informations- und verwandte Technologien)
CRISP DM	Cross Industry Standard Process for Data Mining
mDAU	monetisable Daily Active Users (dt. monetarisierbare tägliche Nutzer)
NLP	Natural Language Processing (dt. Computerlinguistik)
PCA	Principal Component Analysis (dt. Hauptkomponentenanalyse)
POS tagging	„part of speech“ tagging (dt. Wortartbestimmung)
p-value	Signifikanzwert (eng. "probability value")
R	Bestimmtheitsmaß der Regression
RSS	Residual Sum of Squares (dt. Quadratsumme der Residuen)
SSE	Sum of Squared Errors (dt. Summe der quadratischen Abweichung)
SQL	Structured Querry Language (dt. strukturierte Abfragesprache)
SVM	Support Vektor Maschine (dt. Stützvektormaschine (nicht gebräuchlich))

1 Motivation

Daten bilden die Grundlage jeglicher akademischer Forschung und wirtschaftlicher Entscheidung. Dabei ist die Verfügbarkeit, der Umfang und die Qualität der genutzten Daten entscheidend für die Aussagekraft, Verlässlichkeit und Reproduzierbarkeit wissenschaftlicher Erkenntnisse. Daten bilden den Ausgangspunkt vieler Modelle, unter anderem in dem vielfältig referenzierten Prozess “Knowledge Discovery in Databases” [1]. Daten lassen sich auf vielfältige Weise erheben. Während im naturwissenschaftlichen Umfeld vorrangig Daten durch Experimente erhoben werden, liegt der Schwerpunkt in der sozialwissenschaftlichen Forschung traditionell auf Beobachtungen und Interviews. Diese Erhebung ist mit erheblichem Zeit und Kostenaufwand verbunden und stellt eine ernsthafte Herausforderung für jeden Analysten dar. Die verfügbaren Daten sind im Laufe der Zeit immer vielfältiger geworden. Eine der wichtigsten Triebfedern für die Entwicklung des Internets, der eMail und von Datenbanken in den 1970er Jahren war es, Daten und Datenquellen einfach zu reproduzieren und allgemein verfügbar zu machen. Die Möglichkeiten des Internets zur einfachen weltweiten Kommunikation führten auch dazu, neue Datenquellen für die sozialwissenschaftliche Forschung und wirtschaftliche Nutzung zu erschließen. Soziale Medien sind gekennzeichnet durch von Nutzern generierten Inhalten z.B. Bewertungen, Kommentaren, Blogs, persönlichen Profilen in vielfältigen sozialen Kontexten und Bildern /Videos. Viele Daten werden eigenständig durch die einzelnen Teilnehmer freiwillig und öffentlich via den “sozialen Medien” zur Verfügung gestellt.

Die Nutzung dieser Daten ist vielfältig: im Verborgenen ablaufende Analyse der Nutzerdaten können zur Erstellung individualisierter Werbung und Empfehlungen genutzt werden, das Erkennen gesellschaftlicher Trends ermöglichen eine schnelle und effiziente Reaktion und Steuerung seitens der Wirtschaft, Wissenschaft oder gesellschaftlichen Organen. Aber auch der einzelne Nutzer wird in seiner individuellen Selbstverwirklichung durch die Möglichkeiten des Austauschs mit Gleichgesinnten oder durch die Teilhabe am gesellschaftlichen Diskurs unterstützt.

Der Prozess der Wissensgenerierung aus klassischen Datenquellen setzt nicht unerhebliche Aufwände voraus, und unterliegt einer aufwendigen Planung und Validierung. Während die Wissensgenerierung und der gesellschaftliche Austausch darüber seit Langen in der Literatur diskutiert und geübte Praxis für weite Teile der Gesellschaft ist, muss dies im Gegensatz dazu für die sozialen Medien und den damit verbundenen generierten Daten erst eruiert, adaptiert und etabliert werden. Die in der vorliegenden Arbeit vorgestellten Einzelfallstudien zeigen grundsätzliche Herausforderungen, Möglichkeiten und Limitationen auf und helfen den Prozess der Wissensgenerierung analog zu den klassischen Datenquellen zu etablieren und zu standardisieren,

Unterschiede zu verdeutlichen und somit das Vertrauen auf sozialen Medien basierende Analysen zu stärken.

In der vorliegenden Arbeit soll beantwortet werden, wie sich die neuen Datenquellen für die Wissensgenerierung eignen, aufzeigen welche Herangehensweisen dafür notwendig sind, wie sich Methoden zur Analyse großer heterogener und unvollständiger Daten seit dem Aufkommen der sozialen Medien entwickelt haben und welche spezifischen Herausforderungen sich dabei ergeben. Dabei wird sich im Rahmen dieser Arbeit neben der theoretischen, akademischen Betrachtungen auch um die praktische Verknüpfung bemüht.

1.1 Aufbau der Arbeit

Social Media Daten fallen in riesigen Mengen an und die Bearbeitung dieser Daten kann in den seltensten Fällen manuell erfolgen. Die automatisierte Analyse der gewonnen Daten ist daher üblich und wird als Data Mining bezeichnet. Data Mining ist ein sehr heterogenes Feld und umfasst Beiträge aus Mathematik, Statistik, Physik, Sprachwissenschaft und anderen verwandten Wissenschaftsgebieten.

Somit ist es notwendig, zuerst einen Überblick über die unterschiedlichen Bereiche des Data Mining zu erhalten. Einen Überblick über die Social Media Typen gibt Kapitel 2.1. Aus diesen Typen werden mit verschiedenen Algorithmen und Methoden Daten extrahiert und analysiert. Die Wirksamkeit dieser Algorithmen wird aus wissenschaftlicher Sicht primär an „time honoured“, gut verstandenen Datensätzen demonstriert. Dies ist aus Gründen der Vergleichbarkeit und Reproduzierbarkeit wissenschaftlich gut nachvollziehbar. Praktische Daten weisen zahlreiche Eigenschaften auf, die diese „wohl-konstruierten“ Eigenschaften nicht aufweisen. Die Kapitel 2.2 und 2.3 geben einen Überblick über die Eigenschaften die die Datenqualität im Allgemeinen und speziell im Kontext der Sozialen Medien beeinflussen. Kapitel 2.2 zeigt die grundlegenden Eigenschaften von Social Media Daten auf. Ein wesentlicher Aspekt und Ausgangspunkt sämtlicher Analysen ist die Bewertung der Qualität der zugrundeliegenden Daten. Die Datenqualität muss stets im Kontext des Analyseziels betrachtet werden - Daten welche im Rahmen einer Analyse den Qualitätsanforderungen genügen, können für weitere Analysen völlig ungeeignet sein. Das Kapitel 2.3 gibt einen Überblick über die zu beachtenden Aspekte der Datenqualität.

In Kapitel 2.4 wird ein Überblick über die klassischen Bereich des Data Mining gegeben. Im Rahmen dieser Arbeit kann dabei nur auf wesentliche Ansätze eingegangen werden, die für das Verständnis von Kapitel 2.6 notwendig sind. Es würde aber zu kurz greifen, nur einen Überblick über die reinen Methoden zu geben, da alle Methoden in Abhängigkeit von den zugrundeliegenden Metriken der Daten unterschiedliche Maße zur Darstellung der Ähnlichkeit oder Unähnlichkeit verwenden oder aus einem Pool das Angebrachteste auswählen müssen. Dabei kann die Wahl des Ähnlichkeitsmaßes entscheidend zum Erfolg der Analyse beitragen. Kapitel 2.5 gibt einen repräsentativen Überblick über die wichtigsten Maße, eine vollständige Erläuterung würde sowohl den Rahmen dieser Arbeit sprengen, als auch keinen signifikanten Beitrag zum Verständnis liefern.

Nachdem die theoretischen Grundlagen zur Analyse sozialer Medien Daten gelegt wurden, wird anhand verschiedener Untersuchungen dargestellt, wie einzelne Methoden, basierend auf dem klassischen Kreislauf der Wissensgenerierung, zur Wissensgenerierung jenseits der reinen

Grundlagenforschung beitragen können. Aufgrund des Umfangs der möglichen Social Media Typen und Methoden kann dies im Rahmen dieser Arbeit nur an Einzelfällen demonstriert werden.

Die Auswahl der drei folgenden Forschungsfragen deckt den Kreislauf der wissenschaftlichen Wissensgenerierung ab. Deduktion als klassische Methode der Wissensgenerierung leitet basierend auf einer bestehenden Theorie neue Erkenntnisse ab, welche empirisch validiert werden können. Induktion im Gegensatz liefert durch die Analyse empirischer Daten neue Theorien. Um die Erkenntnissen beider Herangehensweisen zu akzeptieren sind wissenschaftliche Methoden und deren korrekte Anwendung elementar. Dies lässt sich unter Vertrauen subsumieren.

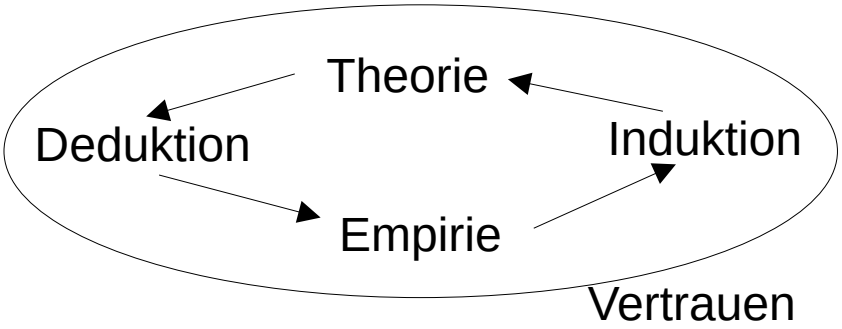

Abbildung 1: Kreislauf der Wissensgenerierung (eigene Darstellung)

In der deduktiven Herangehensweise (Kapitel 3.1) wird eine bekannte Tatsache („Spezialisten“ vs. „Generalisten“) als Grundlage für eine Analyse heutiger IT Fachleute angewendet. Durch den Nachweis der grundsätzlichen Gültigkeit, konnten weitere, detailliertere Erkenntnisse abgeleitet werden. Zeitgleich können die im theoretischen Teil genannten Herausforderungen der Datenqualität praktisch demonstriert werden.

Für die induktive Herangehensweise (Kapitel 3.2) wurde eine Studie zur Studentenmigration, welche auf klassisch gewonnenen Panel Daten basiert, mit Hilfe von Daten, welche aus Social Media Quellen extrahiert wurden, als Inspiration genutzt und die Unterschiede und Gemeinsamkeiten herausgearbeitet.

Trotz der aufgezeigten Möglichkeiten wird dem Data Mining noch immer durch verschiedene Stakeholder Misstrauen entgegengebracht. Das Kapitel 3.3 beleuchtet, welche Faktoren das Vertrauen in das Data Mining und die daraus gewonnen Erkenntnisse beeinflussen können. Dabei wurden bestehende, theoretische Konzepte, welche das Vertrauen beeinflussen, aus Volkswirtschaft,

Psychologie und der Informatik zusammengeführt und praktisch anhand von Umfragen angewendet.

Um dem Vorwurf der Glasperlenspielerei entgegenzutreten, wird gezeigt, wie aus der Anwendung von Data Mining Methoden und der Analyse von Social Media Daten praktische unternehmerische Entscheidungen abgeleitet werden können (Kapitel 3.4). Die folgende Abbildung verdeutlicht den Aufbau der Arbeit graphisch (Abbildung 2).

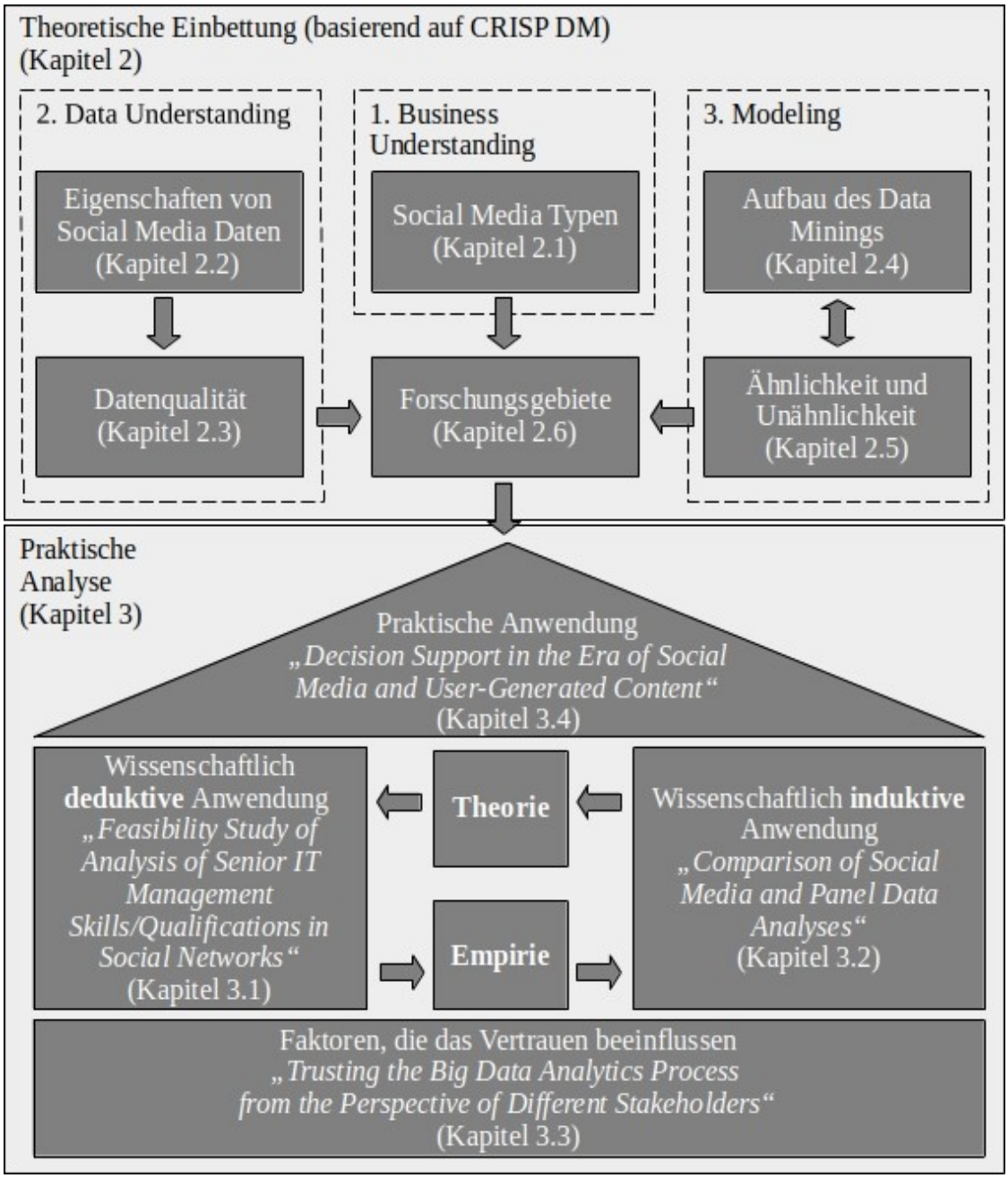

Abbildung 2: Aufbau der Arbeit

2 Theoretische Einbettung

Das Aufkommen von Big Data oder Social Media Data zieht die Aufmerksamkeit von Wirtschaft, Wissenschaft und des Staates auf sich. Beispielsweise investierte die US-Regierung 2012 200 Millionen US-Dollar, um die „Big Data Research and Development Initiative“ zu starten [2]. Das renommierte Wissenschaftsmagazin „Nature“ veröffentlichte allein im Jahr 2021 6317 Forschungsartikel (!) zum Stichwort „social media data“ [3], 2682 Forschungsartikel davon zum Stichwort „big data“. Das namhafte Wissenschaftsjournal „Science“ veröffentlichte im Jahr 2021 197 Artikel (mit Peer-Review) zu Social Media Data [4]. Unter den Top 10 Tech Aktien 2021 befinden sich Alphabet (unter anderem „Google Suchmaschine“) und Meta (unter anderem „Facebook“), beides Unternehmen deren Geschäftsmodell unter anderem auf der Analyse Social Media Daten, Verknüpfung von Information basierend auf angenommenen individuellen Neigungen und schlussendlich passgenauer Werbung basiert. Dies ist ein Hinweis auf die heutige Bedeutung von Big Data und Social Media für Administration, Forschung und Wirtschaft. Durch das Vorhandensein und die Analyse von Big Data aus verschiedensten Quellen gewonnen und mit unterschiedlichen Verwendungszwecken haben Forscher und Entscheidungsträger in Politik und Wirtschaft erkannt, dass diese riesige Menge an Informationen zahlreiche Möglichkeiten beispielsweise für das Verständnis der Kundenbedürfnisse, die Verbesserung der Servicequalität und die Vorhersage und Vermeidung von Risiken von Entscheidungen bietet. Die Nutzung und Analyse von Big Data muss jedoch auf genauen und qualitativ hochwertigen Daten basieren, was eine notwendige Voraussetzung für die Wertschöpfung aus Big Data ist. Dabei ist die Analyse der Qualität der gesammelten Daten ein wesentlicher Aspekt zur Bewertung der späteren Analyseergebnisse. Basierend auf den Daten können die vielfältigsten Algorithmen angewendet werden, die im Spannungsfeld von Geschwindigkeit, Genauigkeit und Generalisierbarkeit stehen.

Im Nachfolgenden werden durch den Autor einige repräsentative Techniken dargestellt und kurz erläutert. Dies ist wichtig zum Verständnis der Social Media Analysen und dem aktuellen Stand, der nachfolgend dargestellt wird (Kapitel „2.1 Überblick über Social Media Typen“).

Wichtig für die Analyse der Qualität, ist die Abgrenzung zwischen Daten und Modell oder Analysequalität. Dem CRISP DM Data Mining Model folgend wird zuerst auf das „Data Understanding“ eingegangen. Dementsprechend wird im Rahmen dieser Arbeit zuerst ein Einblick in die Kriterien gegeben, die die Qualität der Daten beschreiben und strukturieren (Kapitel „2.2 Eigenschaften von Social Media Daten “ und „2.2 Datenqualität“).

Nachfolgend wird entsprechend des CRISP DM Schrittes „Modeling“ auf typische Data Mining Methoden eingegangen (Kapitel „2.4 Data Mining Techniken (Überblick)“ und „2.5 Ähnlichkeits- und Distanzmaße“). Für die Qualität des Analysemodells existieren viele spezifische Qualitätsmerkmale, die hier nicht beschrieben werden können und ein vertieftes Verständnis des jeweils genutzten Modells voraussetzen. Neben dem Überblick über die klassischen Data Mining Methoden im Social Media Umfeld, ist es relevant für das Verständnis einen Einblick in die, den Methoden zugrundeliegenden, Ähnlichkeits- und Distanzmaße zu erhalten. Die Kombination der verschiedenen Methoden und Distanzmaße im Kontext der verschiedenen Social Media Gebiete ermöglicht es, das volle Potential der zur Verfügung stehenden Daten auszuschöpfen.

Nachdem die Voraussetzungen erörtert wurden, wird anschließend ein Überblick über den gegenwärtigen Stand der Forschungsgebiete, welche auf Social Media Daten basieren, gegeben (Kapitel „2.6 Social Media Data Mining Forschungsgebiete“). Die nachfolgende Grafik visualisiert die thematischen Zusammenhänge.

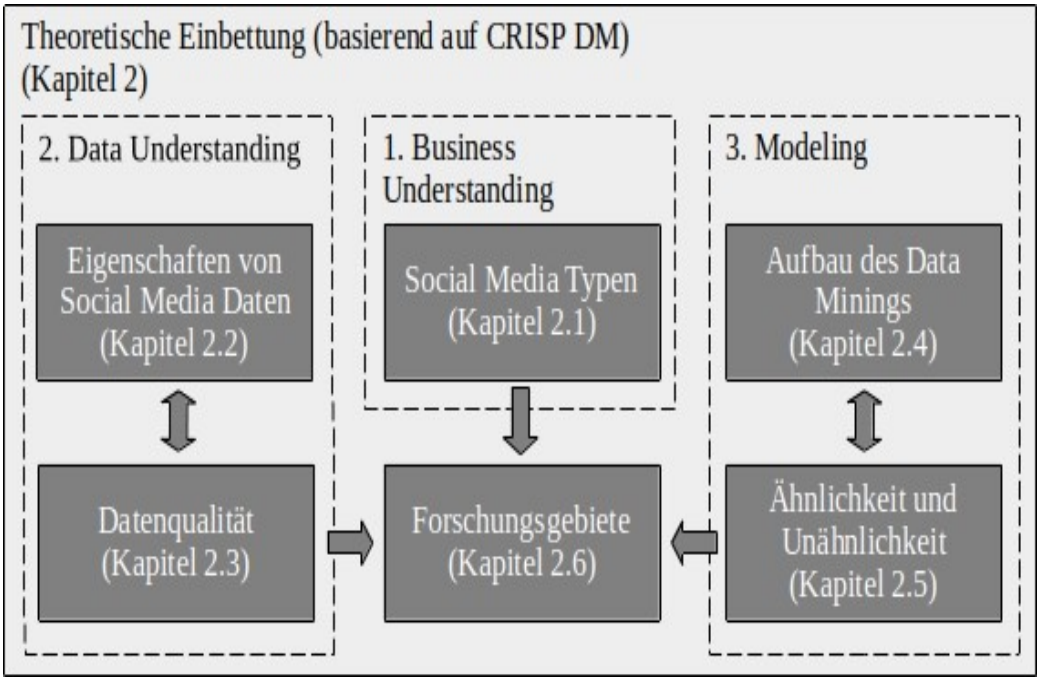

Abbildung 3: Aufbau der Arbeit (Ausschnitt des Zusammenhanges der theoretischen Themen)

Im praktischen Teil der Arbeit (Kapitel „3 Publikationen“) werden diese theoretischen Aspekte im Rahmen von vier Untersuchungen angewendet und veranschaulicht.

2.1 Überblick über Social Media Typen

Social Media werden definiert als Internet basierte Anwendungen, welche auf den Ideen und Technologien des Web 2.0 aufbauen und die Erstellung und den Austausch von Inhalten erlauben, die durch den Benutzer selbst erstellt wurden [5]. Der Benutzer hat die Möglichkeit auf einfache Art und Weise mit anderen Nutzern zu kommunizieren und Netzwerke aufzubauen. Dabei mischen sich traditionelle Medien wie Zeitungen, Radio und TV mit modernen Medien wie XING, Facebook und Twitter. Die Soziale Medien lassen sich, wie die folgende Tabelle zeigt, typisieren, wobei die Grenzen teilweise fließend sind.

Tabelle 1: verschiedene Soziale Medien Typen (eigene Darstellung basierend auf Gundecha et al [6])

Typ	Eigenschaften
online social networks	Netzwerke erlauben realen Personen sich virtuell zu verbinden und Inhalte auszutauschen. z.B. XING, LinkedIn und Facebook.
blogging	Blogs sind zeitungsähnliche Plattformen, in denen die Nutzer vergleichbar mit klassischen Journalisten Beiträge erstellen, bearbeiten und kommentieren. Oftmals auch spezielle Interessen und Einstellungen bedienen z.B. Huffington Post, indymedia, Breitbart (politisch) oder TechRepublic, Wired und TechCrunch (technisch).
micro blogging	Sind ähnlich zu Blogs, jedoch mit stark reduziertem Umfang z.B. Twitter, Weibo, snapchat.
wikis	Wikis dienen der gemeinsamen Bearbeitung von Wissensdatenbanken sowohl global z.B. Wikipedia als auch firmeninterne Plattformen.
social news	Social News sind Aggregatoren zum Sammeln von Nachrichten basierend auf individuellen Präferenzen z.B. Reddit.
media sharing	Eine Sammelbezeichnung für Plattformen, welche speziell darauf ausgerichtet sind Multimedia Inhalte auszutauschen z.B. Youtube, Flickr oder Instagram.
opinions, reviews and ratings	Dies sind Plattformen die es den Nutzern ermöglichen Einschätzungen zu Produkten, Personen oder Firmen zu teilen z.B. Yelp, kununu oder tripadvisor. Große Online Marktplätze haben oftmals ähnliche Funktionen implementiert z.B. Amazon und Alibaba.
problem solving	Diese Plattformen ermöglichen den Austausch von Erfahrungen, Tipps und Tricks für spezielle Themen z.B. stackoverflow.

Als Beginn der Social Media Plattformen kann der micro blog ICQ im Jahre 1996 gesehen werden. Das erste soziale Netzwerk SixDegrees ging 1997 online, jedoch erst mit dem Start von Facebook im Jahr 2004 wurde eine breitere Bevölkerung angesprochen. Die täglichen Nutzerzahlen steigen noch immer und überschreiten teilweise die Einwohnerzahlen der größten Länder.

Tabelle 2: Anzahl der monetarisierbaren täglichen Nutzer (mDAU)

Plattform	mDAU in Millionen (September 2021 [7])
Facebook	1.93
Snapchat	306
Weibo	246
Twitter	192

Amazon, welches 1994 gegründet wurde, gibt den weltweiten Quartalsumsatz 4/2021 mit 137,41 Milliarden US Dollar an [8]. Neben dem reinen Ertrag aus dem Warenverkauf, generiert das Unternehmen eine Vielzahl von Nutzerdaten wie bspw. Besuchs- und Kauffrequenzen, Präferenzen für Produkte und Marken und Informationen zum „cross selling“ Verhalten. Zusätzlich bewerten Nutzer subjektiv die Qualität der Produkte und Dienste.

2.2 Eigenschaften von Social Media Daten

Daten aus sozialen Medien haben charakteristische Eigenschaften und unterscheiden sich von klassisch erhobenen Daten. Social Media Data werden größtenteils durch jeden der Teilnehmer eigenständig generiert und aktualisiert. Der Zeitpunkt der Erstellung oder Aktualisierung ist abhängig von der genutzten Plattform genau bekannt (z.B. Twitter) oder komplett unbekannt (XING). Dabei sind die Daten meist groß oder “big” (“Big Data”), verrauscht („noise“), verteilt, oft unstrukturiert und dynamisch.

Die Anzahl der Daten ist entsprechend abhängig von der Größe der genutzten Plattform (siehe Tabelle 2). Auf Twitter werden täglich über 500 Millionen Beiträge („tweets“) abgesetzt, der populärste Beitrag („Hashtag“) umfasste 400 Millionen Beiträge. Ein Tweet umfasst dabei 280 Zeichen – ca. die doppelte Größe des bisherigen Abschnittes. Allerdings übersteigt die Anzahl der täglichen Tweets 500 Millionen [9]. Instagram setzt auf Bildinformationen, die maximal 15 Sekunden dauern. Diese genannten markieren das obere Ende der Größenskala. Auf der anderen

Seite stehen blogs oder Wissensdatenbanken (z.B. Wikipedia.com, stackoverflow.com oder towardsdatascience.com), deren Beiträge in Größe und Qualität mit den klassischen redaktionellen Alternativen vergleichbar sind.

Als „Noise“ in Bezug auf Social Media Daten sind computergenerierte Beiträge als auch bewusst inkorrekte Beiträge durch die Nutzer selbst zu verstehen. Dies umfasst sowohl falsche Bewertungen („fake ratings/ reviews“) als auch falsche Nachrichten und Berichte („fake news“). “Fake news“ lassen sich definieren als Nachrichten, welche mit Absicht und überprüfbar falsch sind und das Potential besitzen den Leser fehlzuleiten. Davon abzugrenzen sind Beiträge die zwar falsch sind, jedoch nicht mit der Absicht den Nutzer zu täuschen verfasst wurden [10], [11]. Das Phänomen ist als solches nicht neu und schon zu Zeiten der Ägypter, wurden durch Ramses Berichte über die Schlacht von Kadesch 127 vor Christus verfasst, die nicht der Realität entsprachen [12] - jedoch besitzen die heutigen Sozialen Medien eine ungleich höhere Reichweite. Die Vielzahl an Informationsquellen und häufige Wiederholung der Beiträge in unterschiedlichen Medien erschwert die Falsifizierung der Beiträge, und kann zu einem „truth effect“ führen. Dabei wird den Informationen geglaubt, obwohl den initialen Nachrichten misstraut wurde [13]. Die Plattformen, als auch die Regierungen dies als Problem erkannt haben, ist der Umfang dieser Beiträge abhängig vom Thema erheblich. In einer Studie aus dem Jahr 2016 fanden Forscher heraus, dass 50 Prozent der Facebook Beiträge falsch oder sehr stark parteilich waren [8]. Als Konterpart zum klassischen „Noise“ muss auch die Abwesenheit von Beiträgen gesehen und in Analysen beachtet werden. Weltweit sind Anstrengungen seitens Regierungen und der Wirtschaft zu beobachten – teilweise in guter Absicht - den Inhalt zu kontrollieren und schlussendlich zu zensieren ([15], [16]. Gegenwärtig sind Techniken zur Erkennung von falschen Beiträgen Inhalt zahlreicher wissenschaftlicher Arbeiten. [17]–[20]. Dies steht nicht im Widerspruch zu der Anfangs getätigten Aussage, dass die Beiträge grundsätzlich verteilt sind – die Aufwendungen die Beiträge zu kontrollieren ist immens und lässt sich nur in bestimmten politischen Konstellationen und im gewissen Umfang realisieren [21].

Beiträge in Soziale Medien sind grundsätzlich verteilt, dass bedeutet in diesem Zusammenhang, das keine zentrale, steuernde Einheit existiert. Daraus folgt, das die Daten nicht zwingend konsistent in Inhalt oder Qualität sind. Weiterhin gibt es keinen uniquen Schlüssel, um die Beiträge eines einzelnen Nutzers zu kombinieren. Dies macht es schwierig bis unmöglich die Daten der einzelnen Plattformen miteinander zu kombinieren. Das Kapitel Publikation 1: „Feasibility Study of Analysis of Senior IT Management Skills/Qualifications in Social Networks“ kann dies praktisch demonstrieren.

Eine weitere Charakteristik der sozialen Plattformen ist es, dass die Daten oft unstrukturiert vorliegen. Dies folgt zum einen daraus, das Texte per Definition unstrukturierte Daten sind. Während journalistische oder vergleichbare Arbeiten Wert auf Rechtschreibung und Orthographie legen, kann dies für Textbeiträge in sozialen Medien nicht vorausgesetzt werden. Dies führt dazu, das klassische Methoden der natural language processing (NLP) nicht die verlässlichen Ergebnisse und die Qualität liefern im Vergleich zu klassischen Medien [22]. Zeitgleich zu der Entwicklung der sozialen Medien setzte sich die Nutzung von Piktogrammen, sogenannte „emojis“ durch. Diese Piktogramme ergänzen nicht einfach den Wortschatz einer Sprache, sondern bilden sprach- und generationsübergreifend neue Ausdrucksformen, die in unterschiedlichem kulturellen Kontext oder „Blasen“ abweichend interpretiert werden müssen. Ohne Kenntnis des Kontextes und der Bedeutung innerhalb der sozialen Gruppe, sind diese Information nicht eindeutig zu interpretieren [23]. Ein weiteres Problemfeld besteht darin, dass wenig Limitationen hinsichtlich der möglichen Eingabemöglichkeiten vorliegen. Beispielhaft lässt sich dies anhand des Netzwerks XING illustrieren. Bei der Angabe des Arbeitsplatzes gibt es keine Vorgaben zum Namen des Unternehmens. Dies lässt sich praktisch begründen, da die Plattform nicht sämtliche existierenden Unternehmen kennt und somit vorgeben kann. Der Nutzer kann nun frei entscheiden, ob er z.B. Siemens, Siemens Medical, Siemens Healthineers oder weitere Bezeichnungen für das gleiche Unternehmen wählt. Somit ist eine Konsistenzprüfungen kaum möglich. In der ersten Publikation wird dies an einem praktischen Beispiel demonstriert (Publikation 1: „Feasibility Study of Analysis of Senior IT Management Skills/Qualifications in Social Networks“). Hier stellt sich ein gravierender Unterschied zu strukturierten Daten und den damit häufig assoziierten SQL Datenbanken dar.

Soziale Medien leben von dem Verhalten ihrer Nutzer. Ändert sich das Verhalten der Nutzer, reagieren die Plattform mit der Umstrukturierung der Inhalte oder der Einführung neuer Funktionen. Beispielsweise führte die gestiegene Bedeutung der Privatsphäre bei den Nutzern und der Umsetzung der DSGVO zu neuen Sichtbarkeitseinstellungen und verändertem Suchverhalten auf den Seiten [24].

2.3 Datenqualität

Betrachtet man die in den vorgestellten Gebieten (Kapitel 2.6) anfallenden Social Media Daten und deren Forschungsrichtungen und -methoden, fällt auf, das die gewonnen Daten äußerst unterschiedliche Qualitäten haben. So sind Tweets schnell hingeworfene kurze Texte, die häufig nur rudimentäre Beachtung der Orthographie und Grammatik aufweisen und der zusätzlichen

Verwendung von Piktogrammen (aka „emojis“). Technische Blogs sind zwar tendenziell sorgfältiger in Orthographie und Grammatik, mischen aber häufig englische, landesspezifische und spezifische Fachbegriffe ein. Soziale Netzwerke sind häufig unvollständig, d.h. nicht alle Knoten (aka „Untersuchungsobjekte“) und Graphen (aka „Beziehungen der Objekte zueinander“) sind erfasst oder im gewählten Kontext oder Zeitpunkt gültig. Recommender Systeme möchten Vorschläge unterbreiten, auch ohne größere Nutzerhistorie. Dabei kann nicht sichergestellt werden, dass alle Produkte, Artikel oder Webseiten nach den gleichen Standards getagt wurden. Die Kontrolle der Einordnung ist im unüberwachten Lernen naturgemäß schwieriger, als im überwachten Lernen. Analysen im wissenschaftlichen Kontext sind weniger zeitkritisch und ermöglichen eine umfangreichere Datenaufbereitung. Die Qualitätsbewertung der Daten ist enorm wichtig und sollte aus dem jeweiligen Kontext beurteilt werden.

Unabhängig von den auf die Daten im Rahmen der Wissensgenerierung angewendeten Methoden sollten die verwendeten Daten hinreichend repräsentativ hinsichtlich Umfang und Verteilung sein, die Messqualität der Forschungsfrage angemessen und reproduzierbar. Im wissenschaftlichen Kontext ist es dabei wichtig, vor der Verwendung der Daten die Qualität der Daten bzw. der Datenquellen zu analysieren. Die Identifikation und Abgrenzung der Dimensionen der Datenqualität wird in einer Vielzahl der wissenschaftlichen ([25]–[28] und praktischen Literatur [29]–[31] besprochen. Ein allgemein akzeptierter Konsens hat sich noch nicht gebildet. Dies lässt sich auch darin begründen, das die Qualität aus Sicht des Beitrages zu den Zielen und dem Kontext des Datenempfängers definiert wird. Damit spiegeln die unterschiedlichen Ansätze immer auch das konkrete Umfeld wieder. Auch wenn übereinstimmende Dimensionen in den unterschiedlichen Konzepten definiert werden, so variiert die Bedeutung und Intensität je nach Anwendungskontext. Dementsprechend wird auf eine einheitliche Definition verzichtet und es werden nur ausgewählte Ansätze dargestellt.

Um der Vielzahl der Begrifflichkeiten durch unterschiedliche Übersetzungsmöglichkeiten keine weitere Komplexität hinzuzufügen, werden die Begriffe im Englischen belassen.

Redman definiert die Dimensionen der Datenqualität im Kontext von allgemeinen Informationssystemen, und gruppiert die Qualitätsdimensionen nach ihrem zeitlichen, prozessualem Verlauf: Konzeption der Daten, operative Konformität und der Präsentation für den Konsumenten der Daten.

Tabelle 3: Dimensionen der Datenqualität nach Redman (1997) [27]

category	dimensions	
quality of design	• content • scope • granularity	• composition • view consistency • reaction to change
quality of conformance	• accuracy • completeness	• consistency • currency
quality of representation	• appropriate format • format flexibilty • interpretability	• portability • efficient use of storage • ability to represent NULL values

Die Kategorie “quality of design” oder die konzeptionelle Perspektive stellt den Zusammenhang von Nutzererwartungen und Spezifikationen des Informationssystems dar. Dagegen stellt die Kategorie “quality of conformance” den Zusammenhang von Nutzererwartung und der Realisation im Informationssystem dar. In der Kategorie “quality of representation” wird die technische Realisation beschrieben. Informationssysteme können hier als Synonym für Datenquelle gesehen werden.

Wang et al schaffen 1996 ein viel zitiertes Framework [32], [33] und nehmen eine Vielzahl der Dimensionen von Redman auf, verändern und erweitern jedoch die Kategorisierung und ergänzen weitere Dimensionen. Dabei beruhen die beschriebenen Dimensionen nicht allein auf theoretischen Überlegungen, sondern wurden zusätzlich empirisch validiert. Somit handelt es sich auch hier um einen allgemeinen Ansatz, der jedoch die Sicht des Anwenders stärker betont.

Tabelle 4: Dimensionen der Datenqualität nach Wang et al (1996)

category	dimensions[1]	
intrinsic data quality	• believability • accuracy	• objectivity • reputation
contextual data quality	• value-added • relevancy • timeliness	• completeness • appropriate amount of data
represential data quality	• interpretability • ease of understanding	• representational consistency • concise representation
accessibilty data quality	• accessibility • access security	

Dabei umfasst „intrinsic data quality“ Qualitätsmerkmale, welche den Daten inhärent sind. Es zeigte sich in Untersuchungen, das neben den technischen Ausprägungen auch die Einschätzung der Glaubwürdigkeit und der Objektivität aus Sicht der *data consumer* relevant für eine hohe Datenqualität ist. Somit hat die Datenqualität immer auch eine subjektive Komponente, die sich final im Vertrauen manifestiert. Somit steht Vertrauen in die Daten am Anfang jeder weiteren Analyse. Neben der intrinsischen Qualitätsdimensionen ist auch der jeweilige Kontext in dem die Daten stehen aus Qualitätsgesichtpunkten zu bewerten. Dies bedeutet, dass die Qualität der Daten kein fixes Datum ist, sondern für jede Analyse separat betrachtet werden muss. Zusätzlich zu den Einflußfaktoren, die sich aus den Daten selbst und dem Kontext ergeben, fokussiert der repräsentative Datenqualität (engl. „representational data quality“) Aspekt darauf, ob der Nutzer (engl. „data consumer“) die Daten kognitiv erfassen kann und ist somit wiederum subjektiv. Die Abgrenzung zu den intrinsischen Konsumentendimensionen erscheint dem Autor schwierig, da das Vertrauen unter anderem auch von der persönlichen Einstellung abhängt [34]. Während ältere Literatur die Kategorie *accessibilty* nicht explizit aufgeführt wird, trägt die neue Kategorie der Tatsache Rechnung, das heutzutage viele Informationen in elektronische Form vorliegen, die damit prinzipiell leichter veränderlich als gedruckte Quellen sind, und demzufolge die Zugriffs-möglichkeiten (engl. „accessibilty“) und die Zugriffskontrolle (engl. „access control“) stärker relevant werden.

Mit den wachsenden Datenmengen und deren Bedeutung hinsichtlich der Digitalisierung, wurde ab 2009 die Norm ISO 8000 veröffentlicht, welche in der Fassung von 2015 die Informations- und Datenqualität definiert [35]. Während oftmals Daten- und Informationsqualität synonym verwendet wurde, greift diese Norm den Gedanken der Trennung von Daten und Information auf und definiert separate Qualitätsdimensionen für Daten: „*provenance*“, „*accuracy*“ und “*completeness*“ [36]. Für Informationen wurde kein generischer Qualitätsstandard veröffentlicht, da die Kriterien stark kontextabhängig sind und dementsprechend definiert werden müssen [37].

Basierend auf den wissenschaftlichen und regulativen Definitionen, wird in der Praxis sechs Dimensionen die größte Bedeutung beigemessen. Dies lässt sich anhand einer Vielzahl von aktuellen, online Diskussionen nachvollziehen. [29]–[31]

Tabelle 5: Dimensionen der Datenqualität in fachspezifischen blogs (2021)
1) [29] 2) [31] 3) [30]

practical definition	dimensions	
	• accuracy [1/2/3]	• uniqueness [1/3]
	• completeness [1/2/3]	• integrity [2]
	• consistency [1/2/3]	• validity [1]
	• timeliness [1/2]	• currency [3]

Während sich viele Dimensionen von Wang et al wiederfinden lassen, konzentriert man sich eher auf die technischen Aspekte. Diese Dimensionen sind quantitativ messbare Dimensionen, was die Bestimmung der Datenqualität verschiedener Quellen leichter möglich macht und technische Strategien zur Verbesserung der Qualität ermöglicht.

„Accuracy“ bezieht sich dabei auf den Grad, mit dem Daten die „realen“ Objekte korrekt darstellen. „Consistency“ bezieht sich darauf, dass Datenwerte in einem Datensatz mit Werten in einer anderen Datenquelle widerspruchsfrei sind. „Completeness“ beschreibt, dass Datenelemente einen Wert haben müssen, wenn sie einen Wert in der „realen Welt“ haben. Dies impliziert die Fähigkeit, sowohl Nullwerte zuzuweisen und identifizieren zu können, als auch Bedingungen unter denen sie gelten. Die Messung der Aktualität oder der Reaktion der Daten auf Veränderungen in der realen Welt erfolgt über „timeliness“ bzw. „currency“. „Uniqueness“, „integrity“, „validity“ tragen der verteilten Datenhaltung Rechnung und betonen dies über die bekannten Dimensionen hinaus.

Rahm et al [38] geben einen Überblick anhand von Beispielen für die Ursachen mangelhafter Datenqualität – bezogen auf die technischen Aspekte.

Tabelle 6: Ursachen für mangelhafte Datenqualität nach Rahm [38] erweitert durch Laranjeiro et al [39]

Problem types)Problem Typen)		Data Quality Problems	Accessibility	Accuracy	Completeness	Consistency	Currency
Source (Quelle)	Level						
Single (Einzelfall)	Instance	Missing data (fehlende Daten)		x	x		
		Incorrect data (falsche Daten)		x			
		Misspellings (Schreibfehler)		x			
		Ambigueous data (mehrdeutige Daten)	x	x			
		Extraneous data (fremde/ nebensächliche Daten)	x			x	
		Outdated temporal data (veraltete temporäre Daten)		x			x
		Misfielded data (falsch eingegebene Daten)	x	x	x	x	
		Incorrect references (falsche Referenzen)		x			
		Duplicates (Duplikate)	x				
	Schema	Domain violation (Domänenverletzung)		x			
		Violation of functional dependency (Verletzung der funktionalen Abhängigkeit)		x			
		Wrong data type (falscher Datentyp)	x			x	
		Referential integrity violation (Verletzung der referentiellen Integrität)	x	x	x	x	
		Uniqueness violation (Eindeutigkeitsverletzung (Schlüssel))		x			
Multiple (Über-greifend)	Instance	Structural conflicts (Strukturelle Konflikte)	x			x	
		Different word orderings (unterschiedliche Wortordnung)	x			x	
		Different aggregation levels (unterschiedliche Aggregationsebenen)	x	x		x	
		Temporal mismatch (zeitliche Diskrepanz)		x		x	x
		Different units (unterschiedliche Einheiten)	x			x	
		Different representation (unterschiedliche Darstellung)	x			x	
	Schema	Use of synonyms (Verwendung von Synonymen)	x				
		Use of homonyms (Verwendung von Homonymen)	x				
		Use of special characters (Verwendung von Sonderzeichen)	x				
		Different encoding format (andere Kodierung)	x			x	

Sind die Daten in strukturierten Datenbanken (z.B. SQL Datenbanken) abgelegt, lassen sich viele der Fehler verhindern bzw. nachträglich detektieren. Liegen die Daten unstrukturiert („flat file“) oder semi-strukturiert („No SQL Datenbanken“) vor, ist das Erkennen von Fehlern ungleich schwerer. Viele produktive Datenbanken in Sozialen Netzwerken gehören zu den beiden letztgenannten Typen [40]. Zusätzlich erschwerend kommt hinzu, dass aufgrund von gewachsenen Anforderungen an den Datenschutz, Daten nicht validiert werden können oder dürfen [41], [42].

2.4 Data Mining Techniken (Überblick)

Maschinelles Lernen ist ein Forschungsfeld, welches die verschiedenen Ansätze des Erkenntnisgewinnes aus Daten formalisiert. Dabei verfolgt es einen interdisziplinären Ansatz und verbindet verschiedene Forschungszweige wie Statistik, Informatik, Ingenieurswissenschaften und andere naturwissenschaftliche und mathematische Forschungsrichtungen [43]. Dabei lassen sich die verwendeten Data Mining Techniken klassisch in zwei Teilbereiche unterteilen, dem überwachten (“supervised learning”) und dem unüberwachten Lernen (“unsupervised learning”). Relativ neu sind Verfahren, die als halb-überwachtes Lernen („semi-supervised learning“) bezeichnen werden.

2.4.1 Überwachtes Lernen (supervised learning)

Die grundlegende Idee des überwachten (engl. supervised learning *)* Lernens basiert auf folgendem Ansatz. Jeder Datenobjekttyp besteht aus einer Reihe von Eigenschaften (engl. *“features”*). Die Matrix aus den einzelnen Datenobjekten und den zugehörigen Attributen wird auch als Daten oder *data* bezeichnet. Mit dem Datenobjekttyp ist ein Ergebnisvektor assoziiert. Dieser Ergebnisvektor besteht aus einem oder mehreren Ergebnisattribute. Häufig sind diese Ergebnisvektoren eindimensional. Die einzelnen features des Datenobjektes bestimmen die Ergebnisattribute. Je nach Typ des abhängigen Ergebnisattributes unterteilt sich das supervised learning wiederum in zwei Teilbereich. Liegen die Ergebnisse in diskreter Form vor z.B. „ja“/ „nein“ oder „groß“, „mittel“, „klein“, wird von Klassifizierung gesprochen. Im Fall von stetigen Skalen wird von Regression gesprochen. Dabei lassen sich die einzelnen Methoden nicht immer eindeutig in Klassifizierung und Regression unterteilen, einige funktionieren nur für eine Aufgabe, andere sowohl für Klassifizierung als auch für Regression. Das Ziel der Verfahren ist es, mit Hilfe von Ähnlichkeiten das wahrscheinliche Ergebnis neuer, unbekannter Datenobjekte zu prognostizieren. Typische Vertreter sind die Algorithmen k- nearest- neighbour, support vector machines und verschiedene Entscheidungsbäume wie CART oder C4.5. Zusätzlich existieren Verfahren, die sich der grundlegenden Methoden bedienen und diese zu sogenannten ensemble methods (dt. Ensemble-Methoden) kombinieren. Ensemble-Methoden lassen sich in drei Meta-Algorithmen unterteilen:

- **Bagging** kombiniert eine hohe Anzahl gleicher Basismethoden, welche unabhängig voneinander trainiert werden und die Einordnung dann per Mehrheitsentscheidung fällen. Der derzeit prominenteste Vertreter ist der random forest.

- **Boosting** kombiniert eine Anzahl gleicher Basismethoden sequentiell, wobei jeder nachfolgende Algorithmus das Ergebnis des Vorgängers zu verbessern versucht. Prominente Vertreter sind AdaBoost und XGBoost.
- **Stacking** ist sehr ähnlich zum bagging, der Unterschied liegt in der Verwendung unterschiedlicher Basismethoden. [44] Aufgrund zahlreicher ungeklärter Fragen, insbesondere der Auswahl der zu kombinierenden Algorithmen hat sich stacking nicht durchgesetzt und wird hier nur der Vollständigkeit halber aufgeführt.

Für die Abschätzung der Güte der gewonnenen Modelle stehen zahlreiche Kennzahlen zur Verfügung. Zu unterscheiden sind Kennzahlen, welche die Güte des Models im Allgemeinen abschätzen und Kennzahlen, welche die spezielle Güte der Klassifikation oder Regression abschätzen. Die generelle Güte und Komplexität des Modells werden mit Kennzahlen wie dem Akaike Informationskriterium (AIC) oder Baysian Informationskriterium (BIC) untersucht. Betrachtet wird die Vorhersagequalität des trainierten Modells in Bezug auf eine Grundwahrheit (aka die klassifizierten Datensätze), d.h. dem Informationsverlust unter Berücksichtigung der Komplexität des Models, d.h. eine höhere Anzahl an verwendeten Parameter wird bestraft. Somit kann eine Über- oder Unteranpassung der Modelle abgeschätzt und entgegengewirkt werden. Sowohl AIC als auch BIC sind jedoch keine absoluten Kennzahlen und können ausschließlich in Relation zu einem zu vergleichenden Modell auf gleichen Datenbasis gesehen werden [45]. Die Vorhersagequalität eines Klassifikationsmodells selbst kann mit Kennzahlen wie die Rand Accuracy, Precison und Recall bewertet werden (keine deutsche Entsprechung). Es existieren Ansätze diese Kennzahlen in Regressionsmodellen zu verwenden [46], jedoch sind statistische Kennzahlen wie die Summe der quadratischen Abweichung („SSE“ oder „RSS“), das Bestimmtheitsmaß („R^2“) oder der Signifikanzwert („p-value“) erprobt und weitaus verbreiteter (vertiefend z.B. bei Draper [47]).

2.4.2 Unüberwachtes Lernen (unsupervised learning)

Unüberwachtes Lernen wird mit wenigen Ausnahmen (z.B. Principal Component Analysis (PCA)) als Synonym für die Cluster Analyse genutzt [48]. Die Bezeichnung „unüberwacht“ ist darauf zurückzuführen, dass keine Informationen über die Zuordnung der Datensätze zu einer Klasse vorliegen, oder generischer ausgedrückt, kein dependentes Attribut vorliegt. Grundsätzlich basiert Clustering darauf, Gruppen zu identifizieren, deren Mitglieder einander ähnlich und möglichst unähnlich zu Mitgliedern (Kapitel 2.5) anderer Cluster sind.

- **Partitionierende** Verfahren zählen zu den bedeutendsten Methoden der Clusteranalyse. Basierend auf der Annahme, dass die Anzahl der Cluster bekannt ist, werden zufällig Clusterzentren bestimmt. Danach wird in iterativen Verfahren die Distanzen eines jeden Punktes zu den Clusterzentren bestimmt und die Punkte dem Clusterzentrum mit der minimalen Distanz zugewiesen. Nach der Neubestimmung der Clusterzentren wiederholen sich die Schritte bis ein Abbruchkriterium erreicht wird. Die gefundene Lösung kann in Abhängigkeit der Startwerte variieren und wird oftmals mit einem bagging Ansatz stabilisiert. Bekannte Verfahren sind k-means++ [49], das expectation-maximisation Verfahren [50] oder das relativ neue Verfahren der affinity propagation [51].
- **Hierarchische Verfahren** teilen sich in einen divisiven (Bottom Up) und einen additiven (Top Down) Ansatz. Im divisiven Ansatz wird ein Cluster entlang der Achse der größten Varianz iterativ geteilt. Im additiven Ansatz wird für jeden Punkt oder Cluster der nächste d.h. ähnlichste Punkt oder Cluster bestimmt und iterativ hinzugefügt. Die Anzahl der Cluster muss dabei initial nicht bekannt sein, sondern kann in gewissen Grenzen während der Berechnung abgeleitet werden. Bekannte Vorgehensweisen unterscheiden sich weniger in der Methodik an sich, vielmehr unterscheiden sich die Verfahren dadurch wie getrennt oder zusammengefügt wird, den sogenannten linkage Verfahren.
- **Dichtebasierte Verfahren** basieren auf der Annahme, dass Punkte zusammengehören die in nahe zusammenstehen, das heißt sich in einem gewissen Radius zueinander befinden. Ein Clusterzentrum wie in den oben beschriebenen partitionierenden oder hierarchischen Verfahren existiert nicht. Vorteilhaft ist, das die Anzahl der Cluster nicht a priori bekannt sein muss. Ebenso vorteilhaft erweisen sich diese Verfahren zur Identifizierung von Ausreißern (engl. outlier), die als kleine separate Cluster erscheinen. Während in den ursprünglichen Ansätzen (DBSCAN [52]) der Dichteparameter vorgegeben werden musste, erlauben neuere Ansätze in Analogie zu den Hierarchischen Verfahren die bessere Steuerung des Parameters [53].
- **Graphen basierte Verfahren** können als Sonderform der divisiven hierarchischen Verfahren gesehen werden. Die Clusterzugehörigkeit wird über Zentralitätsmaße anstelle der klassischen Distanzmaße bestimmt (Kapitel 2.5). Je nach verwendetem Maß, wird der initiale Cluster aufgespalten indem die Graphen zwischen den Knoten mit dem geringstem Beitrag oder einer minimalen Anzahl an Verbindungen getrennt werden. [54]

- **Modell basiertes Clustering** ist ein statistischer Ansatz für das Clustering. Es wird angenommen, dass die beobachteten (multivariaten) Daten aus einer Mischung von parametrischen, multivariaten Wahrscheinlichkeitsverteilungen generiert wurde. Das heißt, ein Clusterzentrum wurde durch ein statistisches Model generiert, die weiteren Punkte im Cluster sind statistische Abweichungen dieses Models. Aufgabe ist es mit Hilfe der Maximum Likelihood Methode die Clusterzentren und die zugehörigen Parameter der Modelle zu schätzen [55]. Ein bekannter Vertreter ist das Expectation Maximisation Verfahren.

Die Verifikation der Ergebnisse der Analyse oder Modellierung mit Hilfe von Kennzahlen ist, anders als beim überwachten Lernen, schlecht bis nicht möglich und muss durch einen mit dem Kontext vertrauten Experten überprüft werden. Kennzahlen können hier Hinweise geben, und die Cluster haben per se keine semantische Bedeutung [48, p. 20 ff.] und müssen aus dem jeweiligen Kontext (aka „Business Understanding“) heraus interpretiert werden. Beispielsweise erfasst die traditionelle biologische Taxonomie Lebewesen und Viren nach morphologischen Merkmalen. Die von Linne eingeführte Klassifikation zielte auf eine einheitliche Benennung nach optischen Merkmalen der Pflanzen ab. Dies ist aus Sicht von Gärtnern, denen es vorrangig darum geht Pflanzen nach optischen Merkmalen zu bestimmen, ideal. Um phylogentische Zusammenhänge zu beschreiben ist diese Art der Klassifikation (oder Clusterung) jedoch ungeeignet und wird durch Genommerkmale, also durch eine auf Verwandtschaften basierte Klassifikationen, abgelöst [56].

Bekannte Kennzahlen sind der Silhouettenkoeffizient [57], der Calinski Harabasz Index [58] und der Rand Index [59]. Der Silhouettenkoeffizient und der Calinski Harabasz Index ermöglichen z.B. die Berechnung eines Wertes für jeden Punkt, der angibt wie gut die Zuordnung zur gewählten Gruppe im Vergleich zu allen deren Gruppen erfolgt ist. Der Silhouettenkoeffizient wird oft auch visualisiert, und zeigt dann die Struktur der einzelnen Cluster an, wobei dann der Wert der Kennzahl an sich weniger relevant ist. Für beide Kennzahlen ist keine Kenntnis der realen, tatsächlichen Zuordnung notwendig, was aber im Gegenzug auch bedeutet, das zwar eine Lösung gefunden wurde, die aber nicht zwingend in der Realität oder in Bezug auf die Forschungsfrage Sinn machen muss. Auch hier gilt, dass die gefundene Clusteranzahl mit dem Business Understanding in Einklang gebracht werden muss. Der Rand Index hingegen vergleicht die Güte zweier Clusterlösungen, wobei die Vergleichslösung entweder ein alternatives Cluster-Modell, oder aber die gleichen, klassifizierten Datenpunkte sind. In diesem Fall berechnet er die Ähnlichkeit zwischen den zwei Clustern und zählt die Paare, welche in den vorab klassifizierten und damit „wahren“ Clustern demselben oder unterschiedlichen Clustern zugeordnet sind.

2.4.3 Halb-überwachtes Lernen (semi supervised learning)

Im überwachten Lernen liegen klassifizierte Datensätze bereits vor. Dies ist im Umfeld von Social Media Daten jedoch häufig nicht der Fall. Die Klassifizierung müsste durch Business User (dt. hängt vom Kontext ab, Übersetzung als „Domänenexperten“ üblich) erfolgen. Da menschliche Einschätzungen verzerrt oder subjektiv widersprüchlich sein könnten, sollte ein 6-Augen oder anderweitig validierender Prozess angewendet werden. Betrachtet man die vorliegenden Datenmengen, so wird klar, dass dies nicht nur kostenintensiv oder zeitaufwendig wäre, sondern in vielen Bereichen unmöglich, insbesondere wenn die Analyse zeitnah Ergebnisse liefern soll. Halb-überwachtes Lernen behebt diesen Engpass, indem es dem Modell ermöglicht, einen Teil oder alle der verfügbaren unbeschrifteten Daten in sein überwachtes Lernen zu integrieren. Das halb-überwachte Lernen nutzt dabei sowohl bereits klassifizierte, als auch die initial unklassifizierten Daten. Halb-überwachtes Lernen basiert auf zwei bis drei Grundannahmen: smoothness assumption (dt. „Ähnlichkeitsannahme“) [60], cluster assumption („Clusterannahme“) [61] und der manifold assumption („Mannigfaltigkeitsannahme“) [60]. Die Ähnlichkeitsannahme impliziert, dass wenn zwei Datensätze in Bereichen mit einer hohen Dichte nahe beieinander liegen, sie sich in ihren Klassen ähnlich sind. Entsprechend ist für Punkte in Bereichen mit geringerer Dichte die Wahrscheinlichkeit höher, dass sie zu unterschiedlichen Klassen gehören. Dies bedeutet wiederum, das in Räumen mit geringer Dichte vermutlich die Klassengrenzen liegen. Die *Clusterannahme*, eigentlich eine logische Ableitung der *Ähnlichkeitsannahme*, postuliert, dass Punkte im selben Cluster zur selben Klasse gehören. Die *Mannigfaltigkeitsannahme* geht davon aus, das sich Daten die sich lokal, in einem geringer dimensionalen Subraum, ähneln, auch in einem hoch-dimensionalen Raum ähneln. Dies ermöglicht die Anwendung von Ähnlichkeitsanalysen im niedrig-dimensionalen Raum und dadurch einem geringerem Einfluß des „Fluchs der Dimensionalität“.

Hinsichtlich der Forschungsrichtungen existieren verschiedene Ansätze.

Halb-überwachtes Lernen mit seinen gegenwärtig 4 Bereichen [62, p. 8 ff.], [63] erfährt im wissenschaftlichen Umfeld momentan große Aufmerksamkeit, sodass zukünftig weitere Ansätze zu erwarten sind [64]:

- Halb-überwachtes Lernen mit **Graphen:** Gemeinsamkeit dieser Methoden ist es, dass die Daten als Knoten und die paarweisen Abstände durch die Kanten abgebildet werden. Die minimale Entfernung zwischen 2 Punkten kann nun als kürzester Pfad zwischen den beiden Punkten beschrieben werden. Die Mannigfaltigkeit wäre somit die Krümmung im geodätischen Raum.

- Halb-überwachtes Lernen mit **generative Modellen**, wobei basierend auf bedingten Wahrscheinlichkeiten (Bayes), iterativ Datenpunkte hinzugefügt werden, d.h. Punkte die eine hohe bedingte Wahrscheinlichkeit mit den bereits klassifizierten Daten haben, werden identifiziert, diese Punkte werden den klassifizierten Daten hinzugefügt und dies wird wiederholt bis ein Abbruchkriterium eintritt. Dieser Iterationsansatz folgt der Idee des expectation maximisation Verfahrens des unüberwachten Lernens.
- Halb-überwachtes Lernen durch **Trennung bei geringer Dichte.** Dabei werden Gebiete mit geringer Dichte von Datenpunkten, entsprechend der Ähnlichkeitsannahme, als Gebiete mit einem wahrscheinlichen Klassenwechsel betrachtet. Grundsätzlich entspricht die Idee den Support Vektor Maschinen (SVM) des überwachten Lernens, wobei die Grenzen zwischen den Klassen iterativ abgeleitet werden. Anders jedoch als bei dem induktiven SVM Ansatz des überwachten Lernens, fließen in dem transduktiven Vorgehen der semi supervised SVM Version (TSVM) die Geometrie bzw. Dichte der unklassifizierten Punkte in die Schätzung der Klassengrenzen mit ein [65].
- Halb-überwachtes Lernen **mit Änderung der Darstellung/ Perspektive**. Diese Bereich verfolgt einen ähnlichen Ansatz wie der Bagging des überwachten Lernens. Hier ist das Ziel, verschiedene Modelle mit unterschiedlichen Ansichten der Daten zu trainieren. Im Idealfall ergänzen sich diese Ansichten und die Modelle können zusammenarbeiten, um die Leistung des anderen zu verbessern. Diese Ansichten können sich auf unterschiedlichste Weise differenzieren, z. B. in den von ihnen verwendeten Funktionen, in den Architekturen der Modelle oder in den Daten, mit denen die Modelle trainiert werden. Grundsätzlich werden in einem ersten Schritt sämtliche Daten mit Hilfe von unsupervised learning Algorithmen analysiert. Dabei sollte eine neue Perspektive auf die Daten bzw. Modell entstehen. In einem zweiten Schritt, wird nun dieses neue Modell mit einem supervised learning Algorithmus auf die klassifizierten Daten angewendet[62, p. 10 ff.]. Ein leistungsstarker Ansatz ist das Tri-Training [66] bei dem 3 Modelle auf dem gleichen - mittels bootstrapping variierten - Trainingsdatensatz, trainiert werden. Datenpunkte werden für klassifiziert erklärt, sobald alle Modelle zur gleichen positiven Klassifizierung kommen. Die Datenpunkte werden dann dem jeweilige Trainingsset hinzugefügt. Eine Erweiterung des Ansatzes, und in gewissem Sinne eine antagonistische Perspektive, fügt die Daten zu dem Trainingsset I nur hinzu, wenn der Klassifikator I den Datensatz mit einer anderen Kategorie klassifizieren würde, als die beiden anderen Klassifikatoren. Somit wird der Klassifikator speziell in seinen

Grenzbereichen optimiert [67]. Dies entspricht der Idee des Boostings innerhalb des überwachten Lernens (Kapitel 2.4.1)

Ein Ansatz nutzt die klassifizierten Datensätze um das Cluster Modell zu identifizieren und die unklassifizierten Datensätze um die Clustergrenzen genauer abzugrenzen [48]. Dies kann später wiederum die Grundlage für weitere Klassifizierungsalgorithmen sein, die dann auf eine breitere Datenbasis zurückgreifen können. Die bisherigen Publikationen mit Bezug auf das halb-überwachte Lernen zeigen viele Potentiale auf, auch wenn auf Algorithmenebene noch viele Unklarheiten herrschen [68].

2.5 Ähnlichkeits- und Distanzmaße

Zur Identifikation zusammenhängender Gruppen oder Muster ist grundsätzlich die Ähnlichkeit oder Unähnlichkeit der Untersuchungsobjekte zu bestimmen. Die verwendeten Metriken bestimmen maßgeblich die möglichen Algorithmen des überwachten oder unüberwachten Lernens. Die Bestimmung der Ähnlichkeit oder Distanz hängt wiederum unter anderem von den Metriken der Daten ab. Das folgende Kapitel gibt einen Überblick über die unterschiedlichen Konzepte und erhebt keinen Anspruch auf Vollständigkeit.

Ähnlichkeits- und Unähnlichkeitsmaße stehen im Zusammenhang. Eine Ähnlichkeitsmaß für 2 Objekte, ob Punkte oder Cluster, hat typischerweise den Wert 0 wenn sie sich nicht ähnlich sind. Je größer der Wert, desto ähnlicher sind sich die Paare. Typischerweise drückt der Wert 1 Identität aus. Unähnlichkeit wird dazu im Gegensatz definiert, d.h. Ähnlichkeit = 1 – Unähnlichkeit. Unähnlichkeit wird häufig auch als Distanz bezeichnet [48, p. 65 ff.]. Ist die Ähnlichkeit aus Sicht beider Paare identisch, spricht man von symmetrischen Maßen. Ein typischer Vertreter ist die Minkowski Distanz. Ist die Richtung entscheidend für die Distanz, spricht man von asymetrischen Distanzen oder Ähnlichkeitsmaßen, wozu bspw. die Jaccard oder die Kullback Leibler Divergenz zählen.

Metrisch skalierte Attribute

Für metrische skalierte features lassen sich Entfernungen definieren und messen. Bekannte metrische Distanzmaße sind die klassische euklidische Distanz (p=2) und die Manhattan Distanz oder taxicab distance (p=1). Generisch lässt sich dies mit der Minkowski Distanz darstellen:

$$X = (x_1, x_2, \ldots, x_n) \quad und \quad Y = (y_1, y_2, \ldots, y_n) \in set\, R^n$$

$$d(X, Y)_p = \left(\sum_i^n \|x_i - y_i\|^p\right)^{\frac{1}{p}} \quad mit \quad 0 < p < \infty \quad [69]$$

Die Minkowski Distanz setzt für eine sinnvolle Anwendung die Unabhängigkeit der Attribute (Orthogonalität) voraus. Korrelieren die Attribute stark miteinander, werden die Distanzen verzerrt und tendenziell zu gering berechnet, und demzufolge wird eine stärkere Nähe suggeriert (Multikollinearität).

Obwohl es sich um ein etabliertes Entfernungsmaß handelt, ist die Minkowski Entfernung nicht maßstabsunabhängig, d.h. dass die berechneten Entfernungen in Abhängigkeit der Einheiten der Features möglicherweise verzerrt sind. Typischerweise sollten die Daten normalisiert werden, bevor

die Distanzmaße verwendet werden können. Darüber hinaus werden diese Distanzen mit zunehmender Dimensionalität der Daten , d. h. Anzahl der Features, weniger nützlich. In hohen Dimensionen tritt das Phänomen auf, welches als „Fluch der Dimensionalität“ bezeichnet wird: Das Verhältnis zwischen dem nächsten und dem entferntesten Punkt nähert sich 1 (bei Gleichverteilung in R^n), d.h. die Punkte sind im Wesentlichen gleich voneinander entfernt. Dieses Phänomen kann für eine Vielzahl von Distanzmetriken beobachtet werden, ist jedoch für die euklidische Metrik ausgeprägter als beispielsweise für die Manhattan-Distanzmetrik. Die Prämisse der Suche nach dem nächsten Nachbarn ist, dass "nähere" Punkte relevanter sind als "entferntere" Punkte, wenn aber alle Punkte im Wesentlichen gleichmäßig voneinander entfernt sind, ist die Unterscheidung bedeutungslos. Im gewissen Umfang kann dem durch ein p<1 entgegengewirkt werden [70], [71] was dann lt. Definition keine Distanzmetrik wäre, da die Dreiecksungleichung nicht mehr erfüllt ist.

Mahalanobis Abstand

Im Rahmen einer Clusteranalyse verwendet man häufig den Mahalanobis-Abstand als Alternative zur Minkowski Distanz [72]. Die Grundidee dabei ist, zu messen wie viele Standardabweichungen ein Punkt vom Mittel entfernt ist. Während die Minkowski Distanz die Unabhängigkeit der Features voraussetzt (orthogonale Achsen), korrigiert der Mahalanobis Abstand dies zusätzlich mit Hilfe der Kovarianz Matrix. Somit korrigiert sie eine Übergewichtung, die bei korrelierenden Achsen den Abstand verzehrt. Wenn die Kovarianz Matrix diagonal ist, entspricht die Mahalanobis-Distanz der euklidische Distanz. Alternativ könnte dies durch eine Vorschaltung einer PCA, vor der Berechnung der Standard Minkowski Distanz, erreicht werden. Problematisch ist die Anfälligkeit gegenüber Ausreißern [73], die gegebenenfalls vorab bereinigt werden sollten.

$C_i = \frac{1}{n_i} \hat{X}^T \hat{X}$ Kovarianz für Cluster i

$S = \frac{1}{n} \sum_{n=1}^{g} m_i c_i$ gepoolte Kovarianz für alle Cluster mit:

m = Anzahl der Elemente in Gruppe i

n = Anzahl aller Element

$D_M = \sqrt{(x-\bar{x})^T S^1 (x-\bar{x})}$ Mahalanobis Distanz

Ordinal skalierte Attribute

Ordinal skalierte Variablen lassen sich nicht mit der Minkowski Distanz bestimmen. In diesem Fall wird in verbreiteten Verfahren die Anzahl der Löschen, Verschieben und Einfügeoperationen betrachtet, die notwendig sind, um zwischen den zu vergleichenden Datenpunkten einen identischen Vektor zu erhalten. Ein typisches Maß ist die ULAM Distanz. Illustrieren lässt sich dies wie folgt:

Gegeben sind fünf Alternativen: A1, A2, A3, A4 und A5.
Präferenz Nutzer 1: X=(A1, A2, **A3**, A4, **A5**)
Präferenz Nutzer 2: Y=(A1, A2, **A5**, A4, **A3**)
Schritt 1:
Löschen: Y=(A1, A2, , A4, **A3**)
Verschieben: Y=(A1, A2, A4, **A3**)
Einfügen: Y=(A1, A2, A4, **A3, A5**)
Schritt 2:
Löschen: Y=(A1, A2,A4, , A5)
Verschieben: Y=(A1, A2, A4, A5)
Einfügen: Y=(A1, A2, **A3,** A4**,** A5)
Die resultierende ULAM Distanz beträgt: 2

Weitere bekannte Maße für ordinal skalierte Skalen sind Hamming Distanz, Kendal Distanz, Cayley Distanz und die Spearman Distanz [69].

Binär und kategorial skalierte Attribute

Für binär skalierte Variablen kann man im Grunde nur die Anzahl der Unterschiede zählen. Dafür existieren eine Vielzahl an Verfahren zur Messung [74]. Die beiden nachfolgenden Distanzen beschreiben die üblichsten Distanzmaße.

Hamming Distanz

Die Hamming-Distanz misst die Anzahl der Werte, die sich zwischen zwei Vektoren unterscheiden. Diese wird gewöhnlich verwendet, um zwei binäre Zeichenfolgen gleicher Länge zu vergleichen, kann aber auch auf ordinale Skalen angewendet werden.

Jaccard-Koeffizient

Der Jaccard-Index beschreibt eine Statistik, die zum Vergleich der Ähnlichkeit und Unähnlichkeit von Stichprobensätzen verwendet wird. Er errechnet sich aus der Schnittmenge dividiert durch die Größe der Vereinigung der Stichprobensätze. Obwohl dieser Ansatz relativ alt und einfach ist,

basieren viele Analysen zur Ähnlichkeit von Gruppen in Sozialen Medien auf ihm [75]–[77]. Empfehlungssysteme (engl. Recommender Systems, siehe Kapitel 2.6.5) bedienen sich seiner zur Bestimmung der Unähnlichkeit um die Empfehlungen vielfältiger zu gestalten [78]. Hier wird der Anteil errechnet zwischen der Anzahl der Artikel oder Produkte, für die zwei Benutzer gemeinsam abgestimmt haben und die Anzahl der verschiedenen Artikel oder Produkte, für die beide Benutzer insgesamt gestimmt haben, d.h. die Schnittmenge dividiert durch die Vereinigung der abgestimmten Punkte [79].

$$S_J(A,B)=\frac{|A\cap B|}{|A\cup B|}$$ Jaccard Ähnlichkeitsmaß/ Koeffizient

Ein bedeutender Nachteil des Jaccard-Indexes ist, dass er stark von der Größe der Daten beeinflusst wird. Große Datensätze können einen großen Einfluss auf den Index haben, da sie die Vereinigung erheblich erhöhen könnten, während die Schnittmenge ähnlich bleibt.

Bei kategorialen Variablen, d.h. die nominale Variable besitzt mehr als zwei Ausprägungen, wird oftmals die binäre nominale Variable mit k Ausprägungen in eine binäre Darstellung mittels der One-Hot Kodierung in k neue Variablen transformiert, dies ist jedoch nicht zwingend. Die Reihenfolge ist dafür unerheblich.

Gemischte Skalen

In der Praxis haben wir nur sehr selten Messskalen eines einzigen Typs. In praktischen Untersuchungen im sozialwissenschaftlichen Umfeld haben die zugrundeliegenden Daten aus gemischten Messskalen mit nominalen, ordinalen und metrischen Skalen. Für einen typischen Fall einer Person, welche Beiträge in einer *problem solving platform* (dt. „Problemlösungsplattform“) teilt, ist das Geschlecht nominal, die Themen welche er liest oder beantwortet sind kategorial, die Häufigkeit in der monatlich Beiträge besucht oder beantwortet werden und die Länge der Beiträge metrisch. Dabei sollte das Geschlecht nur einen geringen Einfluss haben, und die Häufigkeit mit denen Nutzer Beiträge teilen wird in der Regel deutlich kleiner sein als die Anzahl der Wörter der Beiträge. Dies stellt ein Problem dar.

Um die Distanzen zwischen den Objekten trotzdem bestimmen zu können wird meist wie folgt vorgegangen:

- regelmäßig werden nur normalisierte Distanzen oder Ähnlichkeiten für alle metrischen Attribute verwendet

- die Gewichtung wird für jedes Attribut bestimmt (entweder gleichverteilt oder basierend auf individuellen Annahmen)
- die allgemeine aggregierte Ähnlichkeit gibt den gewichteten Wert der Distanzmatrizen aller Merkmalsvariablen an

Kosinus Ähnlichkeit

Die Kosinus Ähnlichkeit bezeichnet man auch als Winkelabstand. Die Kosinusähnlichkeit wird traditionell und häufig im Text Mining (NLP) eingesetzt. Jedoch kann sie generell eingesetzt werden wenn hochdimensionale Daten vorliegen. Der Text wird dabei in einem Korpus abgelegt, der die vorkommenden Wörter und weitere Informationen wie ihre Frequenz (Häufigkeit) oder die Wortart enthält („POS tagging“). Die einzelnen Einträge im Korpus eines Dokuments werden als Attribute aufgefasst und können als Vektor dargestellt werden. Somit kann das Kosinus Winkelmaß zwischen den Vektoren bestimmt werden. Je geringer der Winkelabstand ist, desto ähnlicher sind sich die Texte oder Objekte. Bezogen auf den Iris Datensatz lässt sich der Unterschied und die unterschiedliche Aussagekraft zwischen einer metrischen Ähnlichkeit und der Kosinus Ähnlichkeit anschaulich demonstrieren:

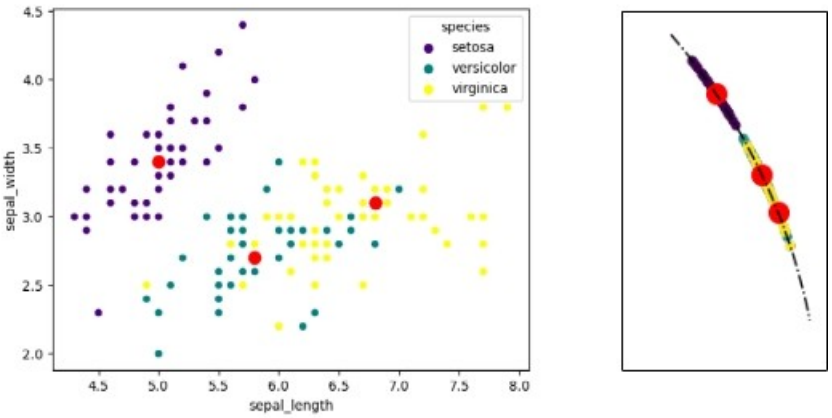

Abbildung 4: Euklidische vs. Kosinus Darstellung (eigene Darstellung)

Die Clusterung mit euklidischem Abstand zeigt, dass die Summen von Blattlänge und -breite bei den violetten und blaugrünen Punkten ähnlicher zueinander sind, als die Summe von Blattlänge und -breite der gelben Punkte. Das bedeutet, dass Flächeninhalt der Blütenblätter ähnlicher zwischen violetten und blaugrünen Punkten als zu den gelben Punkten ist[1]. Die Clusterung nach Kosinus-Ähnlichkeit sagt uns, dass das Verhältnis von Breite und Länge zwischen blaugrünen und gelben

1 Summe der Sepal-Länge und Breite, wenn die Werte logarithmiert sind ($\log_b(P \cdot Q) = \log_b(P) + \log_b(Q)$), alternativ ginge natürlich auch das Produkt, wenn die Werte nicht logarithmiert sind

Punkten im Allgemeinen enger ist als zwischen gelben und violetten. Dies zeigt, dass blaugrüne und gelbe Blüten wie eine vergrößerte Version der anderen aussehen, während violette Blüten im Vergleich eine breitere und kürzere Form haben [80]. Damit zeigt sich, dass der Winkelabstand, obwohl er heute vorrangig im NLP eingesetzt wird, auch in anderen Bereichen des Data Minings interessante Erkenntnisse liefern kann.

Aufgrund der Popularität und breiten Einsatzmöglichkeit wird das Kosinus Ähnlichkeitsmaß unterschiedlich definiert. In der ursprünglichen Form wurde die Größe der Vektoren nicht berücksichtigt, d.h. nur das Vorhandensein eines Merkmals, nicht jedoch seine Anzahl oder Stärke, oder das Verhältnis des Merkmals zur Größe des Gesamtkorpus. Dies stellt einen Nachteil dar, da lediglich die Richtung der Vektoren berücksichtigt wird, nicht jedoch deren Größe. In der Praxis bedeutet dies, dass die Werteunterschiede nicht vollständig berücksichtigt werden und bedeutet z. B. für ein Empfehlungssystem, dass die Kosinus-Ähnlichkeit nicht den Unterschied in der Bewertungsskala zwischen verschiedenen Benutzern oder die Gesamtanzahl der bewerteten Artikel berücksichtigt. Aus diesem Grund existieren Erweiterungen des Kosinus Ähnlichkeitsmaßes, welche die Häufigkeit des Auftreten eines Merkmals und als auch die Größe des Korpus berücksichtigen. Im Kontext der Analyse von Social Media Daten, können mit Hilfe des NLP und dem Kosinus Winkelmaßes ähnliche Texte vorklassifiziert werden und somit klassische Cluster-Analysen ermöglichen. Dabei profitieren die Algorithmen von der rechnerisch geringen Komplexität des Kosinus Winkelmaßes, insbesondere in spärlich besetzten Vektoren [81], [82].

$$S_C(X,Y) = \frac{X * Y}{[\![X]\!][\![Y]\!]} \qquad \text{Kosinus Ähnlichkeit}$$

$$D_C(X,Y) = 1 - S_c(X,Y) \qquad \text{Kosinus Distanz}$$

Relativ neu ist die Verwendung der normierten Winkel Distanz [83], [84], welche die Genauigkeit für kleine Winkelmaße verbessert, aber inhaltlich den Aussagen zur Kosinus Ähnlichkeit entspricht. Obwohl dadurch die Dreieckungleichung erfüllt wird, und es sich somit um eine echte Distanz Metrik handelt, ist sie aufgrund der höheren Rechenkomplexität bisher wenig verbreitet.

$$D_\theta = \frac{\arccos(S_C)}{(\pi)} \qquad \text{Winkel Distanz}$$

Kullback-Leibler-Divergenz

Die Kullback-Leibler-Divergenz, $D_{KL}(P \parallel Q)$ ist ein statistisches Abstandmaß. Sie gibt an, wie sich eine Wahrscheinlichkeitsverteilung Q von einer zweiten Wahrscheinlichkeitsverteilung P unterscheidet.

$$D_{KL} = \sum_{x \in X} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$ für diskrete Wahrscheinlichkeitsverteilungen

und $$D_{KL} = \int_{-\infty}^{\infty} P(x) \log\left(\frac{Q(x)}{P(x)}\right)$$ für kontinuierliche Wahrscheinlichkeitsverteilungen

Die KL Divergenz ist keine klassische Distanz, da sie die Dreiecksungleichung nicht erfüllt. Sie ist asymmetrisch, d.h. $D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P)$. Wenn die beiden Verteilungen Q,P identisch sind, ist $D_{KL} = 0$. Die KL Divergenz ist nicht normiert, d.h. ihre Größe sagt nur im Kontext des gleichen Datensatzes über die Größe der Unähnlichkeit aus [85, p. 6]. Trotz ihrer Schwächen ist die D_{KL} stark verbreitet im Data Mining, z.B. in der Multi-Instanz Klassifikation („information gain“) oder der Visualisierung hochdimensionaler Räume [86, p. 473 ff.], [87]. Die obigen Schwächen werden z.B. durch die Jennsen Shannon Distanz, welche auf der D_{KL} basiert, beseitigt.

$$D_{JS} = \sqrt{\sum p(x) \log \frac{2\,p(x)}{p(x)+q(x)} + q(x) \log\left(\frac{2\,q(x)}{p(x)+q(x)}\right)}$$

für diskrete Wahrscheinlichkeitsverteilungen (kontinuierliche Verteilungen analog)

Zentralitätsmaße

In den Untersuchungen sozialer Netzwerke werden häufig sogenannte Zentralitätsmaße („centrality measures“) z.B. zur Vorhersage der Einflußausbreitung (Kapitel 2.6.1) oder der Verbindungsvorhersage (Kapitel 2.6.4) verwendet. Diese messen zwar keine Ähnlichkeiten oder Distanzen, sondern geben einzelnen Knoten eine Gewichtung oder Rang, und ermöglichen so Knoten hinsichtlich ihrer Bedeutung in dem gegebenen Netzwerk zu beurteilen. Vergleichbar sind sie jedoch nur in ihrem gegebenen Kontext und bei vergleichbarer Datenbasis. Grundsätzlich basieren sie auf der Zählung von Schritten um von einem Knoten zu einem andern zu gelangen. Die Nutzung eines Zentralitätsmaßes wird stark durch den Kontext bestimmt. Sollen zentrale Knoten bestimmt werden, wird **closeness centrality** verwendet. Closeness centrality summiert die Entfernung von einem Knoten zu jedem anderen Knoten, somit haben zentralere Knoten relativ kleinere Werte. Die **stress centrality** misst die absolute Anzahl der kürzesten Pfade, die durch einen Knoten verlaufen, während die **betweenness centrality** den Anteil der kürzesten Pfade misst, die durch einen Knoten verlaufen [88].

2.6 Social Media Data Mining Forschungsgebiete

Verschiedene Untersuchungen beschäftigen sich mit der Frage, welche Data Mining Techniken bzw. Fragen im Kontext von Social Media Data relevant sind. In verschiedenen Publikationen wurden Social Media Plattformen untersucht und folgende Themen identifiziert [89]–[93].

1. influence propagation (Kapitel 2.6.1)
2. community/ group detection (Kapitel 2.6.2)
3. expert findings (Kapitel 2.6.3)
4. link prediction (Kapitel 2.6.4)
5. recommender systems (Kapitel 2.6.5)
6. predicting trust (Kapitel 2.6.6)
7. behaviour & mood analysis (Kapitel 2.6.7)
8. opinion mining (Kapitel 2.6.8)

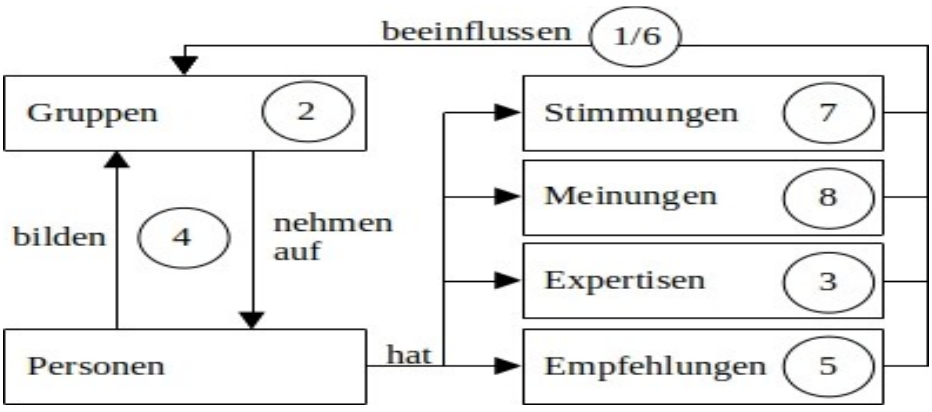

Abbildung 5: Forschungsgebiete des Data Mining in Social Media

Dabei sind die Themen nicht als abgeschlossene, unabhängige Forschungsgebiete zu verstehen, in vielen Gebieten gibt es Überschneidungen und gegenseitige Beeinflussungen.

2.6.1 Einflussausbreitung (Influence Propagation)

Soziale Netzwerke ermöglichen, das sich Personen austauschen und sich durch diesen Austausch in ihren Entscheidungen beeinflussen. Dies hat große Relevanz im Hinblick auf die Meinungsbildung dieser Gruppen (engl. „communities“) und kann für Marketingzwecke genutzt werden. Aber auch im Kontext der gegenwärtigen Pandemie liefert es wertvolle Erkenntnisse und ermöglicht z.B. die Identifikation von Gruppen und Themen um beispielsweise die Impfbereitschaft der Bevölkerung

zu steigern. Weiterhin können die Erkenntnisse der Einflussausbreitung genutzt werden, um andere Forschungsgebiete wie Recommender Systeme (Kapitel 2.6.5 Empfehlungssysteme (Recommender Systems)) oder die Erkennung von communities (Kapitel 2.6.2 Gemeinschaften- / Gruppenerkennung (Community / Group Detection) zu unterstützen und vice versa.

Die Einflussausbreitung (engl. „influence propagation“) basiert auf der Analyse der Verbindungen und Knoten. Diese formen ein Netzwerk aus Knoten (engl. „vertices“) und ihren Verbindungen (engl. „links/ edges“) zueinander. Dabei ist zwischen der Topologie des Netzes und den verschiedenen Aspekten (engl. „facets“) der Verbindung zu differenzieren. Die Aspekte die die Ausbreitung im Netzwerk beeinflussen, können zeitlicher Natur z.B. Häufigkeit des Kontakts, Art des Netzwerkes d.h. Anzahl und Dichte der Verbindungen oder aber auch soziale Merkmale z.B. private oder geschäftliche Kontakte sein.

Die Eigenschaften des Netzwerkes zu messen wird zunehmend durch verschiedene Zentralitätsmaße bestimmt (Kapitel „2.5 Ähnlichkeits- und Distanzmaße“). Die Beziehung zwischen zwei Knoten wird durch die Topologie d.h. die Anzahl, Art und Stärke der Verbindungen beeinflusst. Die Stärke der Beziehung wird bestimmt durch die Anzahl und Art der Knoten, die während der Diffusion betroffen sind. Ob und in welchem Umfang die Verbindung aktiviert wird, hängt wiederum stark vom Aspekt des Netzwerkes ab.

In einer aktuellen Untersuchung kommen die meisten der eingesetzten Methoden, in engem Zusammenhang mit der community detection, aus dem Bereich der unüberwachten Verfahren (z.B. k-means clustering, expectation maximisation clustering, Gaussian Mixture Model) oder Ensemble Methoden [94, p. 1084]. Dabei fokussiert sich die Influence Propagation stark auf die Evaluierung von Kennzahlen. Die Güte der Modelle wird dabei entweder gemessen in Bezug auf die Eigenschaften der Knoten (engl. „accuracy“) oder der Stärke der Verbindungen (engl. „quality“) [94, p. 1100].

Die Accuracy gibt dabei an, in wie weit der identifizierte Knoten korrekt der community zugeordnet wurde. Die Bestimmung der Accuracy setzt eine bekannte Grundwahrheit (engl. „ground truth“) über die korrekte Gruppenzugehörigkeit voraus.

Die Qualitätsmaße setzen keine bekannte Gruppenzugehörigkeit voraus, sondern messen die strukturelle Möglichkeit oder Wahrscheinlichkeit basierend auf der Anzahl und Stärke der Verbindungen zwischen den Knoten. Somit können Qualitätsmaße nur Aussagen zur generellen Wahrscheinlichkeit der Informationsausbreitung innerhalb des Netzwerkes treffen.

2.6.2 Gemeinschaften- / Gruppenerkennung (Community / Group Detection)

Group detection beschäftigt sich mit der Frage, welche Personen sich ähnlicher sind als andere und demzufolge Gruppen bilden. Dies entspricht wortwörtlich der Clusteranalyse im unsupervised learning. Die Gruppenbildung kann auf verschiedenen Ebenen oder Granularitäten erfolgen und ermöglicht es, Gruppengrößen in der realen Welt zu schätzen, oder verschiedene Untergruppen in Clustern zu identifizieren oder auch Überschneidungen sichtbar zu machen. Im Kontext der gegenwärtigen Pandemie ließe sich z.B. die Überscheidungen von „Querdenkern“ und rechts-extremen Gruppen untersuchen. Lancichinetti et al [95] und Leskovec [96] geben einen Überblick über die eingesetzten Methoden. Die Vorgehensweisen lassen sich in drei Gruppen unterteilen: ähnlichkeitsbasierte Verfahren, lernbasierte Verfahren und statistische Probabilitätsmodelle. Wesentlich für die Erkennung von Gruppen, sind nicht nur ihre Verbindungen, sondern auch der Kontext oder die Themen welche über diese links ausgetauscht werden. Beispielhaft für ein statistischen Verfahren ist das Author-Recipient-Topic Model (ART) [97], welches ein themen-basiertes Bayesian network als Grundlage für die Bestimmung der Gewichte der Kanten hat [97]. In einem ähnlichkeitsbasiertem Ansatz kombinieren Zhao et al [98] einen subspace clustering algorithm [99] und eine graphenbasierte Linkanalyse. Mit Hilfe des Clusteralgorithmus wird identifiziert, welche Themencluster grundsätzlich existieren, basierend auf einer Ähnlichkeits-analyse. In nächsten Schritt werden die Individuen den Themenclustern zugeordnet. Abschließend wird eine Link Analyse zur Identifikation der sozialen Gruppierungen durchgeführt.

Gasemi et al verfolgen Ansätze learn- und ähnlichkeitsbasierte Verfahren miteinander zu kombinieren [100]. Dabei werden zuerst relevante Attribute (*features*) der Knoten gelernt, um nachfolgend basierend auf den gewonnen Erkenntnissen mit Hilfe eines subspace clusterings die sozialen Gruppen zu identifizieren.

2.6.3 Expertenanalyse (Expert Findings)

Unter Expert Findings versteht man, wenn versucht wird, Experten in sozialen Netzwerken zu identifizieren. Experten können dabei als Personen aufgefasst werden, die ein tieferes Verständnis eines Themas haben. Verbunden damit sind auch Verfahren, welche das mögliche Themen innerhalb eines Netzwerks *a priori* identifiziert sollen. Als Experten können aber auch technische Systeme verstanden werden, die spezialisierte Dienste oder Dokumente in Suchmaschinen zur Verfügung stellen [101]. Beispielsweise wäre es denkbar auf Plattformen wie „stackoverflow.com“ Personen mit einem tiefen Verständnis einer Programmiersprache zu identifizieren und diese dann hinsichtlich der weiteren Entwicklung der Sprache zu befragen.

Sollen Experten und ihr Rang innerhalb einer definierte Gruppen bestimmt werden, wird häufig auf die Graphentheorie und die Bestimmung via Zentralitätsmaße zurückgegriffen. Zentrale Knoten sind bestimmend für diese Gruppen und können als Experten innerhalb dieser Gruppe identifiziert und anschließend Nutzern vorgeschlagen und durch diese befragt werden. Somit kann den Experten ein Wissensniveau oder ein Rang innerhalb des Netzwerks zugewiesen werden.

Eine weitere Gruppe von Verfahren basiert auf der Analyse der Beiträge verschiedener Nutzer um das wahrscheinliche Thema zu identifizieren. Die *probabilistic latent semantic analysis* (PLSA), untersucht welche Wörter in den Dokumenten gleichzeitig vorkommen. Das Thema ist dabei latent, d.h. es wird nicht explizit erwähnt. Basierend auf den Wahrscheinlichkeiten, wird dann ein Wort für das Thema aus dem Korpus vorgeschlagen [102], [103]. Die Identifikation von Experten kann anschließend nach dem obigen Verfahren innerhalb der Themengruppe erfolgen

Eine andere Verfahrensgruppe beschäftigt sich mit der Optimierung von *peer-to-peer* Netzwerken (P2P). Diese Netzwerke ermöglichen den direkten Austausch von Medien zwischen zwei oder mehreren Nutzern (*peers*). Dabei werden die Medien nicht zwingend von einer Quelle geladen, sondern Medienteile von von verschiedenen Quellen. Zur Identifikation der „besten“ Quelle (aka „Experten“) wird analysiert, welche Teile am seltensten im Netzwerk vorhanden sind, und diese werden bevorzugt geladen. Als Experte wird somit die Quelle bezeichnet, die die seltensten und damit speziellsten Daten zur Verfügung stellen kann. Damit wird die Ausbreitung der Medien auf möglichst viele peers beschleunigt, da die verfügbaren Datenblöcke gleichmäßig zur Verfügung stehen [104].

2.6.4 Verbindungsvorhersage (Link Prediction)

Link prediction beschäftigt sich mit dem Aufbau und Abbau von Netzwerken und ihren Verbindungen. Durch die Analyse von Informationen über einen Knoten/ Nutzer lässt sich ableiten,

welche Interaktionen in der näheren Zukunft wahrscheinlich sind oder fehlen. Damit lassen sich z.B. Aussagen ableiten, wie Gruppen wachsen oder sich anderweitig verändern. Dies kann genutzt werden, um Gruppenveränderungen zu antizipieren oder aber auch Nutzern sozialer Netzwerke mögliche Verbindungen zu anderen für sie interessanten Nutzern vorzuschlagen [105]–[107].

Die Verbindungsvorhersage-Algorithmen werden im Allgemeinen in unüberwachte und überwachte Verfahren kategorisiert. Ähnlichkeitsmessungen basieren grundsätzlich auf unüberwachten Methoden. Überwachte Methoden betrachten Verbindungsvorhersagen als Klassifikationsprobleme. Letztere werden aktuell angewendet und liefern in der Regel die genaueren Ergebnisse [100]. Dabei steht man vor zwei grundsätzlichen Herausforderungen. Soziale Netzwerke sind zum einen oft nicht vollständig abgebildet oder erfasst, da nur ein Teil der Informationen in den sozialen Netzwerken überhaupt zur Verfügung steht und damit gesammelt werden können. Die zweite Herausforderung ist der zeitliche Kontext in dem die Verbindungen bestehen. Der zeitliche Kontext ergibt sich nicht nur aus der zeitlichen Nähe der letzten Aktion eines Nutzers, sondern auch aus dem Lebensalter der Nutzer. Die Nutzung der einzelnen Medien und deren Frequenz ist entscheidend durch das Lebensalter geprägt [108]. Speziell für die Vorhersage von Beziehungen zwischen älteren Nutzern, demonstrieren Kamoun at al an einem Prototypen [109], dass die Einbeziehung von geo-temporalen Informationen die Vorhersagequalität positiv beeinflusst.

2.6.5 Empfehlungssysteme (Recommender Systems)

Empfehlungssysteme ermöglichen es den Nutzern neue Inhalte und Dienste vorzuschlagen. Dies basiert auf Ähnlichkeiten zu Suchanfragen oder verwandten Aktivitäten und basiert grundsätzlich auf der Analyse von Homonymen, Synonymen, Konnotationen sowie Präferenzen von als ähnlich identifizierten Nutzern. Dabei werden Informationen über die Präferenzen ihrer Benutzer für eine Reihe von Elementen (z.B. Filme, Lieder, Bücher, Anwendungen, Websites, Reiseziele etc.) gesammelt. Die Informationen können explizit (in der Regel durch das Sammeln von Benutzerbewertungen) oder implizit (in der Regel durch die Überwachung des Benutzerverhaltens) erfasst werden. Diese Informationen können durch demografische Daten des Nutzers wie z.B. Alter und Geschlecht, sozialen Informationen wie z.B. followers und posts in sozialen Medien oder ortsbasierte Informationen wie GPS Koordinaten erweitert werden. Dabei wird versucht die Balance zwischen genauen Vorschlägen, ähnlichen Vorschlägen und neuartigen Vorschlägen zu finden [78]. Grundsätzlich existieren zwei Ansätze zur Filterung der der relevanten Vorschläge: Kollaboratives (*collaborative) und* inhaltsbasierte (*content-based)* filtern.

Inhaltsbasierte Filterung gibt Empfehlungen basierend auf Benutzerentscheidungen, die in der Vergangenheit getroffen wurden. Dies können Artikel sein die ein Benutzer gekauft, gehört, angesehen und positiv bewertet hat. Die inhaltsbasierte Filterung generiert außerdem Empfehlungen unter Verwendung des Inhalts von Objekten für Empfehlung. Dabei können bestimmte Inhalte analysiert werden, wie z.B. Text, Bild und Ton. Die Ähnlichkeit zwischen Objekten dient als Grundlage für die Empfehlung von Produktion, Seiten etc. [110]. Inhaltsbasierte Filterung hat zwei grundsätzliche Herausforderungen: die sichere automatische Erkennung von Ähnlichkeiten und die Gefahr nur sehr ähnliche Artikel vorzuschlagen, jedoch keine neuartigen [78].

Kollaboratives Filtern gibt Empfehlungen basierend auf Daten von Benutzern mit ähnlichem Interessen. Die grundlegende Annahme ist, dass ähnliche Nutzer ähnliche Interessen an ähnlichen Artikeln haben. Die Korrelation zwischen den Nutzern wird in der Regel anhand von gemeinsamen bewerteten items gemessen (benutzerbasierte kollaboratives Filtern), kann jedoch auch durch soziale Netzwerke, im Internet frei verfügbare verknüpfte Daten (linked open data) oder geografische Informationen bestimmt oder verbessert werden [71]–[74]. Die Genauigkeit von Systemen auf der Grundlage von kollaborativer Filterung hängt von der Bereitschaft des Benutzers ab, Bewertungen abzugeben. Die Bereitstellung von Bewertungen erfordert zusätzlichen Aufwand den Benutzer und Benutzer neigen dazu dies zu vermeiden. Dies verursacht zwei Arten von Problemen. Datensparsität, d.h. für viele Artikel gibt es keine Bewertungen und das Kaltstartproblem, d.h. neue Artikel oder User haben zu Beginn noch keine Bewertungen [111]. Gegenwärtig werden die verschiedenen Ansätze zu hybriden Verfahren kombiniert.

2.6.6 Vertrauensanalyse (Predicticting Trust/ Distrust Among Individuals)

Durch die Verlagerung sozialer Kontakte aus der realen Welt in die Sozialen Medien, ist die Relevanz von Vertrauen und Misstrauen innerhalb von Gruppen in den sozialen Medien gestiegen. Bisherige persönliche Kontakte werden durch virtuelle Kontakte ersetzt und erlernte soziale Muster funktionieren nicht. Auch wenn die persönliche Einstellung zu Vertrauen und Misstrauen gleich bleibt, ist jedoch das strukturelle Vertrauen grundsätzlich anders zu evaluieren [112]. Die Arbeiten fokussieren sich auf die Vorhersage von Vertrauen in ungerichteten Netzwerken (*unsigned networks/ interaction based models*) und gerichteten Netzwerken (*signed networks/ graph based models*). In gerichteten Netzwerken wird das Vertrauen der Nutzer hinsichtlich eines direkten Nachbarschaftsknoten durch drei Graphentypen abgebildet: Vertrauen (*trust)*, Misstrauen (*distrust)* und beziehungslos (*„no relation“)* angegeben. Ziel ist es, über Ketten das Vertrauen/ Misstrauen zwischen zwei Knoten vorherzusagen, die in keinem direkten Zusammenhang stehen. In

ungerichteten Netzwerken ist dies nicht der Fall und muss aus dem Kontext oder via Kennzahlen abgeleitet werden. Opuszko et al nutzt z.B. Aktualisierungsraten von Wikipedia Artikeln um darauf aufbauend Markov Ketten zu generieren und darauf aufbauend die Qualität der Artikel und so indirekt das Vertrauen abzubilden [113]. Das Vertrauen/ Misstrauen ist dabei häufig binär abgebildet (1/0) und entspricht damit nicht der realen Welt. Um dies realitätsnäher zu modellieren, versuchen die verschiedenen Modelle den zeitlichen Kontext oder den inhaltlichen Kontext in ihre Modelle zu integrieren und dementsprechend Ränge oder Gewichte einfließen zu lassen. [114]–[116]. Die Mehrzahl der Arbeiten beschäftigt sich mit gerichteten Netzwerken [117]–[123]. Jedoch scheint das Forschungsfeld insgesamt etwas an Dynamik verloren zu haben. Ghafari et al führen aus, dass zur Validierung der Modelle wenig Benchmark Datensätze zur Verfügung stehen. Ein oft referenzierter Datensatz basierend auf Epinions wird nicht mehr aktualisiert. Auf epinions.com konnten Nutzer Produkte beurteilen und die Verfasser solcher Beurteilungen hinsichtlich ihrer Vertrauenswürdigkeit bewerten. Die Plattform stellte die aktive Beurteilung 2014 ein und schloss die Seite 2018 endgültig [124]. Ein weiteres Problem stellt die Komplexität der Berechnung dar, welche oftmals eine nicht-akademische Anwendung verhindert [125].

2.6.7 Verhaltens und Stimmungsanalyse (Behaviour And Mood Analysis)

Eines der Ziele von Verhaltens- und Stimmungsanalyse (*Behaviour and Mood Analysis*) ist es, festzustellen, wie sich die Nutzer sozialer Netzwerke gegenseitig beeinflussen. D.h. alle Analysen, welche untersuchen, wie ein einzelnes Individuum auf präsentierte Inhalte reagiert (z.B. durch das Klicken auf Werbung oder der Art wie Diskussionen im Internet geführt werden). Im Rahmen der Pandemie z.B. wurde festgestellt, dass die größere Anonymität der sozialen Medien zu verstärkten Beschimpfungen führt [126]. Darüber hinaus versuchen diese Modelle, aufkommende Trends zu erfassen, welche durch die Interaktionen und Entscheidungen einzelner Nutzer verursacht werden. Chameley et al [127] beschreiben 2013, basierend auf einen Gauß-Modell zwei Ansätze. Unter der Annahme, dass jede Meinung/ Einschätzung zu einem Thema grundsätzlich richtig und nur durch einen individuellen Bias verzehrt ist, führt eine große Anzahl von Meinungen zu einem Thema dazu, dass die Varianz des Modells hinsichtlich eines Konsens oder Erwartungswert abnimmt. Dementsprechend nimmt die Unsicherheit für die Vorhersage des Verhaltens ab und wird schlussendlich den wahren Wert annehmen. Dies entspricht dem Ansatz des wissenschaftlichen Realismus [128]. Im zweiten Modell werden die Annahmen dahingehend geändert, das die individuellen Einschätzungen aus einem Pool mit endlichen Möglichkeiten kommen. Jede individuelle Meinung, die vom wahren Wert abweicht, erhöht die Varianz des Modells und damit

die Unsicherheit der Individuen [127]. In aktuellen Beiträgen wird, basierend auf dem Konzept der agent-based-modeling (ABM), um die Gewichte der Verbindungen zu bestimmen jeweils die individuellen Eigenschaften der Knoten (aka Teilnehmer) wie Sturheit d.h. dem Beharrungsdrang des Individuums und den speziellen Kontext berücksichtigt. Damit wird versucht die Beeinflussung durch andere Teilnehmer oder die Kosten der Informationsbeschaffung abzubilden und damit die gegenseitige Beeinflussung zu simulieren. Die Simulationen skaliert grundsätzlich jedoch schlecht und stößt bei Analysen großer sozialer Netzwerke an ihre praktischen Grenzen. Wenn die Anzahl der Benutzer groß wird, wird die Berechnung der Lösung aufgrund der erhöhten Wechselwirkungen und der Rückkopplungen zwischen den Benutzern sehr komplex. Ein aktueller, interessanter Ansatz ist die Einführung der „Mean Field Game Theorie“. Dies basiert auf der „Mean Field Theorie“ der Physik, mit der die Wirkung von vielen Teilchen auf Einzelne analysiert wird. Unter der Annahme einer große Anzahl kaum zu unterscheidender Teilnehmer, wird eher die Gesamtwirkung aller Teilnehmer als die individuelle Wirkung betrachtet [129]–[131].

2.6.8 Meinungsanalyse (Opinion Mining)

Soziale Netzwerke ersetzen oder erweitern die persönlichen direkten Begegnungen. Viele soziale Netzwerke werden genutzt, um die individuellen Meinungen einzelner Teilnehmer zu kommunizieren und sich untereinander auszutauschen. Diskussionen, Beiträge und der Austausch von Ideen einzelner einzelner Nutzer ermöglichen die Untersuchungen von Bedürfnissen und Erwartungen und ihre Reaktion auf Veränderungen. In den Analysen geht es nicht allein um die Untersuchungen von Meinungen zu wichtigen gesellschaftlichen Ereignissen, sondern auch um die demografische Verteilen dieser Meinungen, wie sich die Meinungen im Laufe der Zeit entwickeln, ob sich Meinungsführer identifizieren lassen und welchen Einfluss diese haben. Dies ist nicht nur aus wissenschaftlicher Sicht interessant, sondern natürlich auch aus wirtschaftlicher Sicht .

Beiträge die in Textform vorliegen, lassen sich in Stimmungsanalysen (engl. „sentiment analysis“ untersuchen. Dies ist nicht neu und ist ein Grundbestandteil jeder Sprachwissenschaft, jedoch erlauben die digital vorliegenden Daten und hochperformante Computer die Automation unter Verwendung von Data Mining Techniken. Die manuelle Vorklassifizierung oder Stimmungsanalyse unterliegt dabei oft einem Bias durch die individuell subjektiven Erfahrungen und Einstellungen der Analysten, insbesondere wenn eine Vielzahl an Dokumenten oder Beiträgen untersucht werden muss [132], [133]. Dabei können traditionelle Techniken eingesetzt werden, die auf Bewertung eines initial gebildeten Lexikons basieren, können aber auch durch Methoden des Clusterings unterstützt werden [81]. Die Vorteile der automatisierten Analyse sind offensichtlich, jedoch bleiben

viele Herausforderungen bestehen. Die Sentiments eines Wortes können sich über den Zeitverlauf ändern oder werden in sozialen Gruppen unterschiedlich verwendet. Der Begriff „gay“ bedeutete vor 200 Jahren „fröhlich“, wird heute jedoch mit Homosexualität gleichgesetzt und kann abhängig von der sozialen Gruppe positiv oder negativ besetzt sein [134]. Die Relevanz einzelner Beiträge in Bezug auf den Gesamttext kann problematisch sein, beispielsweise wenn sich innerhalb eines an sich spezifischen, geschlossenen Kontext Unterthemen entwickeln, die vom ursprünglichen Kontext abweichen. Negationen erfordern umfangreiche Modellierungen, Ironie lässt sich ohne die Kenntnis des speziellen Kontextes nicht automatisch erkennen und kann zu Verfälschungen der Ergebnisse führen [135].

2.7 Resümee

Betrachtet man die Gebiete der Social Media Daten und deren Nutzung aus der wissenschaftlichen Perspektive, lässt sich feststellten, dass aufgrund der großen Bandbreite der Thematik viele interessante Anwendungsmöglichkeiten für die Nutzung der Daten bestehen. Viele der genutzten Techniken und Verfahren basieren auf Verfahren die auch bei „klassischen“ Daten genutzt werden. Es besteht jedoch die Notwendigkeit die Verfahren weiterzuentwickeln bzw. zu adaptieren, um speziell auf die Charakteristiken der Social Media Daten (z.B. Größe, Unvollständigkeit, Datenqualität) eingehen zu können. Insbesondere der große Datenumfang, die Aktualität sowie die große Nähe zu den Nutzern als Datenersteller und Nutzer, eröffnen interessante Möglichkeiten. Damit lassen sich neue Anwendungsmöglichkeiten erschließen, welche mit klassisch gewonnenen Daten nur aufwendig durchgeführt werden könnten (z.B. Gruppenidentifikation und Analyse, Stimmungsanalyse, Empfehlungssysteme). So ermöglichen Stimmungsanalysen der Politik und der Wirtschaft zeitnah und kostengünstig auf die Ersteller einzugehen, Empfehlungssysteme bilden das Rückgrat des e-Commerce, Gruppenidentifikation erlauben neben zielgenauer Ansprache von Kunden auch Soziale Netzwerke.

Welche der vorgestellten Verfahren sich in der relativ neuen Domäne praktikabel erweisen und durchsetzen, lässt sich Stand heute noch nicht abschätzen. Jedoch zeigen Anzahl und Art der Veröffentlichungen ein umfangreiches Interesse der Wissenschaft an diesen Themen. Somit muss man Frau Dr. Merkel zustimmen, das auch 50 Jahre nach Einführung des Internets und 30 Jahre nach Einführung Web 2.0 und damit der Sozialen Medien wir uns noch immer im Neuland befinden.

3 Publikationen

Die einzelnen publizierten Arbeiten versuchen das Thema Qualität und die Nutzung von Social Media Daten aus verschiedenen Perspektiven zu beleuchten. Die empirische Sozialwissenschaft basiert auf einer Kette aus Deduktion, Empirie, Induktion und Theoriebildung. Dabei ist Deduktion der „Schluss vom Allgemeinen zum Speziellen“, oder die Ableitungen von allgemeingültigen Aussagen aus einer Theorie, welche dann durch Messungen oder Beobachtungen an konkreten, praktischen Daten validiert werden können. Umgekehrt können durch Induktion oder Schlussfolgerung aus konkreten, realen Daten (Empirie) neue Schlüsse auf allgemeingültige Gesetze oder Theorien erfolgen.

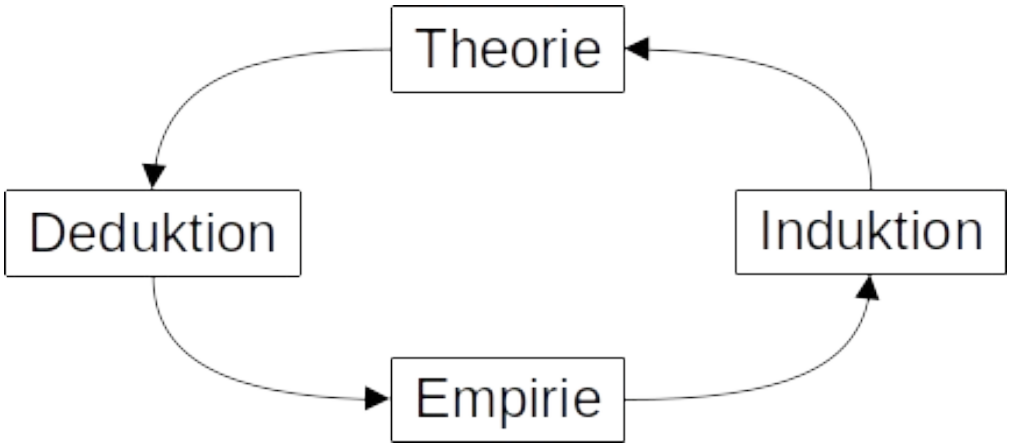

Abbildung 6: Deduktion und Induktion als Verbindung von Theorie und Praxis

Die einzelnen Publikationen folgen dem oben dargestellten Kreislauf. In der ersten Ausarbeitung wird gezeigt, dass von der Theorie kommend, Aussagen durch empirisch erhobenen Social Media Daten bestätigt werden können. Damit wird es möglich, sozialwissenschaftliche Theorien zeitnah und kostengünstig zu validieren. In der zweiten Arbeit wird eine bestehende empirisch, induktive Studie, welche auf Panel Daten beruhte, durch eine induktive Studie auf Social Media Daten bestätigt und die unterschiedliche Aussagekraft der beiden Datenquellen thematisiert. Nachdem der gesamte Kreis der Erkenntnisgewinnung positiv unter Verwendung von Social Media Daten demonstriert werden konnte, wendet sich die dritte Ausarbeitung den Vertrauen beeinflussenden Faktoren in das Data Mining aus Sicht der wichtigsten Anspruchsgruppen (Stakeholdern) zu. Dazu wurden mehrere Interviews mit den einzelnen Stakeholdern durchgeführt und qualitativ analysiert. In der abschließenden Untersuchung wird eine mögliche praktische Anwendbarkeit der bisher

hauptsächlich wissenschaftlich, theoretischen Ausarbeitungen demonstriert und damit die Praxisrelevanz im Rahmen der betriebswirtschaftlichen Entscheidungsfindung aufgezeigt.

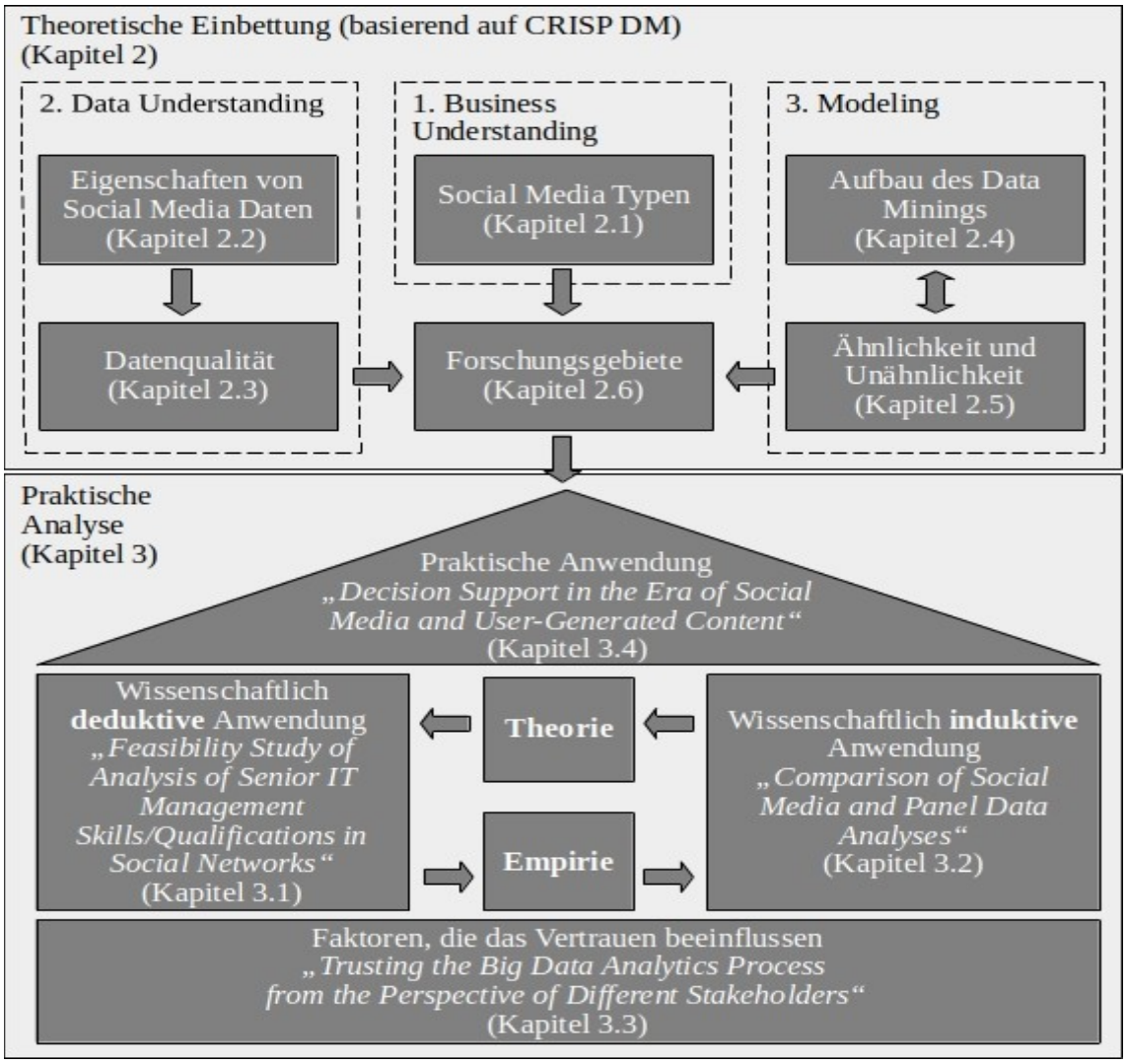

Abbildung 7: Aufbau der Arbeit (Ausschnitt des Zusammenhanges der praktische Analyse)

Die im praktischen Teil beschriebenen Untersuchungen greifen die im theoretischen Teil beschriebenen Techniken auf.

In der ersten Veröffentlichung „Feasibility Study of Analysis of Senior IT Management Skills/Qualifications in Social Networks“ (Kapitel 3.1) lag eine besondere Herausforderung in Analyse der Datenqualität (CRISP DM Phase: „Data Understanding“) und der Datenaufbereitung (CRISP DM Phase: „Data Preparation“). Aufgrund der für Social Media Daten typischen nicht-relationalen Daten (Kapitel 2.2), mussten ähnliche Ausprägungen eines Features (dt. „Attribut“) mit Hilfe von Verfahren des NLP (Kapitel 2.5) aufwendig auf Schlüsselbegriffe verdichtet werden.

Erstaunlicherweise war die Varianz der Schlüsselbegriffe gemessen an der Gesamtanzahl der Werte relativ gering. Dies spricht für eine hohe Qualität der zugrundeliegenden Daten für diese Analyse (Kapitel 2.3). Um die unterschiedlichen Qualifikationen weiter sinnvoll verdichten zu können, wurden die Handlungsfelder des Cobit Frameworks herangezogen. Die eigentliche Modellierung nutzt die klassische hierarchische Clusteranalyse (Kapitel 2.4.2) (CRISP DM Phase: „Modeling“).

In der Publikation „Comparison of Social Media and Panel Data Analyses“ lag ebenfalls eine besondere Herausforderung im Verständnis der Daten und deren Aufbereitung (CRISP DM Phasen: „Data Understanding/ Data Preparation“. Hier zeigte sich, das ein relativ großer Prozentsatz von Datensätzen verworfen werden musste, da der aktuelle Wohnort nicht sichtbar war. Zur Verdichtung der Einzelausprägungen des Standortes mussten öffentlich verfügbare PLZ Daten mit den Social Media Daten verknüpft werden. Hier war die Datenqualität erstaunlich hoch, relativ wenige Standorte mussten manuell nachbearbeitet werden (Kapitel 2.3). Bei der Analyse der Entfernung vom Studienort war die Frage des angemessensten Abstandsmaßes schwierig (Luftlinie vs. „KFZ Distanz vs. Anbindung an den ÖPNV“ – dieses Problem teilen sich jedoch sowohl die Panel als auch die Social Media Daten. Für den Vergleich der Verteilungen wurde wiederum ein klassisches Verfahren der Verteilungsanalyse genutzt (Kullback Leibler Divergenz).

Beide Veröffentlichungen demonstrieren, das die Modellierung mit Hilfe von klassischen Ansätzen durchaus zu Ergebnissen führt. Es zeigte sich aber, das im Vorfeld die Analyse und Aufbereitung der Daten und die Verknüpfung zahlreicher Techniken zwingend notwendig ist. Insgesamt waren für die Untersuchungen die Verknüpfung zahlreicher Techniken aus verschiedenen Wissenschafts-disziplinen notwendig. Dementsprechend ist zum Nachvollziehen der Untersuchung entweder ein breites Fachwissen notwendig oder das Vertrauen in den zugrundeliegenden Prozess (hier CRISP DM). Dieser Fragestellung widmet sich die Publikation „Trusting Big Data Analytics Process from the Perspective of Different Stakeholders“. Es zeigte sich, das Vertrauen interdisziplinär unterschiedlich definiert wird und in den Data Mining Frameworks keine prominente/ explizite Rolle spielt, obwohl Befragungen unterschiedlichster Stakeholder zeigten, das dies durchaus implizit während des Data Mining Prozesses problematisiert wird.

Während sich die ersten drei Arbeiten stark auf die wissenschaftlichen Themen konzentrierte, wird in der finalen Publikation „Decision Support in the Era of Social Media and User Generated Content“ an Einzelbeispielen versucht, die Relevanz der Nutzung von Social Media Daten aus wirtschaftlicher Sicht aufzuzeigen.

3.1 Publikation 1: „Feasibility Study of Analysis of Senior IT Management Skills/Qualifications in Social Networks“

Die dieser Publikation zugrundeliegende Fragestellung lautet: „Kann eine bestehende Theorie durch die Analyse von empirisch erhobenen Social Media Daten bestätigt werden?“. Wenn sich dies ohne Vielzahl von Vorbedingungen und Einschränkungen erreichen lässt, sollte untersucht werden, welche weiteren Erkenntnisse sich dann auf Basis der erhobenen Daten ableiten lassen. Eine bekannte Tatsache ist es, dass sich Arbeitnehmer ihrer beruflichen Tätigkeit entweder spezialisieren oder generalisieren – sie werden fachlich zu „Spezialisten“ oder „Generalisten“ . In dem bekannten Karrierenetzwerk XING haben die Mitglieder die Möglichkeit ihre beruflichen Qualifikationen öffentlich darzustellen und dieses Netzwerk eignet sich als Ausgangsbasis der Untersuchung. Diese Qualifikationen können sehr vielfältig sein, granular und die Eingabemöglichkeiten sind frei und lassen individuelle Beschreibungen zu. Um die Informationen methodisch und konsistent zu verdichten, wurden die häufigsten Qualifikationen mit Hilfe der COBIT Enabler (Handlungsfelder) verdichtet. Auf dem entstandenen Artefakt wurde nachfolgend eine hierarchische Clusteranalyse (siehe Kapitel 2.4.2) durchgeführt. Die beiden mit Abstand größten Cluster wiesen die erwarteten Ausprägungen auf: die Personen im „Generalisten Cluster“ waren in fast allen Handlungsfeldern beschlagen, die Personen im „Spezialisten“ Cluster konzentrierten sich auf wenige Handlungsfelder, dafür jedoch intensiv. Somit konnte die Ausgangstheorie bestätigt werden. Nach der Bestätigung der grundsätzlichen Anwendbarkeit der XING Daten ließen sich nachfolgend weitere Analysen durchführen.

Neben der Beantwortung der Forschungsfrage zeigte die Bearbeitung Herausforderungen, Vorteile und Risiken bei der Analyse von Social Media Datenquellen auf.

- Die Verdichtung der Qualifikationen mit Hilfe der COBIT Enabler ist nicht eindeutig. Jede Qualifikation konnte mehreren Enablern zugeordnet werden konnte z.B. zeugt die Qualifikation „MS SQL Server“ zum einen von Kenntnissen auf dem Gebiet „Service, Infrastructure and Applikations“ – die Fähigkeit die Software zu bedienen - zum anderen von Kenntnissen auf dem Gebiet „Informationen“ – Datenbanken dienen dazu Informationen zu speichern und widerspruchsfrei zu organisieren. Die subjektive Sicht auf die Zuordnung zu den COBIT Enablern konnte jedoch durch unabhängige Expertenreviews beseitigt werden.
- Bei der Analyse der Arbeitgeber zeigte sich, dass das angegebene Unternehmen prinzipiell einer bestehenden Unternehmen zugeordnet werden konnte, jedoch ist die Granularität der

Einträge sehr stark vom Nutzer abhängig. Am Beispiel des Unternehmens Siemens wurden unter anderem Einträge gefunden wie:

- „Siemens“, „Siemens AG“ - Mutterkonzern,
- „Siemens Med“ - Unternehmensteil (alter Namen),
- „Siemens Healthineers“ - Unternehmensteil (neuer Namen).

Dieses Beispiel zeigt das eine fein-granulare Zuordnung problematisch und nicht automatisierbar ist.

- Neben der generischen Aussage, dass Spezialisten und Generalisten existieren, ließ sich eindeutig ein Zusammenhang zur Unternehmensgröße des Arbeitgebers finden. Das Auftauchen der beiden anderen Cluster „Firmenkulturorientiert“ und „Organisationsorientiert“ war ursprünglich nicht erwartet worden und zeigt, dass neue Einsichten aus der Analyse der Sozialen Daten gewonnen werden können.

Das paper wurde mit dem „Best Paper Award“ der Veröffentlichung bedacht [136].

Feasibility Study of “Analysis of Senior IT Management Skills/ Qualifications in Social Networks”

Sven Gehrke, Dr. Lisa Wenige, Prof. Johannes Ruhland
Chair of Business Informatics, School of Economics and Business Administration, Jena, Germany
sven.gehrke@uni-jena.de
sandra.niemz@uni-jena.de
lisa.wenige@uni-jena.de
johannes.ruhland@uni-jena.de

Abstract Digitalization leads to growing complexity and tight coupling of business processes. This increases the risk of ‘natural accidents’ (Perrow, 1999). An answer to this proneness is a holistic mindfulness culture (Roberts, 1989; Weick and Putnam, 2006). The mindfulness dimension requires a holistic approach, because it includes multiple aspects. This mindfulness should be reflected in the self-assessment of company's employees. It is attempted to clarify whether social media allow an insight into how self-proclaimed qualifications are distributed in current enterprises. Social networks are particularly suitable, because the member try to reflect the perceived demands of the companies - while the characteristics presented are „quality assured” by connected members. More than 3500 profiles of senior IT managers were examined in the social network XING. COBIT has been used for categorization as a proven holistic approach for IT governance and structuring tool. The study shows that executives can be primarily categorized into generalists or specialists, both embracing a holistic perception, and therefore social media are suitable for general studies of qualifications in a business environment.

Keywords: qualification distribution, social media, COBIT, senior IT management

1 Motivation and Related Work

The influence of employee and company qualification on the control of business risks has been studied starting with the Natural Accident Theory (Perrow, 1999) and building on “high reliability organizations” (HRO) (La Porte, 1996; Roberts, 1989) and Mindfulness (Weick et al., 2008). Empirical studies confirm these theories (Smith and Ridoutt, 2007). The practical importance from an IT governance point of view can also be seen from the fact that in the generally accepted IT governance framework COBIT “employee qualification” assigns its own classification dimension (People, Skills and Competencies). It seems that Companies' focus on soft skills to handle uncertainty and problem solving skills (Ridoutt et al., 2002). On the other hand, interest in formal and technical qualifications or systematic analysis seems low and is often driven by compliance. The distribution of qualifications seems to depend on factors such as the size of the company (Ridoutt et al., 2005) and the innovativeness of the companies (Long and Fischer, 2002). Scientific research indicates that general management skills are gaining in importance (Murphy and Zabojnik, 2004) and that specialised executives have little bargaining power in the laboratory market (Custódio et al., 2013) and on the other hand general qualifications facilitate job changes (Giannetti, 2011). All that leads to an increase in interest in general qualifications on the employee side (Baruch et al., 2005). However these previous surveys are either not performed in repetition due to effort, are only based on relatively small samples (n = 446 (Ridoutt et al., 2002); n = 318 (Baruch et al., 2005)) or concentrate only on top level executives (Custódio et al., 2013). The hypothesis of this work is that social media can give detailed insight into the distribution of qualifications on different organisational level with an increased number of samples. It is expected that previous findings on qualification levels can be confirmed and thereby showing the relevance of this analysis method.

For identification of qualification profiles there are numerous approaches. In our analysis, we tried to build on an already common and time proven approach, and choose dimensions established there as a classification scheme. The COBIT framework was finally selected due to its worldwide use and holistic approach. Because of this it has been often applied by auditing firms (Gaulke, 2014; Lainhart, 2000) and recommended from official sides (Information Systems Audit and Control Association, 2014; “Summary of risk methods and frameworks - NCSC Site,” n.d.). Other frameworks like IT Infrastructure Library framework (ITIL) which is more focused on IT operations does not offer classification schemes. On the other hand the framework of Committee of Sponsoring Organizations of the Treadway Commission (COSO) is too strategic and therefore not applicable for surveys on the micro (aka individual) level. By having links to ISO 9000, COBIT is even influenced by non-IT governance issues. (**Figure 1**)

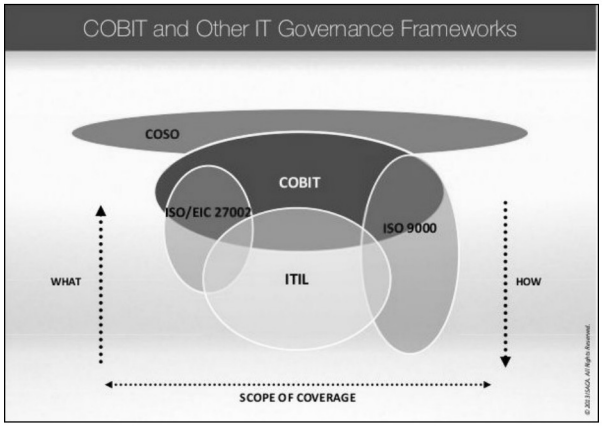

Figure 1 COBIT’s scope and coverage within the generic ecosystem of frameworks and standards

COBIT defines seven areas of action (called “enabler”) which should be looked at if you want to have a holistic approach (Information Systems Audit and Control Association, 2012, pp. 65–88). Their importance has been confirmed by security surveys, where IT security experts identified critical security factors which can be related to COBIT enablers (Hille and Schnedermann, 2017, p. 47).

1. Principles, Policies and Frameworks:
   Concerned with compliance with external and internal regulations and standards or governance specifications i.e. laws and ethical standards.
2. Processes:
   Concerned with structured and standardised approaches for accomplish targets.
3. Organisational Structures:
   Concerned with roles, working organisation and principles, extend and manner of authority.
4. Culture, Ethics, Behaviour:
   Concerned with individual and collective behaviour aka company culture.
5. Information:
   Concerned with intrinsic quality, context sensitive quality, security and availability (security targets) of produced and used information.
6. Services, Infrastructure and Applications.
   Concerned with technical resources and services used to accomplish the IT and business target
7. People, Skills and Competencies
   Concerned with Skills and Competencies like education and qualification levels, technical skills, experience, knowledge and behaviour.(Information Systems Audit and Control Association, 2012)

It can be seen that the COBIT standard does not only comprise technological aspects (as shown in list item 6), but also considers so called “soft areas”, such as cultural factors (list item 4) or qualification of employees (list item 7) since they can play a vital role in solving acute crisis situations. By this means, the standard defines the requirements for a holistic view. The paper at hand investigates if standard IT classification schemas can be

used to categorise qualification and attitudes of senior IT management from a holistic perspective taking into account the seven enablers of the COBIT standard. This is done by determining whether the self-proclaimed skills and competencies of IT executives actually reflect the entire spectrum of the COBIT specification. Therefore we conducted an analysis on the social media platform XING, in which the seven Cobit enablers are mapped to the user profiles of IT personnel having major management positions in German companies. We hypothesize that the usage of this data might give valuable insights into the perceptions of IT managers of what is important and expected by business partners and current or potential employers in terms of business continuity and crises mitigation.

An investigation on individual level doesn't help for generalisations, therefore a cluster analysis has been performed with more than 3600 items.

2 Data-Analysis

2.1 Data-Collection

User profile data of IT executives were collected through the XING API. To remove potential language or cultural biases the feasibility survey was restricted to Germany. The idea of using XING as base for the survey was born because of the disadvantages of classical survey methods: First, an online survey has a low response return rate (Baruch and Holtom, 2008) and we do not know which kind of candidates might answer (Assael and Keon, 1982; Nulty, 2008). The return rate is especially low if the there was no previous contact or if there is no relationship with the candidates. A semi-structured survey can in practise only be used for small number of interviewees (Drever, 1995) and has a different focus (Louise Barriball and While, 1994). Performing interviews without an initial explorative survey bears the danger of biasing the results as the survey proceeds (West et al., 2013) In all cases, retrieving a sufficient amount of subjects is essential for the understanding of the domain and a data source like XING can help to overcome the limitations of classical survey methods. However, subsequent to our analysis, the results can be verified using one of the above methods.

XING ("XING," 2018) is a social network service for professionals. Registered members can befriend business associates and list contact data, professional experiences, work stations, qualifications and skills (aka 'haves') and wishes (aka 'wants'). It is widespread in Germany with more than 10 million users and competes with the similar service LinkedIn. It is one of the major websites used by professional recruiters and companies ("Social Media Recruiting Report 2012 - ICR, Institute for Competitive Recruiting," n.d., "State of social platform use in Germany in 5 charts - Digiday," n.d.). We assume that the candidates - by trying to present themselves in the best possible manner - reflect the perceived expectations of relevant stakeholders (i.e., business associates, recruiters and potential employers) with regards to which qualifications are required and useful in their current or potential future position. By doing so, we link the micro-level to the meso-level in terms of sociological science or the individual level to the team or company level in terms of organisational science. This differentiation is a standard approach in institutional or social research.

Whether the IT executives really possess these attributes, cannot be verified by us. However, a certain quality control can be assumed since the IT managers are also visible to colleges, business contacts and previous employers and may therefore shy away from overstating their skills.

The XING platform provides an API for developers ("XING plugins | XING Developer," 2018). However, there are technical limitations. First, only 90 queries per day are allowed and, more severe, only a maximum of 100 hits per query are returned. Therefore, we had to split our queries into useful subqueries to retrieve the competencies of German IT executives. For this reason, we combined the actual search string with the name of each major cities (≥ 100.000 inhabitants) and medium-sized cities (< 100.000 and ≥ 20.000

inhabitants) in Germany and queried (in total 678 cities) (“Liste der Groß- und Mittelstädte in Deutschland,” 2018). This resulted in the following search strings:

Search string 1: *“Leiter IT AND [city name]”* (translated in English „Manager IT“) and

Search string 2: *“CIO AND [city name]”*

In total we pulled 19.587 users over a period of two months (November / December 2017).

2.2 Data Preparation

We anonymised, normalised and stored the data in an SQL database upfront of the data analysis. The attribute objects *employment_status, professional_expierence, business_address, educational_background, birth_date* (and contained fields) offer predefined list values and are well-structured. The *haves* category is the most important feature of our analysis. Here, IT executives state their competencies, skills and qualifications in a keyword-based text form. Therefore, the *haves* feature needed to be transformed with the help of natural language processing (NLP) operations**.** We realised that not only German words were used, but British and American English keywords were quite frequently stated as well alongside, with occasional spelling errors. We decided to deal with this irregularities manually at a later phase (2.3 Results) and apply no stemming or lemmatisation at all. To increase the quality and to analyse only the core information we applied a public stop word list (“GitHub - solariz/german_stopwords: Extended list of German stopwords for use in Web Projects, Search Engines or everything else.,” 2018), which was iteratively amended by own words after visual reviews. The resulting *haves* were normalised and stored in the SQL database for further analysis.

For all users or samples we found 1730 different characteristics for the *haves* feature. However single characteristics had a high variance in their distribution. For the analysis characteristics with a count <10 have been neglected. Each of these characteristics have been classified and assigned to one or more COBIT enablers. We found, that most of the stated characteristics fall in one of the seven COBIT enablers, while some needed to be grouped into multiple enabler categories (Information Systems Audit and Control Association, 2012, pp. 65–88). In cases where the keyword-based qualification statement spawned several COBIT enablers, it was assigned to all of them with the same weight (for each enabler a column was added that contained either “1” if applicable or “0” if not applicable).

Table 1. Example of attributes and their categorisation

attributes	like_or_equal	Principles, Policies and	Processes	Organisational Structures	Culture, Ethics, Behaviour	Information	Services, Infrastructure and	People, Skills and	description
.net	like	0	0	0	0	0	1	0	Programming Framework
agile transformation	equal	0	1	1	1	0	0	0	agile transformation
bcm	equal	1	1	1	1	1	1	1	Business Continuity Management
begeisterungsfähigkeit	like	0	0	0	1	0	0	0	enthusiasm
itil	like	1	1	1	1	1	0	0	framework

Since the classification has admittedly a certain subjectivity, three different individuals performed this task independently. Afterwards, we merged the results and discussed differences in detail.
After preparing the *category* table, these pattern were mapped to the individual values of the feature *"haves"* of each sample and thereby categorising the user qualifications by COBIT enablers. For the further analysis we had to deal with the different numbers of values for each sample. The count of values in a sample ranged from 4 to 20, some of the users provided only basic qualification, others went into more details. To make the characteristics comparable the values were normalised for each user (row oriented normalisation). That means that the sum of all of the seven enabler for a sample is always 1.

$$X_{\ enabler,user} = \frac{\sum attributes_{enabler}}{\sum attributes_{user}}$$

The result looked like:

Table 2. Example of assigned and normalised characteristics

user_id	Principles policies framework	processes	organisational structures	culture ethics behaviour	information	services infrastructure application	people skills competenc	sum
10002222_818bb0	0.1364	0	0.0909	0	0.0909	0.6364	0.0454	1
10003003_bd325a	0.1765	0.1765	0.2352	0.1765	0	0.0588	0.1765	1
…	…	…	…	…	…	…	…	

To remove unnecessary influences furthermore we restricted the samples of the observation (aka users) by several conditions. First, we included only executives, i.e. people in charge of a business unit, since we were solely interested in candidates in a leading position. Furthermore, we considered only people who are working in companies with more than 50 people – due to the fact that below that the company's organisational structures normally do not call for expressive hierarchies or division of labour in the IT departement. That reduced the amount of people to 3680.

Table 3. Filter conditions for survey

Filter	Possible values	Included values
employment_status	'employee', 'entrepreneur', 'executive', 'freelancer', 'public_servant', 'recruiter', 'retired', 'student'	'executive'
company_size	'1', '1-10', '10001+', '1001-5000', '11-50', '201-500', '5001-10000', '501-1000', '51-200', 'None'	'10001+', '1001-5000', '201-500', '5001-10000', '501-1000', '51-200',

2.3 **Results**

After having performed the pre-processing operations, we undertook an explorative analysis and applied a hierarchical cluster analysis algorithm on the data.

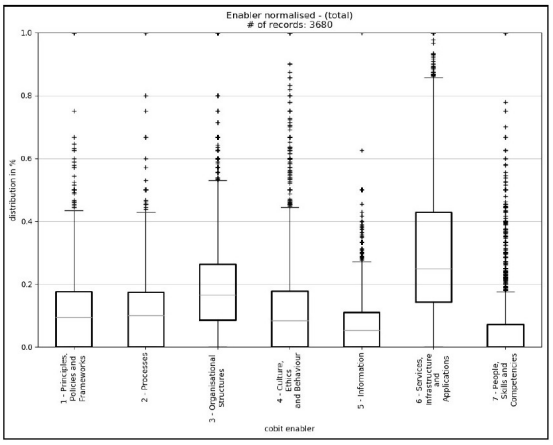

Figure 2. Histogram for all candidate characteristics with all COBIT dimensions

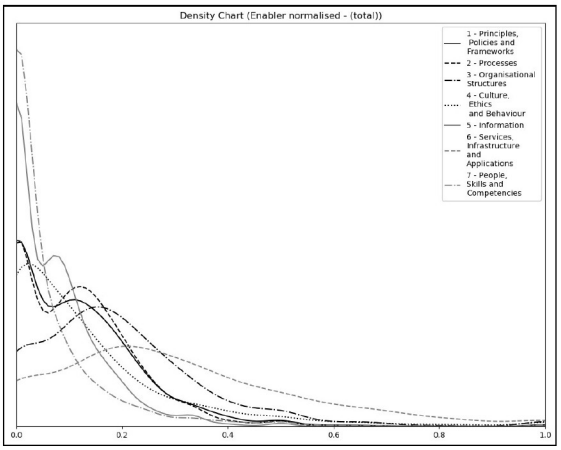

Figure 3. Density chart for all candidates with all COBIT dimensions

As starting point we created a histogram using all samples aka candidates and investigated the distribution of qualifications over all COBIT enabler aka dimensions (). Not surprisingly there is a relatively high proportion of counts in the technical dimension (enabler 6). All other bins are roughly equally distributed with each dimension being covered. Compared with the lower whisker we see a very high upper whisker and a considerable amount of outliers. For all bins we can see extreme outliers. That can further be confirmed by looking at the density chart - within the bins we can see left skewness and very long tails to the right ().

Afterwards, a hierarchical cluster analysis (distance measure “ward”) was performed to see if additionally certain patterns are hidden in the data. In order to identify the best cluster solution we created a dendrogram (Figure 4) and the related elbow chart (Figure 5).

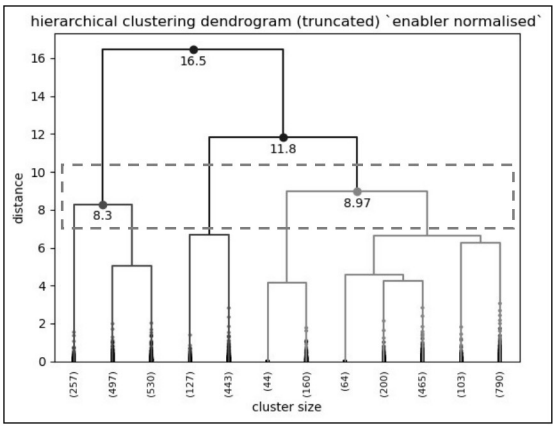

Figure 4. Dendrogram for all candidates for all COBIT dimensions

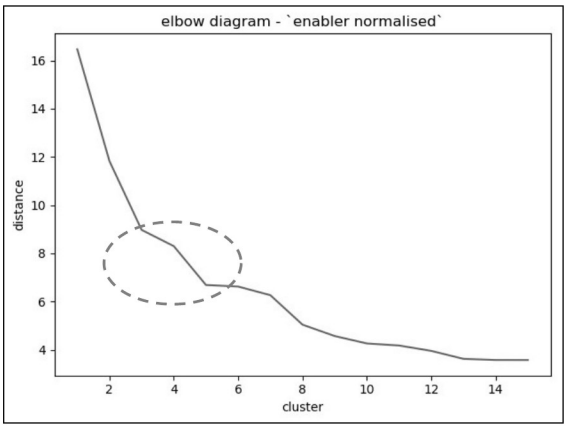

Figure 5. Elbow diagram for all candidates for all COBIT enablers

Both figures indicate that a good cluster solution might consist of 3 to 5 clusters. We decided that 4 clusters constitute the most significant and explainable cluster solution. With only 3 clusters we would miss the opposed slopes in the distribution over different company sizes as seen later in (). In a 5 cluster solution however, as was revealed by a detailed investigation of the histograms, the first cluster (Cluster 1 in) would get divided on the 6th COBIT dimension (“Services, Infrastructure and Applications”) into a medium and an extreme shape - quite often an indicator of overfitting. For these reasons, we decided that the 4 cluster solution represents the underlying patterns in the data most accurately. After having performed the hierarchal cluster analysis we calculated the same histogram, as initially done for all candidates, for each cluster.

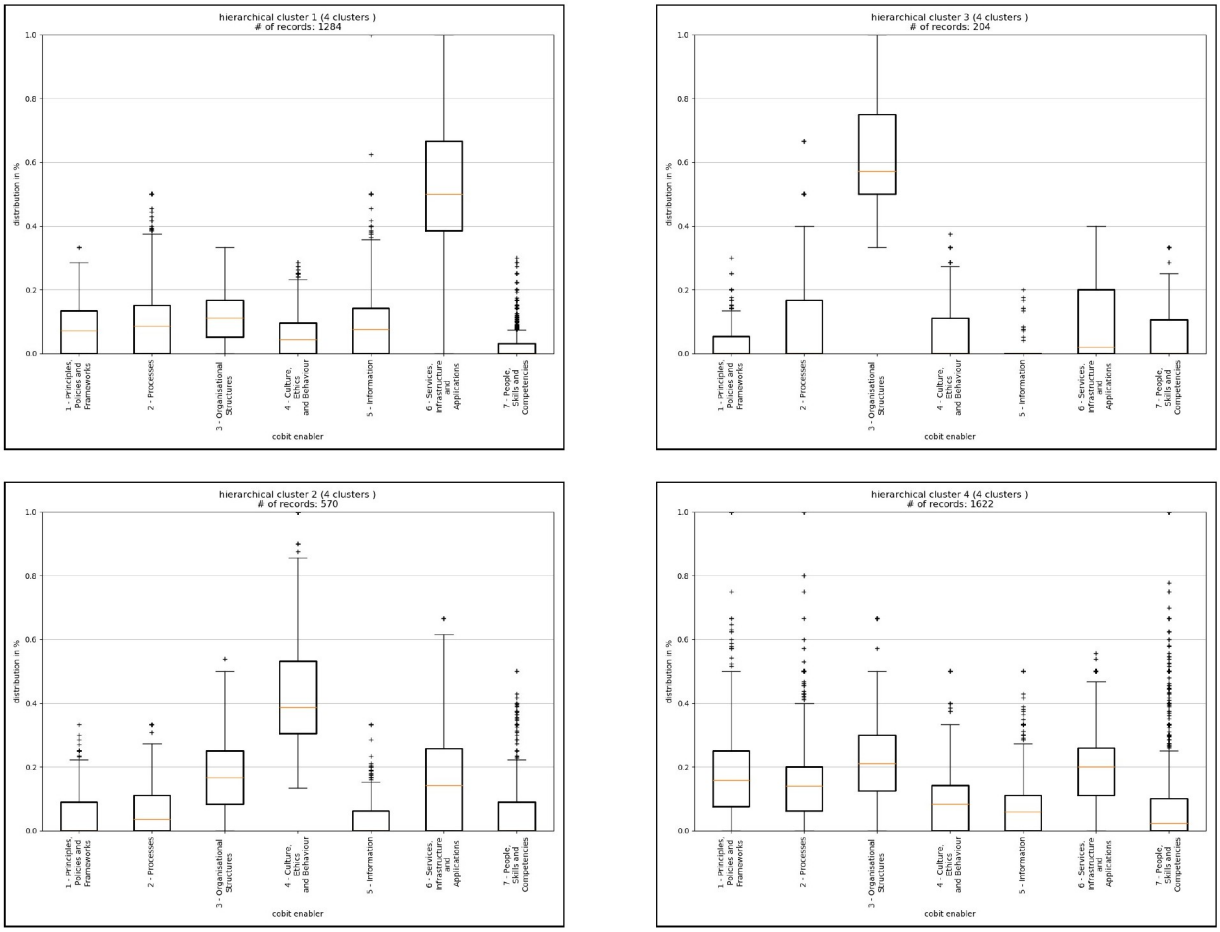

Figure 6. Histograms for each cluster (4 cluster)

For ease of explanation, we gave all clusters “speaking names” and will subsequently refer to those. The largest cluster (named “generalist”, cluster 4).with 1620 out of 3680 samples shows a fairly homogeneous distribution over all dimensions (the median in all dimension not exceeding .21), demonstrating not only technical but a holistic range of skills. The second largest cluster (named “technically focused”, cluster 1) shows a strong concentration on technical skills with a median of .5 in the dimension “Service, Application and Infrastructure”. The number belonging to that cluster is 1284 out of 3680 samples. It might not be surprising that an IT Manager has certain technical knowledge, but for executive managers we did not expect such a large of amount of samples with a strong focus in this area. With 570 of 3680 samples (named “socially focused”, cluster 2) quite surprisingly we found a type that emphasizes culture, ethics and behaviour (median of .39 in the dimension “Culture, Ethics and Behaviour”), but does not neglect organisational and technical skills at the same time. The smallest cluster contains only 204 of 3680 samples (named “organisationally focused”, cluster 3). Its members are very focused on organisational skills and the median there has the highest value of all medians (.57), even by far exceeding the technical skills median of the technically focused cluster.

The data do not allow insights into the depths of knowledge in the specified area. It can be only stated that the distribution in diverse areas is different. For the technical dimension it’s a fair assumption, that in smaller firms, employees need to cover a broader spectrum of technical skills than in larger firms. Furthermore, it can only be assumed that the concentration on less different products or technical skills leads to more expert knowledge in the specific technical area which in our investigation results in smaller numbers of characteristics. That might be the underlying reason of the broad inter quartile range (IQR) of the 6th enabler of cluster 1 (“technically focused”). In cluster 4 (“generalists”), besides a weak emphasis on the technical skills, we can see a second peak in the organisational skills, which goes well with the assumption of a “generalist” profile.

We wondered if the distribution of these clusters exhibits additional patterns. An additional interesting question arises, if the distribution of the clusters depend on the age or better the work experience of the samples. If different means could be identified, it might indicate an evolution from one cluster into another. Our first guess was that with growing work experience the members move from one cluster into another cluster. Therefore we grouped all clusters by work years of the members and compared the distribution. We might see a slight skewness to the right for the ‘organisationally’ focused cluster (cluster 3). Tus skewness could imply that membership in this cluster comes with certain seniority (Figure 7).

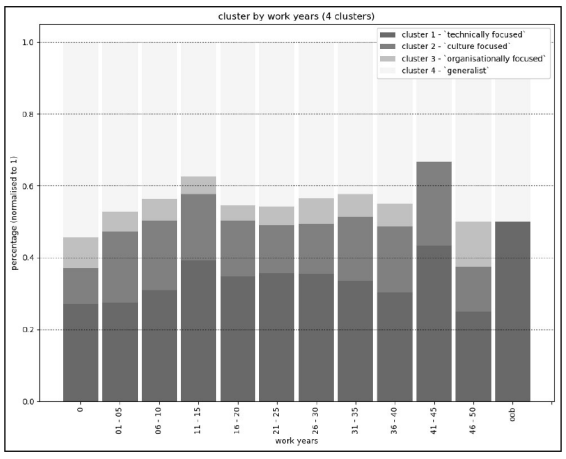

Figure 7. Distribution of work year for each cluster

We combined the analysis of the company size and work years feature, which indicates that with more work experience the member work in small companies and vice versa. However the population in each bin is not sufficient and for a detailed analysis additional

data sources must be linked, which is out of scope of the paper at hand. Therefore the analysis neither can prove a strong concentration in any work experience bin for any cluster nor could we identify a clear trend. Therefore, based on the data at hand the clusters are independent from work years. A second idea was to investigate, if the cluster depends on certain regions, indicating local trends. Again - no concentration could be identified and the clusters can be regarded as region independent.
However, if we group the relative cluster distribution by company size (Figure 8), a clear trend becomes visible. The percentage of the ‘generalists’ cluster (from a share of 0.376 for small to 0.520 for major companies) as well as the ‘culture focused’ cluster (from a share 0.156 for small to 0.172 for major companies) increases for larger company sizes. At the same time the ‘technically focused’ clusters and the ‘organisationally focused’ clusters decrease (from 0.396 for small to 0.270 respective 0.071 for small to 0.038 for major companies) with more employees in an enterprise.

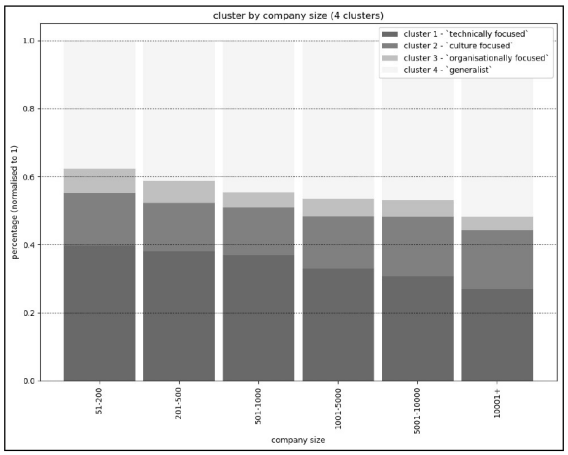

Figure 8. Distribution of clusters by company size

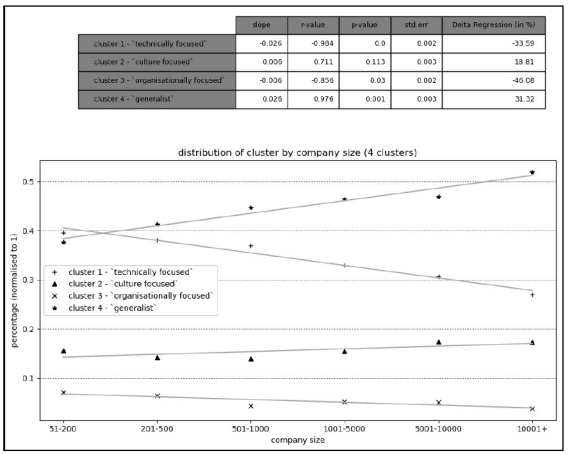

Figure 9.Slope of regression lines of clusters by company size

Looking at the slope () and the number of samples in a group, we can state that the bigger the firm the less technically focused people are found there. These seem to be replaced by generalists, but still, as later analysis will show, holistic minded people. Unfortunately, the data allow no insight on department size, but again it’s quite likely that teams are larger in companies with more employees. The ‘culture focused’ cluster is quite stable and shows only a small positive slope the larger the firms are. Overall, we can see a trend in senior management positions to have a more generalist and holistic profile for larger firms. Not so obvious on first sight is the cluster ’organisationally focused’ – the slope is much smaller but - taking into account the percentage change – the ‘loss’ from the bin for the smallest firms to the bin for the largest firms is quite drastic– we see that the distribution shrinks by more than 46 %. To explore the reasons are beyond the scope of the paper and must be postponed.

After analysing the primary features and identifying the clusters, evaluating secondary features we explored how the distribution of the distinct values for the primary feature (aka enablers) looked like within each cluster. In other words we counted the number of enablers of each candidate and aggregated the results for the whole cluster. By doing so, we investigated how holistic minded the samples are overall (Figure 10) and within each clusters (Figure 11).

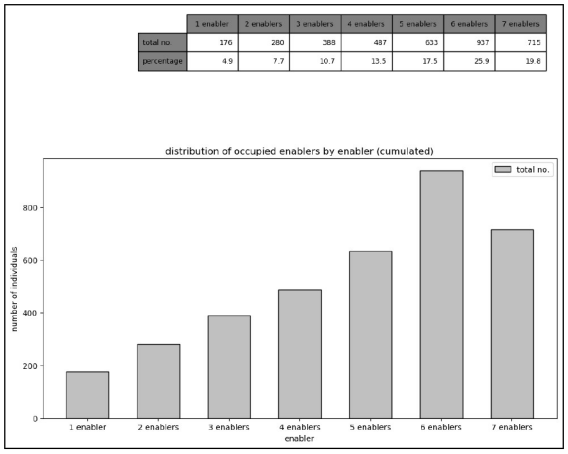

Figure 10 Distribution of occupied bins (enabler) by enabler

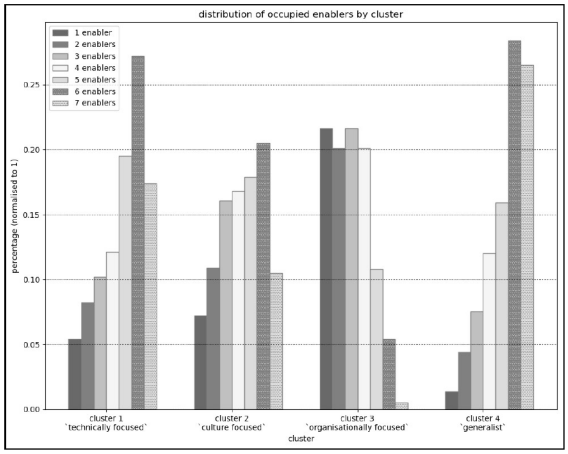

Figure 11 Distribution of occupied bins (enabler) by clusters

If we look at the overall representation, it could be reasoned that all samples are holistic minded with approx. 46% have occupied 6 or 7 enabler, and approx. 63% have occupied 5 to 7 enabler. If we divide the data by the identified clusters the picture changes. It is no surprise that the majority of the “generalists” (i.e., more than 50%) occupy between 6 and 7 enablers. More surprising is, that the “technical focused” cluster comes second with approx. 44% of the members that stated 6 or 7 enablers in their XING competence portfolio. This means that in spite of having a very strong technical focus, they still tend to demonstrate qualifications in many different areas and are holistic minded. On the other hand, the majority of the members of the organisational focused cluster (more than 60%) cover only 1 to 3 enablers.

Therefore Vaughan’s concerns regarding focusing in one area leads to negligence in others (Vaughan, 1999) cannot be confirmed for the majority of samples and clusters - based on the data at hand.

3 Conclusion

As the initial explorative analysis demonstrated, applying the COBIT enabler to the characteristics of IT senior executives certain patterns could be identified. These patterns confirm that two major groups – “generalists” and “specialist” – can be found. Therefore it has been proven, that COBIT and its enabler in conjunction with social networks can be used for classification for types of working attitudes.

Our analysis has revealed that company size influences competence profiles, whereas regional aspects or work years do not seem to be important. We have identified two major clusters – the technical oriented generalist IT managers. Smaller companies seem to expect from their IT management rather technical skills (more than 40%) which declines to less than 30% for large companies. At the same time generalists rise from less than 40 % to well above 50 %, thus making a shift in difference from -5 % to more than 20 %. The cluster with people having a focus on organisational skills decrease as well from small to larger companies, but with a more shallow slope. The socially or company culture focused people increase at the same time with an equal slope. Therefore it can be assumed, that the bigger the company the more general and social aspects become relevant. The analysis cannot make any predictions about the depth of knowledge, or the kind of company culture – but we can make conclusions what the people itself think is important.

A very interesting result is that both major clusters still have a broad coverage of all COBIT enablers. Hence, among these individuals a holistic mind-set can be assumed. While for the “generalist” cluster this is no surprise, it is quite unexpected that the members of the technical focused cluster still exhibit strong managerial skills. Further

research efforts need to be undertaken to determine how these commonalities and differences have influence on the organisation of companies and their culture.

Our study has shown that the analysis of social network data can give valuable insights into the characteristics of senior IT management in Germany, confirming previous studies with different approaches and is therefore a suitable method for research employee qualifications.

Assael, H., Keon, J., 1982. Nonsampling vs. Sampling Errors in Survey Research. Journal of Marketing 46, 114. https://doi.org/10.2307/3203346

Baruch, Y., Bell, M.P., Gray, D., 2005. Generalist and specialist graduate business degrees: Tangible and intangible value. Journal of Vocational Behavior 67, 51–68. https://doi.org/10.1016/j.jvb.2003.06.002

Baruch, Y., Holtom, B.C., 2008. Survey response rate levels and trends in organizational research. Human Relations 61, 1139–1160. https://doi.org/10.1177/0018726708094863

Custódio, C., Ferreira, M.A., Matos, P., 2013. Generalists versus specialists: Lifetime work experience and chief executive officer pay. Journal of Financial Economics 108, 471–492. https://doi.org/10.1016/j.jfineco.2013.01.001

Drever, E., 1995. Using semi-structured interviews in small-scale research: a teacher's guide, Practitioner minipaper. Scottish Council for Research in Education (SCRE), Edingburgh.

Gaulke, M., 2014. Praxiswissen COBIT: Grundlagen und praktische Anwendung in der Unternehmens-IT, 2., aktualisierte und überarbeitete Auflage. ed. dpunkt.verlag, Heidelberg.

Giannetti, M., 2011. Serial CEO incentives and the structure of managerial contracts. Journal of Financial Intermediation 20, 633–662. https://doi.org/10.1016/j.jfi.2011.04.003

GitHub - solariz/german_stopwords: Extended list of German stopwords for use in Web Projects, Search Engines or everything else. [WWW Document], 2018. URL https://github.com/solariz/german_stopwords (accessed 4.9.18).

Hille, M., Schnedermann, E., 2017. Security by Design - Die Rolle von IT-Sicherheitsstrategien in der Digitalisierung.

Information Systems Audit and Control Association (Ed.), 2014. IT control objectives for Sarbanes-Oxley: using COBIŤ 5 in the design and implementation of internal controls over financial reporting, 3rd edition. ed. ISACA, Rolling Meadows, IL.

Information Systems Audit and Control Association (Ed.), 2012. COBIT 5: a business framework for the governance and management of enterprise IT: an ISACA® framework. ISACA, Rolling Meadows, Ill.

La Porte, T.R., 1996. High reliability organizations: Unlikely, demanding and at risk. Journal of contingencies and crisis management 4, 60–71.

Lainhart, J.W., 2000. COBIT™: A Methodology for Managing and Controlling Information and Information Technology Risks and Vulnerabilities. The Journal of information systems 14, 21–25.

Liste der Groß- und Mittelstädte in Deutschland, 2018. . Wikipedia.

Long, M., Fischer, J., 2002. Project 2000-3 Leading edge enterprise: insights into employment and training practices. MONASH UNIVERSITY - ACER CENTRE FOR THE ECONOMICS OF EDUCATION AND TRAINING, Melbourne.

Louise Barriball, K., While, A., 1994. Collecting data using a semi-structured interview: a discussion paper. Journal of Advanced Nursing 19, 328–335. https://doi.org/10.1111/j.1365-2648.1994.tb01088.x

Murphy, K.J., Zabojnik, J., 2004. CEO pay and appointments: a market- based explanation for recent trends. American Economic Review Papers and Proceedings, New Developments in human-capital theory 94, 192–196.

Nulty, D.D., 2008. The adequacy of response rates to online and paper surveys: what can be done? Assessment & Evaluation in Higher Education 33, 301–314. https://doi.org/10.1080/02602930701293231

Perrow, C., 1999. Normal accidents: living with high-risk technologies, Princeton paperbacks. Princeton University Press, Princeton, N.J.

Ridoutt, L, Ridoutt, Lee, Australian National Training Authority, National Centre for Vocational Education Research (Australia), 2002. Factors influencing the implementation of training and learning in the workplace. National Centre for Vocational Education Research, Leabrook, S. Aust.
Ridoutt, L., Smith, C.S., Hummel, K., Cheang, C., National Centre for Vocational Education Research, L. (Australia), 2005. What Value Do Employers Give to Qualifications? Distributed by ERIC Clearinghouse, Place of publication not identified.
Roberts, K.H., 1989. New challenges in organizational research: high reliability organizations. Industrial Crisis Quarterly 3, 111–125. https://doi.org/10.1177/108602668900300202
Smith, C.S., Ridoutt, L., 2007. The importance employers attach to employee qualifications. Asia Pacific Journal of Human Resources 45, 180–199. https://doi.org/10.1177/1038411107079115
Social Media Recruiting Report 2012 - ICR, Institute for Competitive Recruiting [WWW Document], n.d. URL http://www.competitiverecruiting.de/ICRSocialMediaRecruitingReport2012.html (accessed 3.29.18).
State of social platform use in Germany in 5 charts - Digiday [WWW Document], n.d. URL https://digiday.com/marketing/state-social-platform-use-germany-5-charts/ (accessed 3.29.18).
Summary of risk methods and frameworks - NCSC Site [WWW Document], n.d. URL https://www.ncsc.gov.uk/guidance/summary-risk-methods-and-frameworks (accessed 9.3.18).
Vaughan, D., 1999. THE DARK SIDE OF ORGANIZATIONS: Mistake, Misconduct, and Disaster. Annual Review of Sociology 25, 271–305. https://doi.org/10.1146/annurev.soc.25.1.271
Weick, K.E., Putnam, T., 2006. Organizing for Mindfulness: Eastern Wisdom and Western Knowledge. Journal of Management Inquiry 15, 275–287. https://doi.org/10.1177/1056492606291202
Weick, K.E., Sutcliffe, K.M., Obstfeld, D., 2008. Organizing for high reliability: Processes of collective mindfulness. Crisis management 3, 81–123.
West, B.T., Kreuter, F., Jaenichen, U., 2013. “Interviewer” Effects in Face-to-Face Surveys: A Function of Sampling, Measurement Error, or Nonresponse? Journal of Official Statistics 29. https://doi.org/10.2478/jos-2013-0023
XING plugins | XING Developer [WWW Document], 2018. URL https://dev.xing.com/ (accessed 4.9.18).
XING [WWW Document], 2018. . Wikipedia. URL https://en.wikipedia.org/w/index.php?title=XING&oldid=829074295 (accessed 4.9.18).

3.2 Publikation 2: „Comparison of Social Media and Panel Data Analyses“

Im folgenden Abschnitt versucht der Autor anhand einer praktischen Untersuchung den Kreis der Wissensgeneration zu schließen und verfolgt dabei den induktiven Ansatz. Durch Beobachtungen können neue Erkenntnisse gewonnen und Theorien gebildet werden. Die Gewinnung von Daten durch Beobachtungen und Befragungen ist ressourcenaufwändig und zeitintensiv. Somit stellt sich die Frage, ob dieser Prozess durch die Nutzung von Sozial Media Daten unterstützt werden kann.

Basierend auf einer existierenden, empirischen Studie wurde versucht, die Ergebnisse mit Hilfe von Social Media Datenquellen zu reproduzieren. Darauf aufbauend wurden Unterschiede, Herausforderungen und Möglichkeiten beim Einsatz von Sozialen Medien Daten im Vergleich zur Studie auf Panel Daten untersucht. Die Basisstudie untersuchte die Migrationsbewegungen von Studenten 5 Jahre nach Abschluss ihres Studiums. Als Datenbasis diente eine klassische Datenquelle - ein Zensus Panel. Dieses Panel wurde zentral und aufwendig für Deutschland geplant, organisiert und wird seit 1990 in einem 5-jährigen Rhythmus wiederholt. Zu diesem Zweck, werden die Prüfungsämter sämtlicher deutscher Hochschulen mit der Bitte die Umfrage an die bekannten Alumni zu verteilen angeschrieben. Die Ergebnisse werden anonym gesammelt, bereinigt und der Wissenschaft zur Verfügung gestellt. Der Fragenkanon wurde einmalig entwickelt und bleibt über den gesamten Erhebungszeitraum konstant. Die dem Vergleich dienende Social Media Datenquelle war erneut das soziale Netzwerk XING. Für den Prototyp wurde die Erhebung der Daten auf vier Universitäten beschränkt.

Die im Rahmen dieser Arbeit vorgenommene Analyse konnte, anhand der gewählten Beispielen, ähnliche Migrationsmuster wie die Basisstudie zeigen.

In der durchgeführten Untersuchung konnte die Ergebnisse der Baseline Untersuchung bestätigt werden. Jedoch zeigten sich deutliche Unterschiede hinsichtlich der möglichen zu verwendenden Datenbasisgröße, als auch in den Möglichkeiten Aussagen abzuleiten. Final stellt man in der Untersuchung fest, dass beide Analysen sich sinnvoll ergänzen und bereichern und somit nicht alternativ sondern als komplementär zu sehen sind.

Das Publikation wurde mit dem „Second Best Paper Award“ der Veröffentlichung bedacht [136].

Comparison of Social Media and Panel Data Analyses

Sven Gehrke, Dr. Marek Opuszko, Sandra Niemz, Prof. Johannes Ruhland

Chair of Business Informatics, School of economics and Business Administration Jena, Germany

sven.gehrke@uni-jena.de
marek.opuszko@uni-jena.de
sandra.niemz@uni-jena.de
johannes.ruhland@uni-jena.de

Abstract

In the past, social research data was mainly collected through interviews or surveys. Officially collected or available data collections (so called “panel” or “census data”) must be anonymized. The requirement for anonymization results within the EC from the GDPR. In practice, this means that the data is often only available in aggregated form, which anonymises the interviewees but ultimately prevents the use of data mining methods. Currently people focus on social media data sources and classical data sources loose their attraction. Social media platforms store a large number of data that are provided actively by the examined persons themselves. The paper tries to advocate for a mutual reinforcing usage of both methods and therefore releasing the full potential. The authors show based on a practical example, where both approaches have their strengths and weaknesses.

The presented paper attempts to recreate the results of an analysis (Haussen and Uebelmesser, 2018) based on classic panel data by using social media data instead. The method used in the article is based on design science (DSR) approach, that is, based on the previous survey, a prototype has been created and the artefacts are compared. To structure the comparison of baseline and prototype, CRISP DM was applied and every single step analysed and compared. In the evaluation step, the argumentation focuses on the four V-characteristics of big data analysis in order to do justice to both approaches. Based on the results, the authors advocate the mutual use of both data sources, whereby it can be shown that the mutual use strengthens the quality and informative value of the results of each individual approach and that neither approach is a substitute for the other.

Keywords: data mining, social media data, panel data

Motivation

Studies generally require data and today data are seen as the new gold or oil, both in science and business(Espinosa and Armour, 2016; Lohr, 2012; Singh, 2013). The gathering of data is generally associated with great expenditure of time and money. Social Media platforms like Facebook (founded 2004), Twitter (founded 2006) and XING (founded 2003) collect huge amount of data, which are provided and maintained by the users itself. Much of the data is publicly available and available for analysis purposes. Before that, data have been collected through tedious interviews or derived from similar sources. That process required sound funding in people and money and took considerable time. These should be sufficient in scope and type in accordance with the purpose of the investigation. Furthermore officially collected or available data collections (so called “panel” or “census data”) must be anonymized. The requirement for anonymisation results within the EC from the GDPR (article 5) and local law (“Art. 5 GDPR - Principles relating to processing of personal data,” 2018; “Data anonymization and GDPR compliance,” 2019; “Datenschutz,” n.d.; Meyermann, n.d.). Operational analyses are often subject to more stringent regulations depending on the works council. In practice, this means that the data is only available in aggregated form, which ultimately prevents the use of classic data mining methods as described in standard literature (e.g. Hsiao, 2014).

Another issue with panel data is that the data often do not correspond directly to the actual research goal. The canon of questions rarely cover the entire investigation area, the time of the survey and the repetition intervals are given, meaning that the use is problematic or not possible for longitudinal studies (repetition over a given period). On the other hand, current data is available in many social media platforms. This data can partly be read out via special APIs (e.g. Twitter etc.) or directly from the websites (e.g. XING). In principle, this data is publicly available and can be used for

analyses, especially in the academic research (data collected indirectly by the platform provider is not in focus of this paper).

The question that arises here is firstly whether the data from the social media platforms leads to similar results as studies based on the above census data (referred to as "veracity" or "validity" in the analysis of big data), if the amount can keep up with the census data (referred to as "volume" in the analysis of big data) and can be collected in an acceptable time (referred to as a part of "velocity" in the analysis of big data) and secondly whether the underlying data structure is suitable for the analysis (in the Analysis of Big Data called "Variety") (Kitchin and McArdle, 2016; Madden, 2012; Zikopoulos, 2012)et al. In addition, it must be clarified whether there are further advantages such as the mentioned topicality and repeatability and thus an equivalent or extended value compared to census data can be found.

Discussion of the Survey

General Approach

The motivational questions were investigated using an approach close to the design science research methodology (DSR) (Peffers et al., n.d.). The "netnography" method (Kozinets, 2015) is only partly applicable, since the paper concentrates mainly on the underlying data and less on the cultural and social aspects of the investigated subjects. The analysis is basically structured by the phases described in the CRISP DM data model, which has become the de-facto standard in data mining for decades (Shafique and Qaiser, 2014; Wirth and Hipp, n.d.). The artefacts are then compared with the results of the underlying study and the possibilities and limitations of using social media platforms are examined. The comparison is structured based on the process model for data mining analyses CRISP DM. Finally both results have been compared and differences and common grounds structured by the four "V" of Big data shown.

The used baseline study was "No Place Like Home? Graduate Migration in Germany" (Buenstorf et al., 2016; Haussen and Uebelmesser, 2018) and the underlying data source (Briedis et al., 2020), which examines the migration movements of university graduates. The underlying data source are panel based, anonymised and aggregated by the data collector and therefore applicable as baseline study in the above described scenario. For our purpose we concentrate on the first part of the study, especially how graduates migrate five years after their graduation.

An attempt is made to provide similar data with the help of a prototype. The prototype for the DSR approach uses the XING as an example of a social media data source. XING is an SM platform for professionals, which is extremely popular in Germany ("XING," 2020). The XING data were collected contemporary to the panel investigation (2016). One assumption is, that the candidates are at least as honest about their data as in any other interview or survey, especially the above mentioned baseline study. If the members really possess these attributes, cannot be verified by us. However, a certain quality control can be assumed since the participants are also visible to colleges, business contacts and previous employers and may therefore shy away from distort the information.

A complete analysis of all universities was omitted for reasons of time and against the background that only the basic feasibility should be examined. Four universities[1] were selected for the study, which are comparable in size (medium-sized) and degree programs ("full university" ("Volluniversität," 2020)). The universities were chosen from four different provinces (Bavaria, Hesse, Saxony-Anhalt, Thuringia), two each in the area of the former FRG and GDR. The relevant statistics are shown below (Table 1).

Table 1: Gross Migration Patterns(Haussen and Uebelmesser, 2018, p. 446)

	Stayer	Onward Migrants	Return Migrants
Bavaria	67.2	25.9	6.9
Hesse	63.2	29.8	7
Saxony-Anhalt	28.4	61.9	9.7
Thuringia	26.1	66.8	7.1

1 The universities are:
- Justus Liebig University, Gießen
- Friedrich Schiller University, Jena
- Otto von Guericke University, Magdeburg
- University Regensburg, Regensburg

Not directly taken from the underlying data but linked to it in the study itself, is the statement that graduates tend to stay close to their university and that urban areas are more generally attractive (Haussen and Uebelmesser, 2018, p. 443; Krabel and Flöther, 2014).

The derived hypothesis are

H1: The numbers can be reproduced using social media data.
H2: numbers should differ between urban and rural areas, with significant higher numbers in urban areas (not only in absolute numbers but relatively to the population density)

If the above hypothesis cannot be rejected, the prototype tries to emphasize on differences and mutualities, and if both approaches can profit from each other.

CRISP DM: Business Understanding

Baseline Survey

The data of the baseline paper are based on the „DZHW Graduate Study Series“ (“Data Search for Higher Education Research and Science Studies,” n.d.) , which use standardised questionnaires to gather information on graduates, their career entry, career development and further qualification of university graduates. The same questionnaire has been used for each cohort. The Panel contains four surveys starting in 1989 and then every four year until 2013 (the actual collection of the data took place within two years after the start (i.e.: 2013 → 2013-2015). No further survey has been conducted since. The latest graduate cohort has been selected by the above study and therefore these data serve as baseline. The study itself is interested in migration movements of university graduates within Germany. A distinction between different university courses / degrees was not made and is hardly possible due to the underlying data.

Prototype

The Social Media Platform XING is widespread in Germany with more than 10 million users and competes with the similar service LinkedIn. It is one of the major websites used by professional recruiters and companies (“State of social platform use in Germany in 5 charts,” 2017). Registered members can befriend business associates and list contact data, professional experiences, work stations, qualifications and skills, wishes and list the the educational background. The platform is not restricted to university graduates.

Since the prototype serves only for a proof of concept, we omitted a complete analysis of all universities.

CRISP DM: Data Understanding

Baseline Survey

The underlying questionnaires (“Panel Questionnaire 2009,” n.d.; “Panel Questionnaire 2013,” n.d.) and the description of the package (Baillet et al., n.d.) are available in detail. Basically each survey cohort or wave took one year to collect the data. In each survey the data have been collected in three steps. The first and second step uses a standardized postal survey. Since the address information are only available at the university of each student, the questionnaire has been sent to the examination office of each participating university, and then from there to all students. Several university-like education sites did not take part in the survey like universities of cooperative education (“Berufsakademien”), universities of the armed forces (“Bundeswehr”). The responsibility for the distribution lay within each examination office. How they know the contact details of the graduates is unknown, and quite likely based on the voluntary sharing of the details by the students. During the first and second step, the graduates have been asked to provide contact information, which in the third step then had been used to conduct an online survey. The return rate on the contacted graduates varies between approx. 25% in the first step up to approx. 60% in the third step. So the overall survey population contained 8477 samples in total in the first step and considerable less in the third step (for comparison: 2005 Germany had approx:> 200,000 university graduates annually (“Studienanfaenger-Absolventen.pdf,” n.d.)). A large number of universities/ students from Northrhine-Westfalia refused to participate. Basis data preparation has been done already during the survey, therefore except grouping and filtering nothing else needs to be done before the modelling.

Prototype

In principle, the data on the platform are in non-normalized text form. This means that while the user of the platform is entering the data, there are no restrictions with regard to the field contents, nor is an exact structuring specified by the platform. Only central points are given. The underlying data is only described by the name. The query to select specific

candidates carries out a full text search, which means that intensive follow-up tasks are necessary. The platform operator limits data collection to a maximum of 200 data records per query. To get around this limitation, the question about the university was combined with one current, official degree. The degrees were taken from the websites of the relevant university. With this approach “200 by number of degrees per university” hits are possible. To further extend the result set, in a second step the friends/ contacts of the first hits were examined too until sufficient candidates could be identified.

CRISP DM: Data Preparation

Baseline Survey

Most of the data preparation has already been carried out by the panel data collector. Thus no further processing is necessary.

Prototype

The collected data sets (> 70,000) were filtered further after the initial retrieval. In the process, all data records were deleted from which it could be inferred that they came from people who are currently studying (“current occupation”). Furthermore, all data sets were eliminated which did not contain any references to one of the 4 universities under the node "Education". This is potentially the case if the name of the university appears throughout the document, e.g. if the person concerned is employed at the university but did not study there. Records that did not contain valid time information for the university visit were also excluded. In total, around 7,000 data records remained. Approximately 1,800 valid data sets were selected for each university. Data relevant for the analysis potentially included the education, the place of study, the field of study, the degree and the time of study. The data records also contain the name and would enable identification, which is, however, uninteresting for the analysis and problematic in terms of data protection laws and was therefore not used.

Linking the location data with an additional, high-quality data source (“PLZ Download,” n.d.) showed that only a minimal amount of records had to be excluded due to spelling errors or ambiguities (duplicate place names / unknown place names). However, it is unclear whether the current location represents the location of the current employment, the headquarters of the company or the place of residence of the employee. This applies to the same extent for all selected universities and is for the intra prototype comparison negligible. For the the comparison between baseline and prototype it might be a reason for divergences.

Overall, a previous study have shown that the quality of the data and its truthfulness can be described as good. It is assumed that the users of the platform through the network of relationships within the platform, the users, represent themselves relatively truthfully (a slight bias in the qualifications should affect all users equally). (Gehrke et al., 2019)

For all remaining data sets the Euclidean distance between the university and the current position was calculated. At the same time the postcode of the current position was determined. The results were evenly divided into bins in order to correspond more closely to the reality of the postcode area and not to distort the comparison with local peaks (20 km). Furthermore, the data set to be examined was combined with statistical geo-data for an exploratory data analysis. This means that the number of graduates per university has been standardised with the total population of the district in which they are located (proportion of graduates in the population of the postcode area).

Further data preparation has not been applied.

CRISP DM: Modelling

Baseline Survey

The baseline model draws its findings mainly from basic statistic. Data mining techniques are not applicable since all samples are anonymised in such a way that no conclusion about the person behind can be made. Unfortunately, all interdependencies have been removed along the process. The main findings are reflected in hypotheses H1/ H2 and can be found in detail described in detail in the baseline paper (Haussen and Uebelmesser, 2018).

Prototype

The figure (Illustration 1) shows the distribution of graduates across the districts. The graduates were normalized by the number of residents in order to represent the proportion of graduates in the population of the districts. For better visibility, the districts have been combined (the 5-digit postcodes have been combined into the first two digits). It became apparent that the graduates settled generally in circles around the place of study, on the one hand, and that large

cities are very attractive on the other. However, the presentation is not suitable for further analyses; in particular, quantitative statements are difficult to generate and interpret. The concentrations in large cities can also hardly be recorded visually.

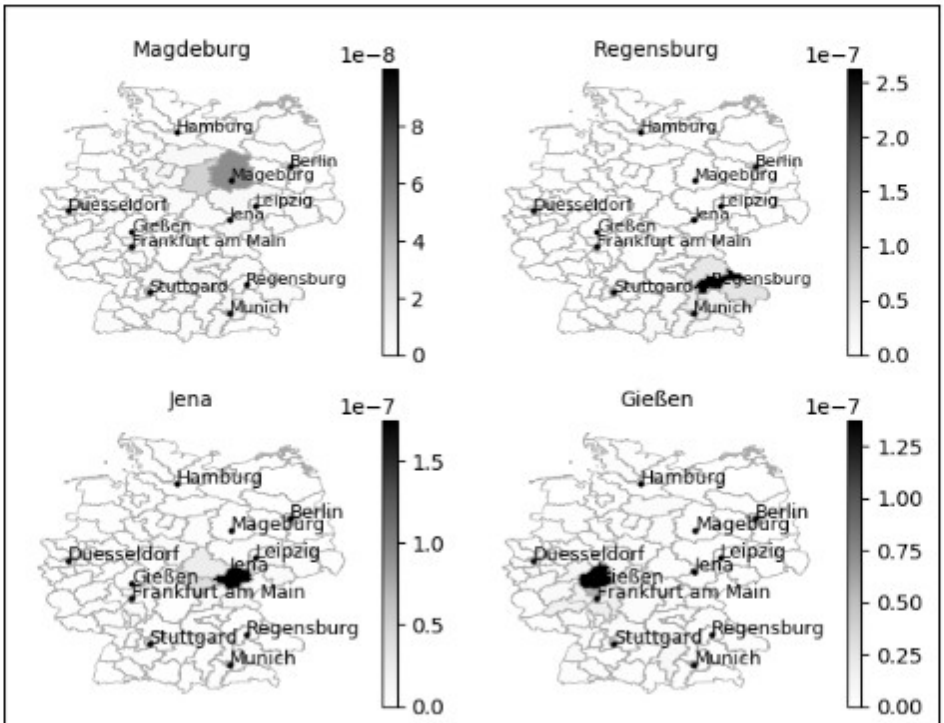

Illustration 1: distribution of graduates for 4 universities

In order to make the migration movements of the graduates quantitatively measurable within the prototype, the distance between the current place of residence and the place of study was measured (Euclidean distance). The following can be seen from the illustration (Illustration 2,Illustration 3):

1. the majority of the graduates are still based near the place of study.
2. The density decreases with increasing distance. In order to do justice to geographical conditions and for better evaluation, the distances were divided into equally sized bins (20km). An exponential distribution can be "guessed", in any case there is no linear distribution. A regression coefficient for a linear regression confirms this. In order to compensate for the increasing surface area while keeping the radius constant, a formation of the bins (root (n) * radius) representing the same area was carried out. However, this processing did not show any linear dependency either. This approach will not be pursued further in the course, since no additional knowledge is expected and the interpretation would only be more difficult. In particular, the radius decreases so much with increasing distance that the bins formed become very sparse populated.
3. The general exponential distribution is "interrupted" by local clusters / outliers

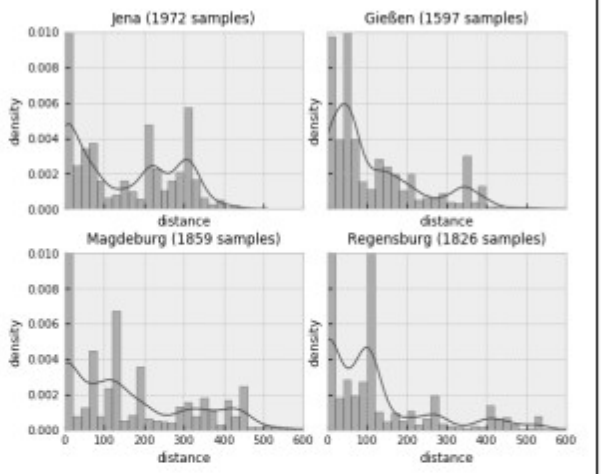

Illustration 2: distances university - current location (equal bins by radius)

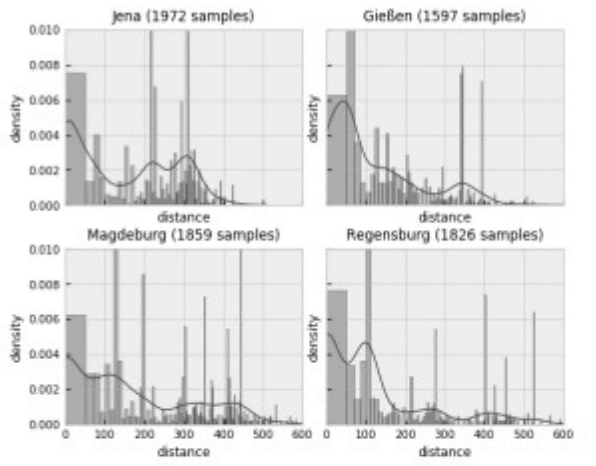

Illustration 3: distances university - current location (equal bins by area)

Further investigations of the local clusters / outliers show that the outliers are often caused by large cities. To illustrate this more clearly, the residuals of the estimated distributions were examined. The following figure (Illustration 4) shows this graphically. It can be seen visually that the estimated distribution only has extreme outliers in a few places[2], otherwise the distribution is close to the median of the residuals.

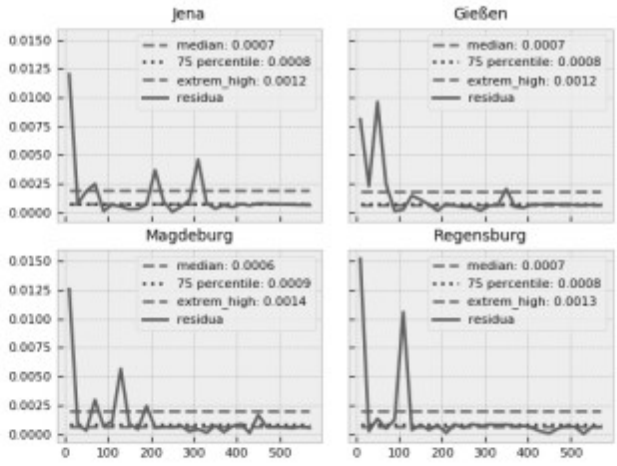

Illustration 4: residua of fitted exponential distributions

The first residue bin deviates strongly in all of the distributions examined, which indicates a very strong concentration of graduates near the place of study. This confirms the previous visual assumption. Otherwise the distribution can be approximated well with an exponential function. The estimated distributions were compared using the Jensen-Shannon Divergence (JSD [23]). The JSD is a modification of the Kullback Leibler divergence and, in contrast to this, is symmetrically and [0: 1] normalized, where 0 stands for identical probabilities. The calculated values show relatively similar and comparable distributions for the probabilities (Illustration 5, Illustration 6). Having comparable distributions for all investigated universities, it indicates that the results can be generalised.

2 Upper Extrem/ outlier values are commonly defined as: Outlier > Q3+ k(Q3-Q1) (Tukey, 1977)

	Jena	Gießen	Magdeburg	Regensburg
Jena	0.000			
Gießen	0.332	0.000		
Magdeburg	0.340	0.371	0.000	
Regensburg	0.386	0.440	0.381	0.000

Illustration 5: JSD for fitted distribution

	Jena	Gießen	Magdeburg	Regensburg
Jena	0.000			
Gießen	0.332	0.000		
Magdeburg	0.340	0.371	0.000	
Regensburg	0.386	0.440	0.381	0.000

Illustration 6: JSD for residua

In order to explain the difference, the extreme values for each distribution were examined more closely. The cities in the bins, which „disturbed“ the curve have been counted and sorted sorted in descending order. The result can be seen in the following table (Illustration 7).

university	bin	location city	inhabitants	alumni by density
Jena	[200, 220)	Berlin	3291932.0	0.0167
Jena	[40, 60)	Erfurt	200963.0	0.0405
Jena	[60, 80)	Leipzig	503235.0	0.0242
Jena	[300, 320)	München	1367641.0	0.0083
Jena	[300, 320)	Hamburg	1726063.0	0.0108
Gießen	[40, 60)	Frankfurt am Main	668327.0	0.0316
Gießen	[340, 360)	München	1367641.0	0.0054
Gießen	[340, 360)	Hamburg	1726063.0	0.0075
Gießen	[20, 40)	Marburg	78549.0	0.028
Magdeburg	[120, 140)	Berlin	3291932.0	0.0211
Magdeburg	[180, 200)	Hamburg	1726063.0	0.0164
Magdeburg	[60, 80)	Wolfsburg	120030.0	0.0902
Magdeburg	[440, 460)	München	1367641.0	0.0082
Magdeburg	[60, 80)	Braunschweig	242670.0	0.0169
Regensburg	[100, 120)	München	1367641.0	0.0358
Regensburg	[100, 120)	Passau	48699.0	0.0135
Regensburg	[100, 120)	Augsburg	267780.0	0.0023
Regensburg	[100, 120)	Erlangen	106852.0	0.0023

Illustration 7: top 5 extreme values by university and city

The extremes considered are always large cities, which are obviously more attractive to graduates or their employers. It should be noted that not only the number of graduates was considered, but also their proportion of the population. A more detailed analysis with the underlying data is not possible because a more detailed analysis of the major cities would have to use the distribution of all graduates and therefore all universities.

CRISP DM: Evaluation

First, following the DSR approach, we have to establish that the prototype can reproduce the same results as the baseline.

For Hypothesis **H1** the results of the of baseline and prototype are shown below (Table 2). The distinction between migrants leaving for good and migrants returning within a five years period cannot be made by the prototype. However if we sum the “stayer” and “returning migrants” to “total stayers”, a comparison can be made. The numbers are not met exactly, but the trend is still recognizable. The clear exception is Gießen, which cannot be explained by looking at the data. Looking at the geographical location, you can see that the university of Gießen is very close to two other provinces. It can be assumed that adding the data of the other three major universities in the province would amend the accuracy to the same level as the other samples.

Table 2: Comparison baseline and prototype numbers

	onward migrants (%)			total stayers (%)		
	Baseline	Prototype	Diff.	Baseline	Prototype	Diff.
Bavaria/ Regensburg	25.9	21.7	4.2	74.1	78.3	-4.2
Hesse/ Gießen	29.8	48.8	-19.0	70.2	51.2	19.0
Saxony-Anhalt/ Magdeburg	61.9	66.7	-4.8	38.1	33.3	4.8
Thuringia/ Jena	66.8	62.8	4.0	33.2	37.2	-4.0

From our point of view, hypothesis **H1** might not be rejected even if its not an exact match.

For the second hypothesis **H2** we were able to trace back the peaks in the residua of the distribution of graduates migration to urban areas (Illustration 4, Illustration 7).

After demonstrating, that our hypotheses are valid we concentrate on strengths and weaknesses of both approaches, and advocating for a mutual complementary use. To structure the comparison we use the four "V"s.

Volume

Both surveys have similar number of samples. However the baseline survey has to cover whole Germany, whereas the prototype covers only four universities and could gather more samples for the remaining universities. The questionnaire return rate is quite impressive, but compared to the total number of graduates it does not exceed 2%. Since it requires the possession of contact information of the graduates at their Alma Mater and their willingness to participate the number of participants is basically limited to the current numbers.

In principle, the consent of all stakeholders to participate is required, which results in a clear advantage when using social media.

Variety

Variety as in linking additional data sources is not really in the focus of both surveys. The mechanism or sources used in the baseline survey to enrich the sample data are not known.

The prototype however uses two additional data sources – the universities internet presences for retrieving the possible graduation degrees and the official internet presence for post codes to detect the county of the current location of each graduate to link it to a county and therefore to enrich the data with the size and position of each county and the population of each county.

In general, the possibility to link the data is not dependent on the source, but how the key fields used for linking the sources can be controlled. In panel data the answers can be controlled by providing a set of possible answers. In social media platforms no control is possible. Therefore both methods face similar challenges, with panel data having an advantage if planed in advance.

Velocity

The velocity for collecting the data differs considerable. In the baseline survey the process of collecting the data took one year, the time for preparing the process is unknown but might not be negligible. It can only be assumed, that there is not much room for speeding up. The collecting of the data for the prototype took four weeks for a single person/ computer. It could scaled by increasing the number of people collecting the data. It might even be possible to contact the platform operators to retrieve the data directly.

In terms of velocity there is a clear advantage in using social media data over panel data.

Veracity

Veracity is by far the most interesting characteristic. The prototype could establish similar results as the baseline survey. That leads to the understanding, that user of social media platforms are at least truthfully about their public profiles as they are truthfully in panel surveys. Using a range of data mining techniques, different insights into migration patterns have been obtained, in particular that graduates migrate within Germany with a negative exponential distribution with the university as centre. Furthermore the indications for both hypotheses could be found in one study.

One advantage is of the use of social media data is, that this knowledge generation is possible retrospectively - which means that it does not depend as much on advance planning as panel data. However, if it becomes evident, that if very specific questions needs to be answered, either in granularity or area, that advantage diminishes. In an iterative process

finally it depends on the willingness of the objects of the investigation to participate – or the possibilities to link additional data.

Another important fact is that using two different approaches leading to similar results and it strengthens the fundamental informative value of both approaches. Since the prototype can confirm basic statements of the baseline survey, further statements of the prototype should lead to valid results as well. That makes the application of data mining techniques applicable, which are not possible with panel data. On the other hand, the results of the panel survey with little sample size in comparison to the prototype are more justified transferable to the total survey population.

Conclusion

It could be shown, that the motivational questions could be answered. Panel Data and Social Media Data can answer research questions to a certain degree equally. However as the prototype could demonstrate, its not a decision for one and against the other, both have advantages and disadvantages. Both kind of survey can be considered mutually reassuring. Panel data based on questionnaires can be more specific and can actively controlled and thus answer research questions with greater granularity. On the downside panel survey definitively require strong ahead planning, while social media analyses are more flexible and allow initial, exploratory surveys. Social media analyses enable data mining techniques and the linkage of further data sources, however in most of the times the researcher has little control and acts passively. All statements about social media data are, of cause, under the assumption that there is a social media platform actually covering the object of investigation. The present positive study can serve as a pilot for further, broader and more specific investigations.

Art. 5 GDPR - Principles relating to processing of personal data [WWW Document], 2018. . GDPR.eu. URL https://gdpr.eu/article-5-how-to-process-personal-data/ (accessed 3.5.21).

Baillet, F., Franken, A., Weber, A., n.d. DZHW-Absolventenpanel 2005 49.

Briedis, K., Fabian, G., Landers, G., Redeke, S., Rehn, T., Schulz, J., Trennt, F., Deutsches Zentrum Für Hochschul- Und Wissenschaftsforschung (DZHW), 2020. DZHW Graduate Panel 2013DZHW-Absolventenpanel 2013. https://doi.org/10.21249/DZHW:GRA2013:1.0.0

Buenstorf, G., Geissler, M., Krabel, S., 2016. Locations of labor market entry by German university graduates: is (regional) beauty in the eye of the beholder? Rev. Reg. Res. 36, 29–49. https://doi.org/10.1007/s10037-015-0102-z

Data anonymization and GDPR compliance: the case of Taxa 4x35 [WWW Document], 2019. . GDPR.eu. URL https://gdpr.eu/data-anonymization-taxa-4x35/ (accessed 3.5.21).

Data Search for Higher Education Research and Science Studies [WWW Document], n.d. . Res. Data Cent. High. Educ. Res. Sci. Stud. URL https://metadata.fdz.dzhw.eu/en/search (accessed 11.29.20).

Datenschutz: Darauf sollte man bei der Anonymisierung achten [WWW Document], n.d. . ComputerWeekly.de. URL https://www.computerweekly.com/de/ratgeber/Datenschutz-Darauf-sollte-man-bei-der-Anonymisierung-achten (accessed 3.5.21).

Espinosa, J.A., Armour, F., 2016. The Big Data Analytics Gold Rush: A Research Framework for Coordination and Governance, in: 2016 49th Hawaii International Conference on System Sciences (HICSS). Presented at the 2016 49th Hawaii International Conference on System Sciences (HICSS), IEEE, Koloa, HI, USA, pp. 1112–1121. https://doi.org/10.1109/HICSS.2016.141

Gehrke, S., Wenige, L., Ruhland, J., 2019. Feasibility Study of Analysis of Senior IT Management Skills/Qualifications in Social Networks, in: In ECSM 2019 6th European Conference on Social Media. Presented at the ECSM, Academic Conferences and publishing limited, pp. 324–334.

Haussen, T., Uebelmesser, S., 2018. No Place Like Home? Graduate Migration in Germany. Growth Change 49, 442–472. https://doi.org/10.1111/grow.12249

Hsiao, C., 2014. Analysis of Panel Data. Cambridge University Press.

Kitchin, R., McArdle, G., 2016. What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets: Big Data Soc. https://doi.org/10.1177/2053951716631130

Kozinets, R.V., 2015. Netnography, in: The International Encyclopedia of Digital Communication and Society.

Krabel, S., Flöther, C., 2014. Here today, gone tomorrow? Regional labour mobility of German university graduates. Reg. Stud. 48.10, 1609–1627.

Lamberti, P.W., Majtey, A.P., 2003. Non-logarithmic Jensen–Shannon divergence. Phys. Stat. Mech. Its Appl. 329, 81–90. https://doi.org/10.1016/S0378-4371(03)00566-1

Lohr, S., 2012. The Age of Big Data (Published 2012). N. Y. Times.
Madden, S., 2012. From Databases to Big Data. IEEE Internet Comput. 16, 4–6. https://doi.org/10.1109/MIC.2012.50
Meyermann, A., n.d. Hinweise zur Anonymisierung von qualitativen Daten 17.
Panel Questionnaire 2009 [WWW Document], n.d. URL https://metadata.fdz.dzhw.eu/public/files/instruments/ins-gra2009-ins2$-1.0.1/attachments/gra2009_W2_Main_PAPI_Questionnaire_de.pdf (accessed 10.15.20).
Panel Questionnaire 2013 [WWW Document], n.d. URL https://metadata.fdz.dzhw.eu/public/files/instruments/ins-gra2013-ins1$-1.0.0/attachments/gra2013_W1_Questionnaire_de.pdf (accessed 11.29.20).
Peffers, K., Tuunanen, T., Gengler, C.E., Rossi, M., Hui, W., n.d. THE DESIGN SCIENCE RESEARCH PROCESS: A MODEL FOR PRODUCING AND PRESENTING INFORMATION SYSTEMS RESEARCH 25.
PLZ Download [WWW Document], n.d. URL https://www.suche-postleitzahl.org/downloads (accessed 12.2.20).
Shafique, U., Qaiser, H., 2014. A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA) 12, 6.
Singh, A., 2013. Is Big Data the New Black Gold? Wired.
State of social platform use in Germany in 5 charts, 2017. . Digiday. URL https://digiday.com/marketing/state-social-platform-use-germany-5-charts/ (accessed 7.7.20).
Studienanfaenger-Absolventen.pdf [WWW Document], n.d. URL https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2005/2005_10_01-Studienanfaenger-Absolventen-2020.pdf (accessed 11.29.20).
Tukey, J.W., 1977. Exploratory data analysis, Addison-Wesley series in behavioral science. Addison-Wesley Pub. Co, Reading, Mass.
Volluniversität, 2020. . Wikipedia.
Wirth, R., Hipp, J., n.d. CRISP-DM: Towards a Standard Process Model for Data Mining 11.
Zikopoulos, P., 2012. Understanding big data: analytics for enterprise class Hadoop and streaming data. McGraw-Hill, New York.

3.3 Publikation 3: „Trusting Big Data Analytics Process from the Perspective of Different Stakeholders“

Während in den ersten beiden Untersuchungen der Kreislauf der Wissensgeneration untersucht wurde, geht die folgende Publikation der Frage nach, welche Faktoren für das Vertrauen in die Ergebnisse der Wissensgenerierung durch Data Mining aus Sicht der unterschiedlichen Beteiligten, den Stakeholdern, wichtig sind. Der Begriff Vertrauen ist im täglichen Sprachgebrauch weit verbreitet, es zeigt sich jedoch, das eine konkludente wissenschaftliche Definition schwierig ist und von den einzelnen Wissenschaftsdisziplinen durchaus unterschiedlich formuliert wird. Auch wenn je nach Disziplin der Fokus unterschiedlich gesetzt wird und die Terminologie nicht einheitlich ist, wird doch ein gemeinsamer Kern im Verständnis erkennbar. Als methodischer Rahmen wurde wiederum die im Data Mining weit verbreitete Methode CRISP DM genutzt. Inspiriert vom COBIT Konzept und der IT Governance wurden typische Stakeholder indentifiziert. Neben den offensichtlich Beteiligten, den Datenanalysten (aka „Auswertungsspezialisten“) und Business Usern (aka „Fachgebietsspezialisten“), wurden die Projektsponsoren (aka „Auftraggeber“) der Data Mining Studien sowie die Konsumenten (aka „Nutzer“/ „Betroffene“) des Data Mining Projektes befragt. Dabei wurde soweit möglich, auf Experteninterviews zurückgegriffen. Um dem umfangreichen Potential der Antworten gerecht zu werden, wurden sowohl offene als auch semi-strukturierte (Leitfadengestützte) Interviews geführt. Basierend auf diesen Interviews wurden die Antworten qualitativ analysiert und die wesentliche Einflußfaktoren identifiziert. Dabei wurde eine Informationsasymetrie zwischen den Anspruchsberechtigten (engl. „Stakeholdern“) in den einzelnen Phasen der Datenanalyse identifiziert. Die Arbeit zeigt, das Transparenz für alle Phasen des Data Mining wichtig ist, jedoch die bisherigen Ansätze dem nur bedingt gerecht werden. Wichtige Kennzahlen oder Aussagen, die den primär für die einzelnen Stufen des CRISP DM Modells verantwortlichen Stakeholdern Hinweise auf die Qualität der Artefakte dieser Stufe liefern, werden von anderen Stakeholdern nur bedingt verstanden. Dies erschwert die Vertrauensbildung bzw. den Transfer von Vertrauen zwischen den beteiligten Stakeholdern und damit insgesamt das Vertrauen in die Ergebnisse der Datenanalyse. Insgesamt zeigten sich Visualisierungen gegenüber Kennzahlen für das Verständnis vorteilhafter. Wobei wiederum simple gegenüber komplexeren Darstellungen bevorzugt wurden je weiter die beteiligten Stakeholder inhaltlich von ihrem eigentlichen Fachgebiet entfernt sind.

Im Ergebnis lässt sich konstatieren, dass bestehende Frameworks, wie das hier exemplarisch verwendete CRISP DM, weiterentwickelt werden sollten, um den vielfältigen Ansprüchen der

Stakeholder, deren heterogenen Fachexpertisen und der daraus resultierenden Informationsasymetrie gerecht zu werden.

Trusting the Data Analytics Process from the Perspective of Different Stakeholders

Sven Gehrke, Sandra Niemz, Johannes Ruhland
Chair of Business Informatics,
Friedrich-Schiller-University,
Jena, Germany
e-mail: sven.gehrke@uni-jena.de, sandra.niemz@uni-jena.de, johannes.ruhland@uni-jena.de

***Abstract*—The paper at hand shows different aspects of the concept of trust. Using the Cross Industry Standard Process for Data Mining (CRISP-DM) phase model, we stress the information asymmetry of the typical stakeholders in a data mining project. Based on the identified influencing factors in relation to trust, problematic aspects of the current approach are verified. We execute various interviews with the stakeholders and the results of the interviews confirm the theoretically identified weak points of the phase model with regard to trust. Based on the finding, we sketch amendments and future research areas.**

Keywords-trust; data mining; CRISP DM; stakeholder management.

I. MOTIVATION

Big data analyses take up an ever-larger part of our lives or influence them indirectly. This can be derived from the number of scientific publications [1], which can be assessed as an indicator of the researchers' interest in the subject. Another indicator is the trend in Internet searches on the subject, e.g., for the term "Data Science" [2], which not only reflects academic interest, but also a broader public interest. In addition to the theoretical interest, the number of mass market products that are essentially based on big data analyses has been increasing for years [3]. The same interest can be seen for the term “social media” [4].

Data mining algorithms are largely based on heuristics, i.e., finding probable solutions with limited knowledge and time. This goes hand in hand with probabilities and trust. While there is an extensive literature canon on trust in general - differences and similarities in trust in general and in specific technology - the interactions between people who request, create, operate and use a specific data mining application with regard to trust and its elements have been little explored [5]. The present paper reports on various studies of the relationship between big data analyses and, in particular, their representation and trust depending on the stakeholders involved. Based on the interviews with major stakeholders of the data mining process, the paper points out open issues and challenges found during the survey.

The rest of the paper is structured as follows. In Section 1, we provide an overview on standard data-mining procedures and, based on a literature review, we examine different concepts on trust depending on the field of study. Both these concepts – data-mining methodology and the components influencing trust – are then linked together. In Section 4, we present the interview results from several identified main stakeholders. Since the concepts behind the survey were not equally known by all those involved, mainly semi-structured interviews were chosen. The results were analyzed qualitatively using Mayring's approach [6]. In Section 5, we give a conclusion and list the main findings.

II. LITERATURE REVIEW

A. Big Data Analytics (BDA) / Data Mining (DM)

Big data analyses are not just about the underlying data and the resulting analyses but, as with other information systems, about the organization of the processes and organizational views [7]. This requires the presentation in a holistic end-to-end model that connects and coherently maps the individual elements. Phase models are traditionally used in project management [8] [9], while process models, such as Business Process Model and Notation (BPMN) [10] have established themselves to map the various aspects involved.

Regardless of the learning type or the methods, the Framework Cross Industry Standard Process for Data Mining (CRISP DM) has established itself as the standard procedure in data mining projects [11] [12]. The model describes the basic sequence of individual phases, the relationships between one another (see Figure 1) and the tasks contained therein (see Figure 2).

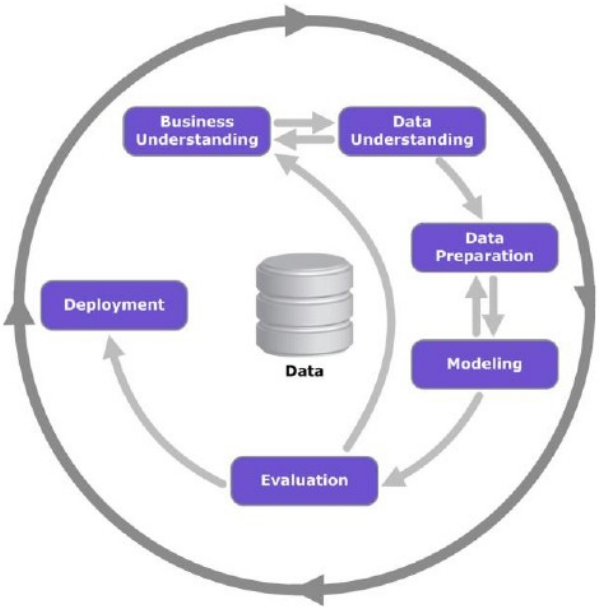

Figure 1. CRISP DM model [12 p. 5].

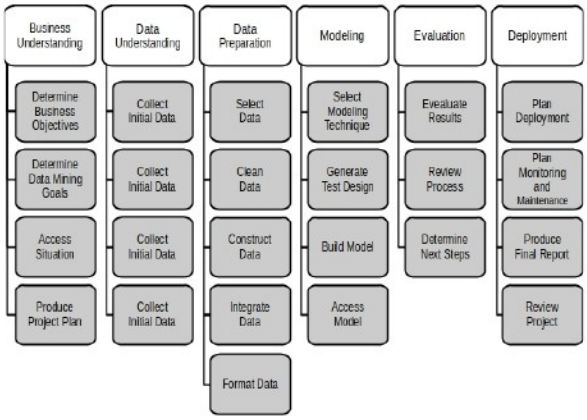

Figure 2. CRISP DM - detailed phases/ tasks [12 p. 6].

The organizational view deals, among other things, with the balancing of interests and information of the stakeholders involved. Only if all those involved find their needs taken into account and respected, they will use the results of the analysis. Obviously, an elementary component like trust has to be treated at every step of the process and cannot be added after the fact. A transfer of trust between different stakeholders is, therefore, necessary in many process steps for trust in the result of the data analysis

Interestingly, typical tasks are listed and described, but no stakeholder identification or explicit role assignment is provided for the tasks in the framework. Accordingly, there is also no consideration of the relationship of trust between the individual stakeholders. As part of this paper, in addition to the standard model, the phases within the framework of a stakeholder analysis, the typical roles involved in a RACI matrix and the information flow were analysed (see also [13] [14]). This is certainly open to discussion in detail, but several things become clear. In the individual phases, there is an information asymmetry to be overcome between the stakeholders and thus a situation of trust in the above sense. Thus, not only is the algorithm itself afflicted with probabilities and thus (trust) risks, the roles involved must also build trust among each other in order to resolve the information asymmetry - and this at different times and in different directions. If data mining is used purely in-house, the business user represents the users or consumers of data mining in the company and can manage the transfer of information and trust. If, however, the consumer of data mining is the general public, it is necessary for the “business user” to put themselves in the most varied of perspectives in the best possible way and to create the basis for the transfer of trust for all stakeholders not directly involved. To make matters worse, the general public is very heterogeneous - both in their personal attitudes and experience, as well as in their institutional environment.

The business user should take the perspective of all major representatives and anticipate and manage their expectations. Looking at the model, (see Figure 1) it becomes clear that this management has to be taken into account both when considering the origin of the data and when communicating the evaluation results / key figures.

TABLE 1. ROLE OF STAKEHOLDERS DESCRIBED BY RESPONSIBILITY ASSIGNMENT MATRIX (RACI)

	Project Sponsor (PS)	Business User/ Analyst (BA)	Data Analyst/ Scientist (DA)	Information Ownership/ Flow
Business Understanding	**a**	**r**	c	PS → BA → DA
Data Understanding	**a**	c	**r**	BA/ DA
Data Preparation/ Modeling	**a**	i	**r**	DA
Evaluation	**a**	**r**	c	DA → BA → PS
Deployment	a	r	c	PS

a = accountable; = responsible; c = to consult; i = to inform

B. Acceptance and Trust

Trust is a complex concept that is defined differently in numerous disciplines depending on the specific circumstances. An additional complicating factor is that trust is also used in everyday scenarios, which results in a multitude of meanings without any reference to a concept. In a much-cited 1964 standard by Kaplan, the author goes so far as to recommend that researchers focus on a specific component of trust rather than a generalized view [15]. If, in addition to the use of the term in one language, one also considers the translation into other languages, there is a large number of uses and synonyms with clear deviations in the connotation. Cooperation, confidence and predictability are closely related terms that Mayer et al. used to describe the term in English [16 p. 729]. Trust is the basis for accepting vulnerability from people, technology and, in our case, the use of data analysis results [17].

To measure acceptance of an application or technology in business informatics, the Technology Acceptance Model (TAM) is often used [17] [19].

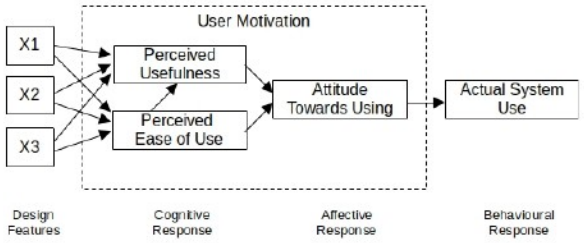

Figure 3. Technology Acceptance Model [41 p. 24].

In Figure 3, the "Attitude Towards Using" is the readiness for use, which is influenced by "Perceived Usefulness" and "Perceived Ease-of-Use". The “Perceived Usefulness” describes the expected benefit, while the “Perceived Ease-of-Use” describes the costs for the user to learn how to use the technology and thus indirectly the costs of building trust. Due to its simplicity, it is easy to use and popular. However, the model focuses on the user and the lack of consideration of the situation / structure is often criticized [20 p. 30]. Furthermore, it does not take time into account and, therefore, a separation of initial trust and continued trust is not explicitly described in the model.

Closely related concepts to trust come from psychology, sociology and social psychology, the latter being understood as a bridge between the former. While psychology focuses on the person-to-person relationship, sociology focuses on the organization-to-organization relationship. What they all have in common is to define trust as "the willingness to take risks" [21 p. 103] or the "intention to accept vulnerability" [22 p. 395]. Basically, Mayer, Davis and Schoorman show that both the personal influence and the organizational or institutional influence of taking risks can be characterized on the basis of competence, benevolence and integrity [16].

After an intensive comparison of the literature, McKnight and Chervany succeeded in bridging the gap between the above-mentioned disciplines and showing the interdependencies [23]. The authors separate between *trusting believes* as the extent to which a target is likely to behave in a way that is "benevolent, competent, honest, predictable in a situation" and the *trusting intentions* as the extent to which a person is willing to make himself vulnerable to another person's actions [23].

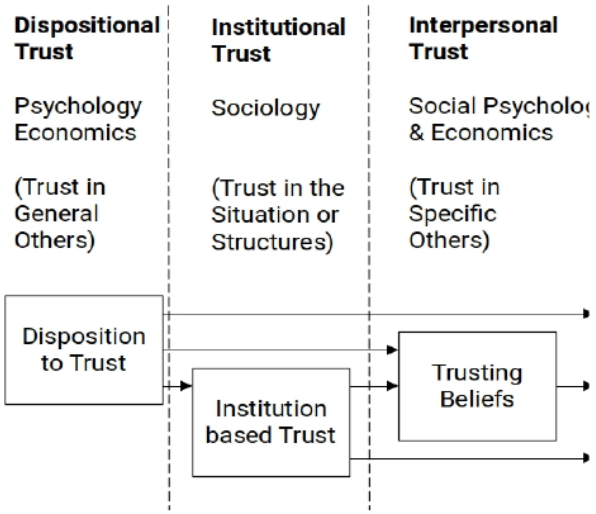

Figure 4. An interdisciplinary Model of High Level Trust Concepts [23]

In a further work, McKnight et al. show that the original characteristics (aka trusting beliefs) of trust in people can be transferred to trust in technologies (functionality, helpfulness, reliability) [5 p. 9]. Thus, the original concept is not limited to natural persons, but can be transferred to people, technologies / objects and even processes [24]. While Figure 4 links different research fields and explains influencing factors from the personal disposition to the trusting intentions, it does not take into account the dimension „time". While continued trust and trusting intentions result from experience and, therefore, the balance of incentives and penalties resulting from trusting, initial trust results from trust transfer - either from person, groups or places [25]–[27].

The microeconomic theory as a further subject area investigates trust in its own branch of research - information theory. It offers a complementary model to the explanatory approaches above. Here, the focus is on trust in goods and the costs of evaluating their properties, less on individual disposition. The underlying assumption is that the information market does not exhibit high degrees of transparency. That is, to evaluate the information, the information must be known, so one has to invest in learning it to evaluate it [28]. In principle, a distinction is made between three types of goods: search goods, experience goods and credence goods. Search goods can be evaluated before purchase or use and, therefore, trusted due to previous experience or easily available product information. Search goods are well known and represent continued trust. Experience goods can be evaluated only after purchase and, therefore, trusted after the purchase and need either a transfer of trust or reduced penalties (e.g., a „refund policy"). Credence goods cannot be evaluated due to prohibitive information retrieval costs or singularity and depend always on external trust transfer [29] [30]. This links the model nicely to the initial model: search goods have a strong linkage towards trust transfer and/ or previous experienced trust, experience goods need initial trust and a positive experience balance for continued usage, and credence goods cannot personally be evaluated over time at all and depend entirely on trust transfer.

From their perspective, all the models presented above are justified and complement each other. The technology acceptance model focuses heavily on the acceptance of a technology without addressing explicitly the wider trust aspect. However, McKnight's model explains the impact of trust in technology and links personal and environmental attitudes, while the microeconomic model deepens insights into how costs affect trust transfer.

C. Linking CRISP DM and Trust

There is only very limited amount of literature about what influences trust in data mining analyses. If one looks at the underlying knowledge and skills of the individual stakeholders in the individual phases of the CRISP DM model (see Figure 1), they will see a strong information asymmetry. To make matters worse, it is noticeable that the responsibility changes in the phases (see Table 1), and external information interests are not explicitly in the focus of the model. This is associated with relatively high costs for obtaining information. Thus, the next step is to evaluate the current practical approaches and problems by interviewing the stakeholders involved.

There are two basic approaches to understand data - compressing the information into key figures/ metrics and visualizing it in graphics. Measures have been around since the dawn of mathematics and are widely used in a wide variety of scientific fields. visualisations of data are just as old as key figures, but have been experiencing increased interest since the 1970s, beginning with Tukey's "Exploratory Data Analysis" [31] and the "Box Plots". Currently, the increased computing power enables complex representations of high-dimensional data and leads to innovative and highly complex forms of representation, e.g., t-distributed Stochastic Neighbor Embedding (t-SNE) [32]. While t-SNE or Uniform Manifold Approximation and Projection (UMAP) help to understand the data directly, graph based visualisation helps to understand hierarchies and dependencies between data or Key Performance Indicators (KPIs) [33]. As an independent specialist discipline, however, visualisation is quite young [34].

Both approaches are to be viewed critically: the key metric α (significance level) used most frequently in statistics and the corresponding p-value (probability) are so often misinterpreted in the scientific literature that the American Statistical Association opposed in 2018 another use of the term "significance" pronounced [35] [36]. On the other hand, visualisations are not problem-free either. A sub-goal of visualisation is to increase the perceived and cognitively processed amount of information and to capture interdependencies in the data structures through aggregation and emphasis. This is intended to support the correctness of decisions and confidence in the decision [37]. Thus, visualisations are not neutral either and depend on the ideas of the creators [38].

Recently, to overcome the aforementioned issues, there is a trend to combine key figures/ metrics with visualisations or allow using interaction in the visualisation. In order to develop a balanced strategy across all critical goals and their respective KPIs, it is particularly important to discover the inherent relationships between all KPIs. In this case, graph-based representations are particularly suitable [39] [40].

D. Identified Research Area

Freedom offered by modern societies, access to loads of sources of information and increased complexities force people to cope with the uncertainty of the global world by themselves. If one considers the factors presented that are important for the trust of the individual stakeholders and if one also considers the CRISP DM phase model, it becomes apparent that the currently leading framework does not explicitly pay attention to the transfer of trust between the stakeholders involved. It should also be clear that trust must be observed from the beginning to the end - interrupting the analysis process would disrupt the transmission of trust. It is important to consistently monitor the level of knowledge of all persons involved. Since the analysis process cannot be explained personally to everyone, it is important to create a framework that passively enables it. A process-oriented, graph-based visual representation as well as additional, generally understandable visualisations should help to show the connections and thus reduce the information asymmetry between the stakeholders. This should be practically substantiated in the following summarized interview with the stakeholders (see Table 2).

III. Practical Survey

As in the theoretical model, trust or the transfer of trust depends on the personal environment ("dispositional trust"), the environment ("institutional trust") and specific influencers ("interpersonal trust"). In order to obtain a representative picture, identified stakeholders were interviewed using different survey methods . In addition to the typical stakeholders already identified (see Table 1), the list was expanded to include "normal consumers" ("consumer A") and "informed consumers" ("consumer B").

In the interviews conducted, the aim was to show which aspects - from the stakeholder's point of view - are necessary in order to develop or transfer trust in data analyses. The questionnaire was based on the identified trust influencing factors during all phases of CRISP DM. Consumer were asked to compare results from data mining analyses with previous expert analyses.

TABLE 2: OVERVIEW OF STAKEHOLDERS AND THEIR INTERVIEWS

Stakeholder	Interview Type	Interview Channel
S1: Data Analyst	semi-structured	face-to-face
S2: Business User	semi-structured	face-to-face
S3: Project Sponsor/ Management	semi-structured	face-to-face
S4A: Normal Consumer	semi-structured / closed	telephone
S4B: Informed Consumer	semi-structured	face-to-face

A. S1 Data Analyst Representative

The data analyst has specialist knowledge of the technical analysis of data, its consistent preparation and the use of appropriate statistical or data mining methods. He needs information about the data used and the business objective of the analysis.

Interview

For the interview, 3 data scientists were asked independently of each other which indicators they believe are relevant in order to trust the data and models. Then they were presented with various business KPIs and visualisations of their department together with the respective business representative and the similarities and differences in understanding were determined.

Main concerns and issues

In the business understanding, it was difficult for the stakeholders involved to interpret the specific KPIs. Concrete examples and the representation of the processes through graphics were essential for understanding.

The interviews showed that the KPIs used to justify the analysis results were rarely understood or misunderstood.

In principle, graphic representations were preferred by the other stakeholders involved. More complex representations were accepted, but required more detailed descriptions, and here again the data analysts often struggled with the business terms. As a compromise for understanding, several simple graphics that build on one another were used.

B. S2 Business Representative

The business analyst represents the business perspective of the departments and has special knowledge of his department. He alone can make sense of the data and explain their origin and meaning and practically validate the results.

Interview

For the interview, 3 representatives were interviewed independently of each other regarding their intentions during the phases in which they are responsible. Then they were presented with various KPIs and visualisations of their department together with the respective data analysts and the similarities and differences in understanding were determined.

Main concerns and issues

The concerns and issues of data analysts reflect the concerns and issues found among business users.

C. S3 Project Sponsor/ Management Perspective

In addition to the roles of data analyst and business user, which are involved operationally, the project sponsor is another relevant role that is more strategically oriented. His trust in the implementation and the results decides initially and finally on the resources used and the use of the results. As a managing role, which is not directly involved but is regularly informed about the analysis and the results, his trust can be seen as the first test of trustworthiness. This role is also responsible to a not inconsiderable extent for the design of the environment, ergo it exerts a great influence on the "institutional trust".

Interview

For the analysis, 10 senior IT managers were asked about their criteria for building trust in a guided interview. The following section summarizes the answers and the underlying intentions.

Main concerns and issues

Looking at the key considerations and underlying intentions, the focus is clearly on promoting institutional trust rather than understanding individual BDA and its metrics. The interviewees emphasized that building a high-quality and transparent data infrastructure is essential for trust in the results. There were different opinions as to whether this should be done step-by-step or with a “big bang”. While the majority emphasizes the "step-by-step" approach and thus the step-by-step understanding of the data, a minority fears that too narrow a focus will limit the reference power of the data too much. All project sponsors emphasize that a common understanding is important. To achieve this goal, KPIs, a commonly understood language - which also includes visualisation, are used. For the most part, however, the internal stakeholders are taken into account, but the perspective of the external stakeholders is primarily included through reference to legal data sources.

D. S4 Consumer

When analysing the consumer, a distinction must be made between two stakeholders. The difference lies in the existing experience with analyses. One group are the consumers who have no experience with data mining analysis, data and procedures (see Table 2: “S4A: Normal Consumer”). On the other hand, there are consumers who were not directly involved in the analysis but have personal experience with similar data mining projects (see Table 2: “S4B: Informed Consumer”).

Consumers use the results directly for their own purposes, e.g., fitness wearables or evaluations in magazines. However, consumers can also be indirectly affected by the results of the analysis, e.g., as bank customers who are subject to a risk classification when requesting a loan.

E. S4A: Normal Consumer

Interview

In a semi-structured interview, 23 people between the ages of 20 and 60 were asked which factors are relevant for them in different contexts in order to trust data mining analyses.

Main concerns and issues

The results of the data mining were accepted to a very limited extent. Without a well-founded justification for the refusal, it was doubted that the data were representative and reflected the personal circumstances. Although trust was positively influenced by the spread of the BDA (e.g., wearables/ web portals) and by certificates, the results are doubted by 70% - 80% of the respondents and used personally. When it comes to acceptance, the personal opinion of a specialist or friends prevails. If trust has arisen through the transfer of trust from third parties, the trust is not shaken by isolated negative examples or experiences. In principle, the respondents do not see themselves in a position to validate the data bases and functionalities and need support from their environment.

F. S4B: Informed Consumer

Interview

In the interviews, three scientists were questioned in a semi-structured interview. They were not actively involved in the analyses, but they were familiar with the environment.

Main concerns and issues

The expert survey revealed that these people generally trust the analyses, but inform themselves about the data collection, data processing and methods used on a random basis. A renowned environment of the BDA reduces the scope of own validations, but is not sufficient.

IV. Conclusion

In principle, the results of the data mining process are accepted by the stakeholders involved in the analysis, but trust in the results correlated strongly with the proximity to the process and the associated costs of information procurement.

In the business and data understanding, the business and data analyst representatives attached great importance to understanding the data and assessing its quality. Less value was placed on a detailed verbal description, the focus was more on use cases and easily understandable key figures.

While the specialists tend to orientate themselves towards the key figures of their specialist area during the evaluation, the other stakeholders involved prefer visual representations in addition to the key figures. Key figures are accepted without understanding their meaning in particular. In order to understand the statements and to trust the results, visual representations prevail. There, too, a trend towards rather simple, well-known representations was discernible. For example, a combination of histogram and box plot was preferred to a violin plot.

The preparation of the analysis process for later users was less of a focus. A fundamental desire was established among all stakeholders to present their findings or interests transparently. However, they often do not realize that the technical terms that describe their special field are not universally understandable. Thus, in data science projects, the focus should be on a more understandable language in advance. In particular, the visualisation seems to have a greater influence on the overall understanding than on specific key figures.

Based on the findings, it would certainly be helpful to add a stakeholder-oriented view to the CRISP DM framework. It should be essential to meet both the information needs of the specialists and to balance the

information asymmetry among the stakeholders. In addition to specific, subject-related key figures, this view should be integrated into the CRISP DM process as well as simple visualized representations and generally accessible key figures. This view should also depict the chronological sequence and, so to speak, represent the information transformation nose to tail. In order to do justice to the different levels of knowledge, it should be able to depict different levels of detail.

References

[1] Y. Zelenkov and E. Anisichkina, "Trends in data mining research: A two-decade review using topic analysis," *Bus. Inform.*, vol. 15, no. 1, pp. 30–46, Mar. 2021, doi: 10.17323/2587-814X.2021.1.30.46.

[2] "Google Trends 'Data Science,'" *Google Trends*. https://trends.google.de/trends/explore?q=%2Fm%2F0jt3_q3&date=all (accessed Aug. 15, 2022).

[3] "Google Trends 'Fitness App,'" *Google Trends*. https://trends.google.de/trends/explore?date=all&q=%2Fg%2F11cny0zppc (accessed Aug. 15, 2022).

[4] "Google Trends 'Social Media,'" *Google Trends*. https://trends.google.de/trends/explore?date=all&q=social%20media (accessed Aug. 15, 2022).

[5] D. H. Mcknight, M. Carter, J. B. Thatcher, and P. F. Clay, "Trust in a specific technology: An investigation of its components and measures," *ACM Trans. Manag. Inf. Syst.*, vol. 2, no. 2, Art. no. 2, Jun. 2011, doi: 10.1145/1985347.1985353.

[6] P. Mayring, "Qualitative content analysis," *Companion Qual. Res.*, vol. 1, no. 2, pp. 159–176, 2004.

[7] A.-W. Scheer and K. Schneider, "ARIS — Architecture of Integrated Information Systems," in *Handbook on Architectures of Information Systems*, P. Bernus, K. Mertins, and G. Schmidt, Eds. Berlin/Heidelberg: Springer-Verlag, 1998, pp. 605–623. doi: 10.1007/3-540-26661-5_25.

[8] G. Garel, "A history of project management models: From pre-models to the standard models," *Int. J. Proj. Manag.*, vol. 31, no. 5, pp. 663–669, Jul. 2013, doi: 10.1016/j.ijproman.2012.12.011.

[9] P. D. de M. Sánchez, C. G. Gaya, and M. Á. S. Peréz, "Standardized Models for Project Management Processes to Product Design," *Procedia Eng.*, vol. 63, pp. 193–199, 2013, doi: 10.1016/j.proeng.2013.08.176.

[10] M. Chinosi and A. Trombetta, "BPMN: An introduction to the standard," *Comput. Stand. Interfaces*, vol. 34, no. 1, pp. 124–134, Jan. 2012, doi: 10.1016/j.csi.2011.06.002.

[11] P. Chapman *et al.*, "Step-by-step data mining guide," *SPSS*, p. 76, 2000.

[12] R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data mining," in *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, 2000, vol. 1, pp. 29–39.

[13] C. J. Costa and J. T. Aparicio, "POST-DS: A Methodology to Boost Data Science," in *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)*, Sevilla, Spain, Jun. 2020, pp. 1–6. doi: 10.23919/CISTI49556.2020.9140932.

[14] L. Wehrstein, "CRISP-DM ready for Machine Learning Projects," *Medium*, Dec. 19, 2020. https://towardsdatascience.com/crisp-dm-ready-for-machine-learning-projects-2aad9172056a (accessed Jun. 11, 2021).

[15] A. Kaplan, *The conduct of inquiry: methodology for behavioral science*. Scranton (Pa.): Chandler, 1964.

[16] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An Integrative Model Of Organizational Trust," *Acad. Manage. Rev.*, vol. 20, no. 3, pp. 709–734, Jul. 1995, doi: 10.5465/amr.1995.9508080335.

[17] B. A. Misztal, "Trust: Acceptance of, Precaution Against and Cause of Vulnerability," *Comp. Sociol.*, vol. 10, no. 3, pp. 358–379, 2011, doi: 10.1163/156913311X578190.

[18] F. D. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *MIS Q.*, vol. 13, no. 3, p. 319, Sep. 1989, doi: 10.2307/249008.

[19] N. Marangunić and A. Granić, "Technology acceptance model: a literature review from 1986 to 2013," *Univers. Access Inf. Soc.*, vol. 14, no. 1, pp. 81–95, Mar. 2015, doi: 10.1007/s10209-014-0348-1.

[20] B. Lunceford, "Reconsidering technology adoption and resistance," *Explor. MEDIA Ecol.*, p. 20, 2009.

[21] N. Luhmann, "Familiarity, Confidence, Trust: Problems and Alternatives," p. 10.

[22] D. M. Rousseau, S. B. Sitkin, R. S. Burt, and C. Camerer, "Not So Different After All: A Cross-Discipline View Of Trust," *Acad. Manage. Rev.*, vol. 23, no. 3, pp. 393–404, Jul. 1998, doi: 10.5465/amr.1998.926617.

[23] D. H. McKnight and N. L. Chervany, "What is Trust? A Conceptual Analysis and an Interdisciplinary Model," *AMCIS 2000 Proc. P382*, p. 8, 2000.

[24] J. B. Thatcher, M. L. Loughry, J. Lim, and D. H. McKnight, "Internet anxiety: An empirical study of the effects of personality, beliefs, and social support," *Inf. Manage.*, vol. 44, no. 4, pp. 353–363, Jun. 2007, doi: 10.1016/j.im.2006.11.007.

[25] D. Belanche, L. V. Casaló, C. Flavián, and J. Schepers, "Trust transfer in the continued usage of public e-services," *Inf. Manage.*, vol. 51, no. 6, pp. 627–640, Sep. 2014, doi: 10.1016/j.im.2014.05.016.

[26] D. Harrison McKnight, V. Choudhury, and C. Kacmar, "The impact of initial consumer trust on intentions to transact with a web site: a trust building model," *J. Strateg. Inf. Syst.*, vol. 11, no. 3, Art. no. 3, Dec. 2002, doi: 10.1016/S0963-8687(02)00020-3.

[27] K. J. Stewart, "Trust Transfer on the World Wide Web," *Organ. Sci.*, vol. 14, no. 1, pp. 5–17, Feb. 2003, doi: 10.1287/orsc.14.1.5.12810.

[28] B. Kahin and H. R. Varian, Eds., *Internet publishing and beyond: the economics of digital information and intellectual property*. Cambridge, Mass: MIT Press, 2000.

[29] S. Fließ, "Qualitätsmanagement bei Vertrauensgütern," *Mark. ZFP*, vol. 26, no. Sonderheft 2004, pp. 33–44, Jan. 2019, doi: 10.15358/0344-1369-2004-Sonderheft-2004-33.

[30] K. Mitra, M. C. Reiss, and L. M. Capella, "An examination of perceived risk, information search and behavioral intentions in search, experience and credence services," *J.*

Serv. Mark., vol. 13, no. 3, pp. 208–228, Jun. 1999, doi: 10.1108/08876049910273763.

[31] J. W. Tukey, *Exploratory data analysis.* Reading, Mass., 1977.

[32] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE.," *J. Mach. Learn. Res.*, vol. 9, no. 11, 2008.

[33] M. Graham, J. B. Kennedy, and C. Hand, "A comparison of set-based and graph-based visualisations of overlapping classification hierarchies," in *Proceedings of the working conference on Advanced visual interfaces - AVI '00*, Palermo, Italy, 2000, pp. 41–50. doi: 10.1145/345513.345243.

[34] H. Reiterer, T. M. Mann, G. Mussler, and U. Bleimann, "Visualisierung von entscheidungsrelevanten Daten für das Management. In: HMD, Praxis der Wirtschaftsinformatik," vol. 212, 2000, pp. 71–83.

[35] V. Amrhein and S. Greenland, "Remove, rather than redefine, statistical significance," *Nat. Hum. Behav.*, vol. 2, no. 1, pp. 4–4, Jan. 2018, doi: 10.1038/s41562-017-0224-0.

[36] R. L. Wasserstein and N. A. Lazar, "The ASA Statement on *p* -Values: Context, Process, and Purpose," *Am. Stat.*, vol. 70, no. 2, pp. 129–133, Apr. 2016, doi: 10.1080/00031305.2016.1154108.

[37] J.-A. Meyer, *Visualisierung von informationen*. Place of publication not identified: Springer, 1999.

[38] E. Bussemas, "Mehr als Balken und Torten. Eine experimentelle Befragung zur Wahrnehmung von interaktiven Datenvisualisierungen im Journalismus," *Medien Kommun.*, vol. 66, no. 2, pp. 188–216, 2018, doi: 10.5771/1615-634X-2018-2-188.

[39] M. P. Brundage, W. Z. Bernstein, K. C. Morris, and J. A. Horst, "Using Graph-based visualisations to Explore Key Performance Indicator Relationships for Manufacturing Production Systems," *Procedia CIRP*, vol. 61, pp. 451–456, 2017, doi: 10.1016/j.procir.2016.11.176.

[40] N. Elmqvist *et al.*, "Fluid interaction for information visualisation," *Inf. Vis.*, vol. 10, no. 4, pp. 327–340, Oct. 2011, doi: 10.1177/1473871611413180.

[41] F. D. Davis, "A technology acceptance model for empirically testing new end-user information systems: Theory and results (Doctoral dissertation)." Massachusetts Institute of Technology, 1985.

3.4 Publikation 4: „Decision Support in the Era of Social Media and User-Generated Content“

Die Betrachtung der Kreislaufes der Wissensgenerierung und welche Faktoren das Vertrauen in Data Mining beeinflussen, waren eher theoretischer Natur. Im folgenden Kapitel werden anhand von praktischen Beispielen positive und negative Auswirkungen auf wirtschaftliche Entscheidungsprozesse beleuchtet. Im Ergebnis entstand eine kurze praktische Checkliste, die zeigen soll, dass die Thematik der Arbeit nicht nur aus theoretischer Sicht relevant ist, sondern konkreten praktischen Nutzen liefert.

Decision Support in the Era of Social Media and User-Generated Content

Kathrin Kirchner, Marek Opuszko, and Sven Gehrke

Abstract Social media and the huge amount of user-generated content offer new possibilities for decision-making in companies, as more data can be acquired easily without extra cost. However, a larger database does not automatically lead to better decisions, and volume, variety, and veracity of data from different sources are often overwhelming and challenging. This chapter provides two cases where we used big social data as basis for decision-making. The first case describes the incorporation of extracted information from social media to decision-making models. The second case focuses on the veracity challenges of social media data. By relying on these two cases, we derive guidelines for tackling the veracity of social media data and provide insights how decision-making can be influenced positively and negatively by social media and user-generated content. With the guidelines, it can be determined how much big social data has an influence on quality and rigor in the decision-making process.

Keywords Social media · User-generated content · Decision support · Big data · Veracity · Multi-criteria decision-making

1 Introduction

Making the right decisions has always been a core component of all business activities. The fundamental components of the decision-making process are data, information, and knowledge. For a long time, decision-makers were dependent on information from a few carefully selected sources. In particular, the process of obtaining information was associated with high costs and efforts. In many cases,

K. Kirchner (✉)
Technical University of Denmark, Kgs. Lyngby, Denmark
e-mail: kakir@dtu.dk

M. Opuszko · S. Gehrke
Friedrich Schiller University, Jena, Germany
e-mail: marek.opuszko@uni-jena.de; sven.gehrke@uni-jena.de

J. Papathanasiou et al. (eds.), *EURO Working Group on DSS*, Integrated Series in Information Systems, https://doi.org/10.1007/978-3-030-70377-6_5

it was even impossible to obtain relevant information at all. This situation has changed dramatically in recent years and the emergence of huge online networks and the increasing availability of ever more user-generated data is nothing less than a revolution.

Social media is one of the biggest sources for a variety of big data generated from users, e.g., on public platforms like Facebook, YouTube, Twitter or LinkedIn, or organization-internal platforms. People are enthusiastic to share, interact, and collaborate via these platforms with other users.

Nevertheless, using such social media data for decision-making can be challenging—data have to be carefully selected among the huge amount of available data, and their quality, credibility, and objectivity might be questionable. Furthermore, the question arises if we can use existing decision-making methods. Maybe it is necessary to develop new decision support methods that are able to cope with this specific type and amount of data.

This chapter aims to provide insights on the impact of social media and user-generated content on decision-making. After discussing related work, we present two case studies that illustrate the challenges in using user-generated social media data for decision-making. Finally, we derive recommendations for the identified challenges.

2 Related Work

Information systems in general and decision support systems in particular support individual, group, and organizational decision-making processes. In the last years, social media platforms like social networks (e.g., Facebook, LinkedIn, Yammer) or microblogs (e.g., Twitter) provide opportunities to share knowledge, create new ideas, express opinions, or even integrate employees, business partners, and customers in decision-making processes [1]. These social media platforms also deliver data that can be used as a basis for a data-based decision-making [2].

Data from social platforms can be analyzed on three different levels: (1) interpretation of the whole network, (2) interpretation of groups and components (socio-centered), and (3) interpretation of individual positions (ego-centered) [3]. The analysis of data from social platforms can thus provide a basis for decisions on different levels. The elements for a multi-criteria decision-making can be obtained from social media data. For instance, an analysis of key terms used in a discussion can help to analyze alternatives to solve a problem. Determining the number of times a concept or alternative is mentioned helps to derive preferences for decision alternatives. By analyzing different opinions and sentiments in discussions about a certain topic, group positions about this topic can be estimated.

Social media data play a role, e.g., in social media marketing and consumers' decision-making based on word-of-mouth [4, 5]. In a study among internet users, social media was even influential for 40% of respondents for their decision-making regarding travel, and 74% rely on reviews posted by others [6]. Social media data

also play a role in emergency cases, where real-time information can be used by both authorities and citizens for making safer decisions [7].

Every minute, users produce new content on social media—with every click, like, share, and contribution in different formats. In 2019, more than 500,000 tweets or 4.5 million YouTube videos were posted per minute worldwide [8], which can be named social big data. The term social big data comprises the domains of big data and social media [9].

This huge amount of available data provides new possibilities for analysis and decision-making. The promises of the Big Data age are indeed great: more data, more information, and more knowledge. However, there is no guarantee that more data will automatically result in better decisions. Many researchers therefore point out the challenges in dealing with Big Data [10], especially regarding the so-called information paradox, according to which decision-makers are thirsty for knowledge but drown in information [11]. This is especially true for data generated by users in social media. However, companies are looking for ways to harness the power of this data to improve their decision-making.

Big data is characterized by 3 V: Volume, Velocity, and Variety [12]. Later, three more V were added: Value, Variability, and Veracity [13].

In the context of data analytics in social media, volume, velocity (the high speed of data transfers) and variety (different types of structured and unstructured data) can be critical in big data projects. However, they can be handled by taking samples and using an intelligent experimental design [9]. Furthermore, data frameworks like Apache Hadoop [14] and Spark [15] or algorithms based on MapReduce [16] allow the handling of big amounts of data for decision-making.

In the literature, and from our experiences, veracity is the most critical characteristic of big social data and requires the biggest effort to handle it. Veracity refers to the correctness and accuracy of data as well as to privacy and legal concerns. Lukoianova and Rubin [17] propose the following main subcategories for veracity: Credibility/Implausibility, Deception/Truthfulness, and Objectivity/Subjectivity. Credibility (i.e., believability) is the perceived quality of simultaneously evaluated trustworthiness and expertise. Deception is connected with the verification of the writer's intention to create a truthful impression in the readers' mind, which is also connected with the credibility of the writer. Objectivity refers to the fact that a message can be objective (fully supported or proven) or subjective knowledge (unsupported or weakly supported).

In the following, we will discuss challenges in using social media data for decision-making based on two own cases. The first case studies the decision-making based on a variety of social media data and the second case studies the veracity of this data.

3 Case 1: Incorporating Different Social Media Data into the Decision-Making Process

The introduction of a new software, e.g., an e-commerce software, is a big challenge for many companies. If the software affects many business areas, it is a strategic decision with sometimes far-reaching consequences [18], as the example of the disastrous ERP implementation of MillerCoors shows [19]. For this reason, great attention is paid to the procurement decision. If we divide the procurement process into four phases as described in [20, 21], the analysis and acquisition, in addition to operation and implementation, comprise 50% of the entire process.

The decision-making process can typically be illustrated as shown in Fig. 1. Starting point is the problem definition, in our case the selection of an (enterprise) e-commerce software. In the following, we will mainly limit ourselves to the phases of selecting possible sources of knowledge and evaluating the alternatives since this is where the use of information from social media can come into play.

To illustrate the process within this case study, we assume that five Open Source e-commerce platforms are available, from which an alternative should be chosen at the end of the decision process. These five options are among the most used Open Source e-commerce platforms according to builtwith.com: Magento, Prestashop, OpenCart, VirtueMart, and WooCommerce.

Companies want to avoid introducing software that becomes obsolete or outdated after a few years and then has to be replaced by another software product at great expense. Since very high switching costs can arise, especially with integrated software, an incorrect decision can have disastrous consequences. In addition, the usually strong network effects of software products play a role [22]. For instance, based on total cost of ownership, a high distribution of software might reflect future stability and longevity. Furthermore, the probability of finding trained personnel is also higher if the software is widely distributed and even represents a quasi-industry standard, as is the case with Microsoft Office, for example [23, 24]. On the other hand, network effects can lead to a lock-in and dependence on a single provider. It is precisely this scenario where information from social media and user-generated content can help to examine these questions more closely. This is especially true due to the real-time aspect of social media.

The choice of adequate knowledge sources is of course dependent on the way the subsequent evaluation is carried out. The type of evaluation and thus the problem

Fig. 1 Decision-making process

definition is often adapted to available data sources. The criteria for evaluating software alternatives are numerous and range from functional aspects such as the functional scope of a software, economic effects such as depreciation, network effects, or possible lock-ins to a holistic cost consideration that goes far beyond the directly attributable acquisition costs. For example, Benlian and Hess [25] describe seven criteria for evaluating software alternatives for the purchase of an office suite:

- Functionality
- Reliability
- Cost
- Ease of use
- Ease of customization
- Ease of implementation
- Software support

The evaluation of the functionality of a software is usually of less difficulty. Since features are implemented directly in the development of the software, they can be determined relatively easily [26]. For this purpose, numerous services and websites exist that compare and rank the functional aspects of software products. For example, the Google search for "comparison e-commerce platform" delivers about 77,200,000 search hits (as of 15 March 2020).

However, it is much more difficult to assess fuzzy criteria such as reliability, costs, and sustainability. This is all the more difficult if the manufacturer is not a software development company, but a community that produces open source software (OSS) on a decentralized basis [20, 27]. Especially when estimating costs, serious errors can quickly occur if only directly attributable costs such as pure procurement costs are used. For a long time, there have been methods for a holistic assessment of the investment costs. The total cost of ownership (TCO) approach allows costs to be recorded and evaluated in full. This is particularly important in the procurement of digital goods that are subject to network effects. Here, the recording of all cost types represents a major challenge. Since the main purpose of this case study is to demonstrate the use of data from social media for decision support, we will focus on reliability, costs, and support.

In many business decisions, costs play a significant, if not the most significant, role alongside the benefits of a product. For this reason, special attention will be paid to the decision dimension of costs. In addition to pure acquisition costs, all cost types that arise during the entire life cycle of the product have to be considered. In the context of a product purchase decision such as that of an e-commerce platform, these are the costs of pre-selection, acquisition, use, care, maintenance, backup, and many more. Therefore, costs can be extremely diverse in nature and decision-makers have to collect and evaluate them within the decision-making process.

This becomes more serious if open source products are among the alternatives for software procurement. OSS has long since reached a level of maturity that makes its use in an enterprise environment possible [27]. In the past, however, the focus has usually been on the criteria for purchasing proprietary software [28]. Only recently

has the criteria for the selection of OSS been examined more closely [25]. This is of great importance because OSS promises cost reduction through obtaining a free copy, yet the majority of the total costs of ownership will arise elsewhere, for instance in staff costs. In addition, global communities can dissolve at any time, so there is no guarantee of support or bug fixes [29]. These risks and costs have to be factored into the purchase decision, and the communities must be adequately assessed. In the context of OSS and TCO assessment, other cost factors are as follows:

- Distribution of the software platform
- Insufficient level of knowledge of the users
- Insufficient maturity of the platform
- Lack of community support/activity
- Poor knowledge base
- Lack of available skilled workers

In order to weigh up influencing factors and reach a decision in a structured manner, a wide variety of methods can be used. A well-known procedure is the analytic hierarchy process (AHP) [30, 31]. In this decision support procedure for complex decisions, data is collected, compared, weighted, and processed in several phases until a decision is reached. In the absence of historical data or studies, the data can also be derived from the assessment of experts [32, 33] or user studies [25]. Within the AHP, the criteria matrix can be supplemented by information from social media and enters the decision matrix of an AHP or a cost-utility analysis as further criteria. The advantage is that the process of decision support remains unaffected and can be carried out as before. Figure 2 shows a snapshot of a possible AHP structure with goal, criteria, and alternatives for the present case study.

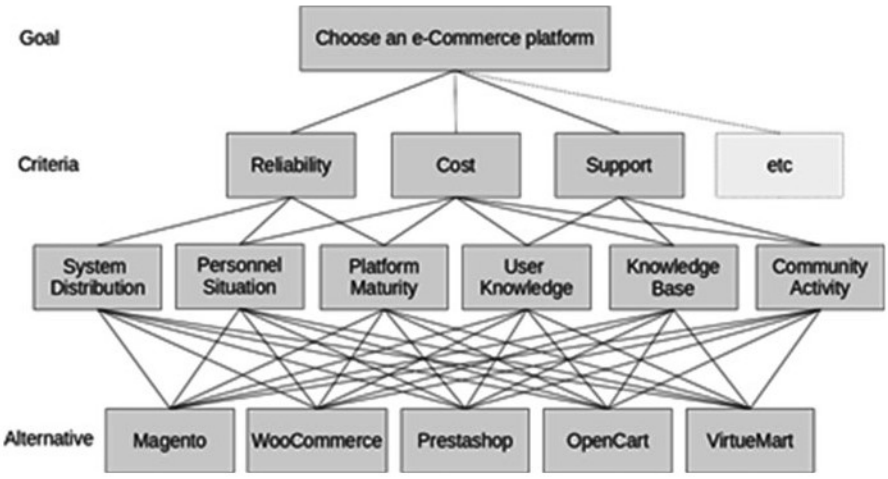

Fig. 2 Exemplary structure of the AHP

In the present case, we focus on the criteria of reliability, cost and support, and subsequential criteria identified based on a TCO approach. The next step is to determine how the criteria can be evaluated. The question arises which indicators exists to measure and assess the criteria and which sources of knowledge can be used to measure these criteria. In a first step, indicators are identified which may reflect the questions or the criteria. For example, the distribution of the software platform can be approximated by the search interest on search engines over time. Other indicators are the number of live systems that use a particular shop system. It becomes clear that these indicators are context dependent and can vary greatly for each decision problem. Therefore, the skill and experience of the decision-maker is required. Once all indicators are defined, the search for possible data sources begins. For example, services such as Google Trends can be used to record online search interest. The number of live systems are shown on websites such as builtwith.com that use a search engine not unlike crawler to search websites worldwide for information about the software used. An interesting question, especially with OSS, is the evaluation of the activity of the community and thus the sustainability of the platform. Here, possible indicators are the general activity of the community, which can be determined by the intensity of communication in message boards and forums.

For this purpose, the online forum Stackoverflow, an internet platform aimed at software developers, is suitable. On this platform, users and developers can exchange information, ask questions, and get answers. The intensity of the number of questions and answers can therefore provide information about the general activity of a community. Furthermore, the quota of help can also be determined by analyzing the answers.

Table 1 shows indicators assigned to the criteria and possible sources of knowledge where relevant information can be found. These too must be collected and structured within the decision-making process and must be context dependent. Fortunately, today there are possible sources of knowledge for almost every question. These results can then act as supplementary information from social media and enter the decision matrix of an AHP or a cost-utility analysis as further criteria. The advantage is that the process of decision support remains unaffected and can be carried out as usual. For reasons of clarity and space limitations, we will limit ourselves to describing the bold indicators and knowledge sources in Table 1 in this case study.

3.1 Results: Distribution of the Systems

Figure 3 shows the search interest of the five alternatives over the last 4 years in the form of search queries to the search engine Google. A decrease in the overall search frequency can be clearly seen. Above all, the longstanding top-ranked shop system Magento is clearly losing interest in searches suggesting that interest is declining and which points to an increasing sustainability risk. If the trend continues and Magento loses interest and distribution in the future, there is a risk of costs

Table 1 Criteria, indicators, and knowledge sources as basis for decision support

Criteria	Indicator	Knowledge source
Distribution of the system	**Online Searches**, Downloads, **Live Websites**	**Google**, Bing, **builtwith.com**
Level of knowledge of the users	Members	LinkedIn, XING, Stackoverflow
Maturity of the development	Release Status	Platform Website, GitHub, SourceForge
Community support/activity	**Tutorials, Videos**, Threads, Questions, Answers, Bugfixes, Release History	**YouTube, Stackoverflow**, Twitter
Knowledge base	Forums, **Tutorials, Courses**	**Udemy, YouTube**, Stackoverflow
Available skilled workers	Employment Service Platforms	LinkedIn, XING, Stackoverflow

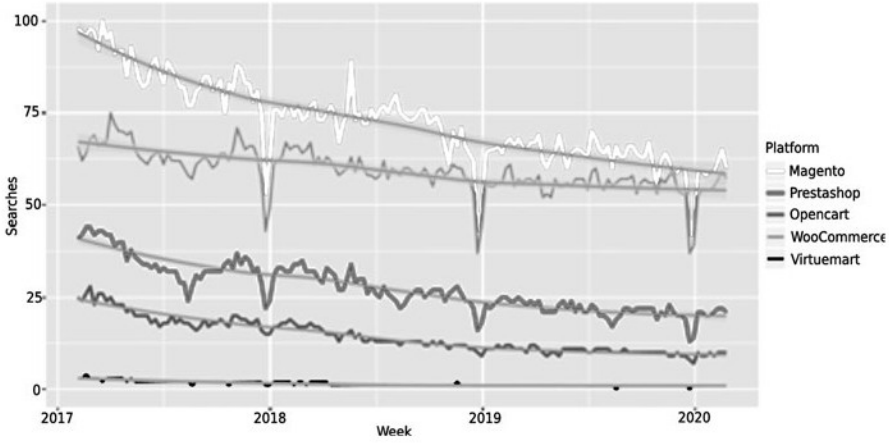

Fig. 3 Search for the five alternatives over the last 4 years in Google search engine

associated with a changeover to another system in the future. In contrast to Magento, the WooCommerce system in particular is showing a constant high level of search interest.

What also becomes clear in the diagram is a generally declining interest in on-premise systems. This also represents a high-cost risk for the future, as cloud-based e-commerce platforms may be the means of choice in the future. However, in the present scenario we assume that a self-hosted solution is a prerequisite, due to the higher flexibility and customizability.

Another key figure is the number of systems installed worldwide. Table 2 illustrates the wide range in terms of the distribution of installed systems. WooCommerce shows with almost four million live websites an absolute top position. This can be associated with a high spread of knowledge about how to use the system. It is

Table 2 Distribution of e-Commerce platforms

Platform	Live websites[a]	6 Month trend	Points	Relative
Magento	190,731	Falling	2	0.05
WooCommerce	**3,876,748**	**Rising**	**5**	**1.00**
Prestashop	250,603	Rising	3	0.06
VirtueMart	56,768	Falling	1	0.01
OpenCart	337,025	Stagnating	4	0.09

[a]According to builtwith.com (accessed 2020-03-02)

Table 3 Indicators on stackoverflow.com

Platform	Threads	Answers per question	Average votes per question	Points	Relative
Magento	**37,666**	**1.37**	**1.09**	**5**	**1.00**
WooCommerce	22,641	1.01	0.8	4	0.60
Prestashop	4835	1.22	0.51	2	0.13
VirtueMart	723	1.14	0.42	1	0.02
OpenCart	5026	1.22	0.49	3	0.13

in turn is a cost factor in the operation of such a system as well as for the recruitment of suitable specialist personnel. The last two columns "Points" and "Relative" show a simple ranking with points and a rating based on the percentage relative to the best result. These values are exemplary evaluations and can be used later in a cost-utility matrix or AHP to evaluate the alternatives.

Table 3 shows an example of the results of the ratios on Stackoverflow. Here, the number of threads, responses, and ratings of the contributions to the respective e-commerce platform were measured. Points and relative points are also listed here as an example of evaluation.

The diagram of the course of the aggregated response frequencies in Fig. 4 illustrates this graphically. It also shows the decreasing response frequency for many e-commerce platforms except WooCommerce. Magento shows the highest overall number of questions and answers, which is displayed in Table 3. However, the graph in Fig. 4 shows a clear trend change in recent months.

Table 4 shows the measured key figures in the area of knowledge base and community support in terms of online courses and tutorials on the platforms Udemy and YouTube. Here, too, a diverse picture emerges. WooCommerce, for example, shows a very large view interest on YouTube with almost 12 million views in comparison with VirtueMart with a view count of 52,184. The data from YouTube was collected via YouTube's own API. For the information from Udemy, a web crawler was used. The evaluation of points and relative points would be analogous to the evaluations of the previous key figures.

Once all criteria are collected, the evaluations of all key figures can be performed. Table 5 shows the ranking points that could be also transformed into relative points (like in Table 3). The results can be used as a basis or component for a utility-cost matrix or AHP and are only shown here as an example. Above all, it is the task of

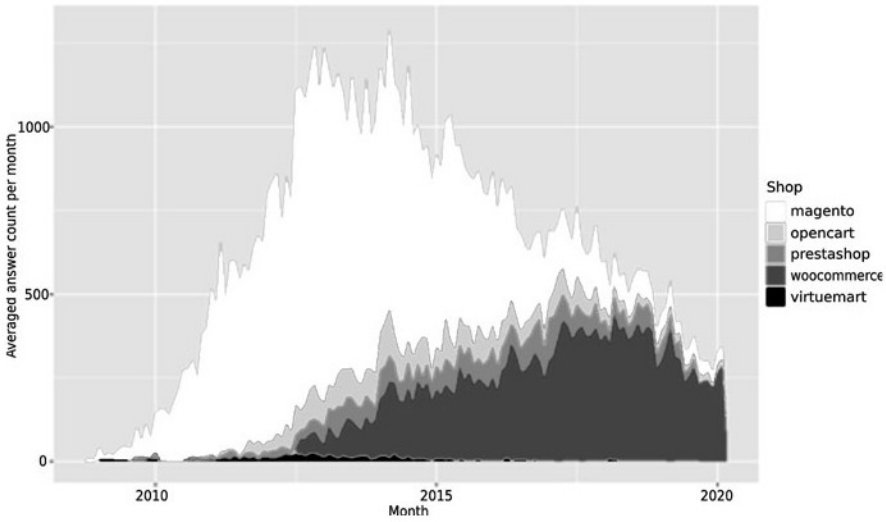

Fig. 4 Aggregated answer count per thread on Stackoverflow

Table 4 Knowledge base and community support indicators on Udemy and YouTube

Platform	Udemy topics	Udemy search results	Tutorial Videos on YouTube	Tutorial views on YouTube
Magento	48	81	562	3,204,093
WooCommerce	**78**	**725**	**570**	**11,929,255**
Prestashop	31	65	550	1,988,160
VirtueMart	0	4	73	52,184
OpenCart	19	54	479	1,979,464

Table 5 Assigned points to e-Commerce Platform

	Live Websites	Stackoverflow			Udemy		YouTube		Total
Magento	2	5	5	5	4	4	4	4	33
WooCommerce	**5**	**4**	**1**	**4**	**5**	**5**	**5**	**5**	**34**
Prestashop	3	2	4	3	3	3	3	3	24
VirtueMart	1	1	2	1	1	1	1	1	9
OpenCart	4	3	4	2	2	2	2	3	22

the decision-maker to weight the individual criteria, such as number of tutorials on Udemy, or number of threads on Stackoverflow.

This case study clearly shows how characteristics and information from different social media can be incorporated into decision support. When the data is carefully selected and aggregated, well-known decision algorithms like AHP can be easily applied without any adaptations to social big data.

4 Case 2: Tackling Veracity in Big Social Media Data

In the past, classic in-house databases and OLAP systems were the primary sources for decision support in companies. With the emergence of Web 2.0 and social media, a comprehensive range of data and information is available for decision support. This includes all resources on the web, from classic websites to discussion boards and social networks; a heterogeneous pool of information has emerged.

In the past, analysts often had to develop their own crawlers to extract information from websites. Even though this is a frequently used procedure, it is still fraught with a number of hurdles and shortcomings. Many websites are characterized by the fact that data is stored in a weakly structured manner. Additionally, websites are subject to frequent changes, e.g., in format and style sheet, which makes a permanent adaptation of the web crawler necessary. According to Glez-Peña et al. [34], the Web scraping process comprises several tasks: Site access through HTTP, HTML Parsing, and Output building.

Especially the parsing of HTML is a big challenge for automated crawlers, which is mainly due to the often low-structural strength of HTML. Although many libraries exist for parsing web pages such as Curl, Scrapy, BeautifulSoup, crawlers are considered weak software due to their vulnerability to changes in the web pages.

Because of the hurdles described above, classical data sources were preferred over the use of web crawlers. An alternative to extracting content from web pages are standardized interfaces. Recently, numerous Application Programming Interfaces (API) have become accessible, especially on the World Wide Web. These interfaces offer a structured and well-defined programming interface to access information on the corresponding platform [35]. Leading social media services like Twitter, Google, or Facebook operate well-defined interfaces in the form of an API. The advantage of APIs is the uniform interface and the strong structuring of the data. Disadvantages are the sometimes-limited access to data, quotas, and limits as well as the dependence to single providers. In general, data sources can be classified according to their structure (Fig. 5). APIs usually have the highest level of structuring, which makes data processing the easiest. Traditional HTML web pages are often unstructured and require a large amount of data preparation. In addition, other sources exist, e.g., databases, wikis, which differ in their structure, according to whether they are operated in-house and allow direct access or are public. This should be taken into account during planning and implementation.

4.1 Analyzing Data from a Social Network

For an educational institution, it is interesting to analyze professional qualifications that people in companies possess. The results help to understand, which types of skills are needed and whether general or very specific qualifications are important.

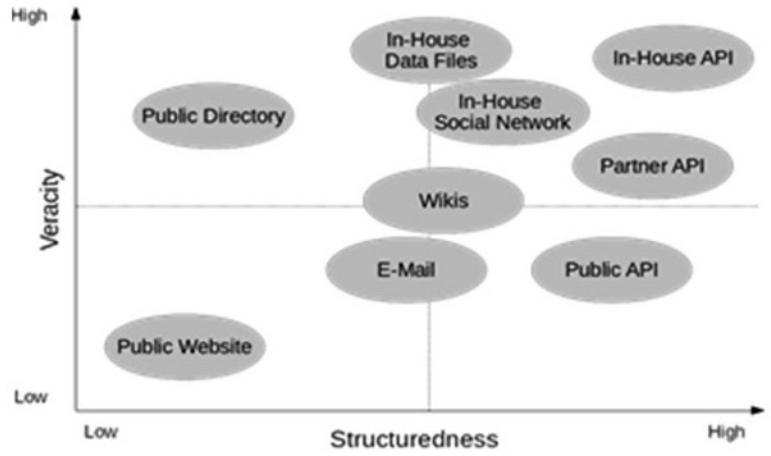

Fig. 5 Structuredness-veracity matrix of social media sources

This can influence decisions in study programs or courses in order to qualify students for the workplaces.

In this case study, we analyzed data from a career-oriented social network in order to identify and investigate distributions of professional qualifications. The aim was to find out whether—besides of the classical generalists and specialist—other types of qualifications can be found with unsupervised learning, and whether these can be found in specific types of companies [36]. Unsupervised learning is a type of machine learning that finds previously undetected patterns in data when class labels are not already known and training data from the past is not available. The data is thus grouped according to similarities and differences between the data.

Profile data was extracted in this case study from the career-oriented network. In a first step, we investigated the collected samples in order to assess whether the samples are representative. The selected social media platform returns only a limited amount of samples per data request, in this case, 200. By a systematic modification of the query, we could collect 7000 data samples. It remained unclear with which criteria the platform selected these samples. Maybe the data was chosen randomly, or the newest, oldest, or most active profiles were selected. Social networks do not provide transparency of how their filtering mechanism decides what data is selected from the social data stream. The selection algorithm can be furthermore biased by the so-called filter bubble, where the user is trapped inside the limited boundaries of his/her interests and cannot be exposed to any surprising, new, and desirable information [37]. The profile selection could therefore be also influenced by the available information of the person who extracted the data (IP-address, own profile . . .).

Based on the data analysis only, it could not be validated how the algorithm selected the data. The only way to validate is indirectly by analyzing the distribution of the qualification types and comparing it to the expected distribution.

4.2 Veracity Challenges

The second step was the classification of qualifications among the collected samples. It turned out, that they were written in different languages: German, English, Chinese, or several languages were used in the same profile. Classic methods of natural language processing (NLP) could be used only in a limited way because tokens (smaller parts of a text like words) especially in English and German can overlap and the learned model would get very inefficient. A learned model is the outcome of a machine-learning algorithm like rules, numbers, or data structures. Such a model was learned on data and can be later applied to new data.

In the investigated data, some qualification information contained writing errors, or skills were written in different ways (e.g., names of software with shortcut or long name or version number). As a solution to this problem, we manually created dictionaries with synonyms. In addition, we also manually created stop words. A stop word is a commonly used word (like “the”, “in”) that is removed before the application of an algorithm because it increases the amount of effort for the algorithm while providing only minimal benefit for the outcome. However, we were aware that parts of the information could be lost by removing these stop words later in the analysis.

Regarding the veracity, and especially the objectivity of the qualification information, we have to distinguish between the perspective of the creator and the reader of this information. In the investigated career-oriented network, we can assume sufficient objectivity of the information because via network connections (contacts) of the profile owner, a person is not anonymous anymore and would therefore not provide wrong information. However, the depth and quality of the given qualifications are subjective. If two people write “SQL” as a qualification, then the perspective, focus, and depth of this knowledge can differ between a database expert and someone who is just able to write simple SQL queries. The information reader can also read and understand such a qualification in a different way.

Another question was, whether we can investigate the development of qualifications over the time. This was impossible because the data samples contained only the current state of the profiles. Neither the timestamp of the last profile update nor the time when a specific information was added to the profile could be investigated. A longitudinal analysis would require a regular repetition of the data collection, which requires a bigger effort. Additionally, changes would only be visible at the time of the collection of new data, and some changes from the past would be still lost.

It was also difficult to analyze the company affiliation because profile owners used different ways of writing their company name in the free-text field. We could identify the following patterns:

- *Different granularity*: The company names are different depending on using the name of a holding (e.g., Siemens) or the name of the subsidiary (e.g., Siemens Healthineers).

- *Different timeline*: Company names were written down when the person was working there once (e.g., Siemens Medical Services) or the new name was used (e.g., Siemens Healthineers).
- *Different conventions*: The name was either written using the official registered name (e.g., Siemens Medical Solutions), or the officially used shortcut (e.g., SMS) or an industry-specific name (e.g., SMED).

We found similar problems in all the free-text data fields of a profile. This leads to a problem if the profile data has to be connected with other data sources (to achieve a bigger variety). Thus, we were not able to connect the carrier network data automatically with other data from company portals in order to see whether certain qualifications are especially needed in certain types of companies.

From the case study, we can conclude that social media data can be used as a basis for decision-making, but the veracity is a challenge. Every additional data source adds to a higher data variety and thus to more problems with veracity that can be only partly, and maybe only manually, fixed.

5 Conclusions from the Case Studies

The chapter approaches how data produced by users on social media can be the basis for decision-making. Social media with its user-generated content provides new possibilities for decision-making through a huge amount of different types of interesting data from different sources. However, some challenges have to be addressed.

The first case described decision-making based on a variety of social media data. This case showed how different data from social media and websites can provide novel, sometimes even contradictory information that were not available before and provide a good basis for decisions. Existing procedures for decision-making such as value analyses or AHP will not be changed by the inclusion of social media data, but additional criteria will become part of the analysis. Nevertheless, we should view the weighting of the individual criteria critically. Furthermore, it is difficult to find a holistic approach due to the sheer number of possible sources. Since social media provide user-generated content, special attention should be always paid to the veracity of the data.

In the second case study, our focus was therefore especially on the veracity. Based on our analysis, we propose the following guidelines in order to tackle the veracity of big social data:

1. Check if the results from the collected social media data (or of a randomly selected subset of this data) can be verified by the results from additional data sources or alternative analytics.
2. Check if it is necessary to create a platform-specific dictionary including stop words. Evaluate how the usage of such a dictionary will influence results.

3. During data understanding and data preparation, several platform-specific assumptions were made. Think about how these assumptions influence your results in terms of precision, recall, and specificity and how this can be measured.
4. Determine when the information was created. Does the platform allow the creator to manipulate timestamps or can invalid information be corrected or removed?
5. Verify if all information readers can understand the information equally. If not, assess the range of individual subjectivity.

With the help of these guidelines, it can be determined how much big social data has an influence on quality and rigor in the decision-making process. All cleansing approaches are associated with a loss of samples, which can easily melt the impressive data set to 10–50%. Overall, the tradeoff between accuracy of the result and sample size should always be estimated.

References

1. Kirchner, K., & Razmerita, L. (2019). Managing the digital knowledge work with the social media business value compass. In *Proceedings of the 52nd Hawaii International Conference on System Sciences* (pp. 6438–6447). New York: APA.
2. Power, D. J., & Phillips-Wren, G. (2011). Impact of social media and Web 2.0 on decision-making. *Journal of Decision Systems, 20*, 249–261. https://doi.org/10.3166/jds.20.249-261.
3. Freire, M., Antunes, F., & Paulo Costa, J. (2015). Exploring social network analysis techniques on decision support. In *Proceedings of the 2nd European Conference on Social Media ECSM* (Vol. 2015, pp. 165–173).
4. Keegan, B. J., & Rowley, J. (2017). Evaluation and decision making in social media marketing. *Management Decision, 55*, 15–31.
5. Wang, J. C., & Chang, C. H. (2013). How online social ties and product-related risks influence purchase intentions: A Facebook experiment. *Electronic Commerce Research and Applications, 12*, 337–346. https://doi.org/10.1016/j.elerap.2013.03.003.
6. DiStaso, M., & McCorkindale, T. (2017). *The science of influence: How social media affects decision-making in the healthcare, travel, retail, and financial industries*. Report retrieved from https://instituteforpr.org/science-influence-social-media-affects-decision-making-healthcare-travel-retail-financial-industries/
7. Conrado, S. P., Neville, K., Woodworth, S., & O'Riordan, S. (2016). Managing social media uncertainty to support the decision making process during emergencies. *Journal of Decision Systems, 25*, 171–181. https://doi.org/10.1080/12460125.2016.1187396.
8. Martin, N. (2019). How much data is collected every minute of the day. In: *Forbes*. Retrieved April 30, 2020, from https://www.forbes.com/sites/nicolemartin1/2019/08/07/how-much-data-is-collected-every-minute-of-the-day/#72daa1c33d66.
9. Bello-Orgaz, G., Jung, J. J., & Camacho, D. (2016). Social big data: Recent achievements and new challenges. *Information Fusion, 28*, 45–59. https://doi.org/10.1016/j.inffus.2015.08.005.
10. Labrinidis, A., & Jagadish, H. V. (2012). Challenges and opportunities with big data. *Proceedings of VLDB Endow, 5*, 2032–2033. https://doi.org/10.14778/2367502.2367572.
11. Orman, L. V. (2015). Information paradox: Drowning in information, starving for knowledge. *IEEE Technology Society, 2015*, 63–73.
12. McAfee, A., & Brynjolfsson, E. (2012). Big data: The management revolution. *Harvard Business Review, 90*, 60–68.
13. Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*, 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007.

14. White, T. (2015). *Hadoop: The definitive guide* (4th ed.). Newton: O'Reilly Media.
15. Zaharia, M., Chowdhury, M., Franklin, M. J., & Shenker, S. (2010). Spark: Cluster computing with working sets. In: *Proceedings of the 2nd USENIX conference on Hot topics in cloud computing*. pp. 1–7.
16. Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. *Communications of the ACM, 51*, 107–113. https://doi.org/10.1145/1327452.1327492.
17. Lukoianova, T., & Rubin, V. (2013). Veracity roadmap: Is big data objective, truthful and credible? *Advanced Classification Reseasrch Online, 24*, 4–15. https://doi.org/10.7152/acro.v24i1.14671.
18. Umble, E. J., Haft, R. R., & Umble, M. M. (2003). Enterprise resource planning: Implementation procedures and critical success factors. *European Journal of Operational Research, 146*, 241–257. https://doi.org/10.1016/S0377-2217(02)00547-7.
19. Beardwood, J. P., & Millar, P. (2019). Failed ERP implementation case study of MillerCoors v HCL. *Computer Law Review International, 20*, 136–142.
20. Larsen, M. H., Holck, J., & Pedersen, M. K. (2004). *The challenges of open source software in IT adoption: Enterprise architecture versus total cost of ownership*. IRIS'27, 2–20.
21. Jansen, A., Müller, C., Prümper, J., & Stein, B. (2005). Software-Einführung in KMU-(kein) Platz für Benutzerbeteiligung?-Eine qualitative Bestandaufnahme. In: *Berichtband des dritten Workshops des German Chapters der Usability Professionals Association e.V*. pp. 108–114.
22. Gallaugher, J. M., & Wang, Y. M. (2002). Understanding network effects in software markets: Evidence from Web server pricing. *Management Information Systems Quarterly, 26*, 303–327.
23. Katz, M. L., & Shapiro, C. (1994). Systems competition and network effects. *The Journal of Economic Perspectives, 8*, 93–115. https://doi.org/10.1257/jep.8.2.93.
24. Liebowitz, S., & Margolis, S. (2001). Network effects and the Microsoft case. In *Dynamic competition and public policy: Technology, innovation, and antitrust issues* (pp. 160–192). Cambridge: Cambridge University Press.
25. Benlian, A., & Hess, T. (2011). Comparing the relative importance of evaluation criteria in proprietary and open-source enterprise application software selection—a conjoint study of ERP and Office systems. *Information Systems Journal, 21*, 503–525. https://doi.org/10.1111/j.1365-2575.2010.00357.x.
26. Huang, Z., & Benyoucef, M. (2013). From e-commerce to social commerce: A close look at design features. *Electronic Commerce Research and Applications, 12*, 246–259. https://doi.org/10.1016/j.elerap.2012.12.003.
27. Fitzgerald, B. (2006). The transformation of open source software. *Management Information Systems Quarterly, 30*, 587–598. https://doi.org/10.2307/25148740.
28. Jadhav, A. S., & Sonar, R. M. (2009). Evaluating and selecting software packages: A review. *Information and Software Technology, 51*, 555–563.
29. Krivoruchko, J. (2007). The use of open source software in enterprise distributed computing environments: A decision-making framework for OSS selection and planning. In *IFIP International Conference on Open Source Systems* (pp. 277–282). Boston, MA: Springer.
30. Saaty, T. L. (1990). How to make a decision: The analytic hierarchy process. *European Journal of Operational Research, 48*, 9–26.
31. Vaidya, O. S., & Kumar, S. (2006). Analytic hierarchy process: An overview of applications. *European Journal of Operational Research, 169*, 1–29. https://doi.org/10.1016/j.ejor.2004.04.028.
32. Tung, Y. A. (1998). Time complexity and consistency issues in using the AHP for making group decisions. *Journal of Multi-Criteria Decision Analysis, 7*, 144–154. https://doi.org/10.1002/(SICI)1099-1360(199805)7:3<144::AID-MCDA180>3.0.CO;2-4.
33. Ulbricht, S., Opuszko, M., Ruhland, J., & Thrum, M. (2017). Towards an analysis and evaluation framework for in-memory-based use cases. In: *The twelfth international multi-conference on computing in the global information technology*. pp. 22–27.
34. Glez-Peña, D., Lourenço, A., López-Fernández, H., et al. (2014). Web scraping technologies in an API world. *Briefings in Bioinformatics, 15*, 788–797. https://doi.org/10.1093/bib/bbt026.

35. Dig, D., & Johnson, R. (2006). How do APIs evolve? A story of refactoring. *Journal of Software Maintenance and Evolution: Research and Practice, 18*, 83–107. https://doi.org/10.1002/smr.328.
36. Gehrke, S., Wenige, L., & Ruhland, J. (2019). Feasibility study of analysis of senior IT Management Skills/Qualifications in Social Networks. In: *ECSM 2019 6th European Conference on Social Media*. pp. 324–333.
37. Nagulendra, S., & Vassileva, J. (2013). Providing awareness, understanding and control of personalized stream filtering in a P2P social network. In *Collaboration and Technology. CRIWG 2013. Lecture Notes in Computer Science* (pp. 61–76). Berlin, Heidelberg: Springer.

4 Schlussbetrachtung

In klassischen Befragungen hat man umfangreiche Möglichkeiten die Untersuchungen zu steuern, d.h. durch gezielte Formulierungen des Fragenkatalogs die Wissensgenerierung zu beeinflussen. Im Gegensatz dazu, ist die Analyse von Social Media Daten ausschließlich beobachtend. Jedoch zeigte sich, dass die aufzuwendenden Ressourcen im Vergleich zur klassischen Befragung deutlich geringer sind und weitaus größere Datenmengen untersucht werden können. Der große Umfang der so gewonnenen Daten besitzt eine reduzierten Bias und Varianz gegenüber manuell gewonnenen Daten. Somit ließen sich die dort bereits gewonnenen Erkenntnisse validieren und auf eine breitere Basis stellen. Dies ist jedoch nicht gleichbedeutend mit der Aussage, dass die Analyse von Social Media Daten als alternative Datenquelle zu sehen ist. Vielmehr lässt sich ihre Funktion als konfirmierend und komplementierend beschreiben.

Auf der einen Seite ermöglicht die Beobachtung und Sammlung der Social Media Daten die Anwendung von Data Mining Techniken, um bisher unbekannte Muster, Gruppen und Verhalten zu erkennen. Durch den, im Vergleich zu klassischen Befragungen, zeitlich und personell deutlich geringerem Aufwand kann die Social Media Analyse gut zur initialen Exploration und nachfolgend als Basis des Designs klassischer Befragungen eingesetzt werden. Dies könnte positiven Einfluss auf die Entwicklung von Befragungskatalogen haben.

Auf der anderen Seite können die Ergebnisse klassischer Befragungen leichter auf ihren Umfang oder Auswirkung in der Gesamtpopulation abgeschätzt werden. Während kleinzahlige Befragung immer den Makel der Nichtrepräsentativität tragen, lassen sich die gewonnenen Erkenntnisse damit auf eine breitere Basis stellen. Dieses Vorgehen folgt der Idee der Plausibilisierung im supervised learning.

Weiterhin besteht die Möglichkeit, durch kleinzahlige, gezielte Befragungen klassifizierte Datensätze zu erhalten. Im nächsten Schritt können die Datenbasis durch Social Media Daten verbreitert werden und mit Hilfe von Clusterverfahren eingeteilt werden. Dementsprechend stünde für Klassifizierungsverfahren eine größere Datenbasis zu Verfügung, welche entweder einen positiven Einfluss auf die Qualität haben oder bestimmte Verfahren, wie zum Beispiel neuronale Netze, erst ermöglichen. Dies entspricht der Idee des semi-supervised learning.

Trotz der positiv einzuschätzenden Ergebnisse, ergeben sich daraus Anforderungen die besondere Aufmerksamkeit bedürfen. Neben dem Ermitteln geeigneter Sozial Media Quellen, stellt die Qualität der dort gefundenen Daten eine zentrale Herausforderung dar. Somit ist den

Qualitätsanforderungen (Tabelle 5) besondere Beachtung zu schenken. Dies gilt nicht nur hinsichtlich der Beantwortung der Frage, welche Datenqualität den Anforderungen der Forschungsfragen genügt, sondern auch ob die Wahl der Social Media Datenquelle damit einhergeht. Dabei ist zu beachten, dass die Qualität der Social Media Quelle zeitlich und inhaltlich nicht konsistent sein muss. In wieweit die Datenquelle den Anforderungen genügt, sollte vor Verwendung entsprechend analysiert und transparent dargestellt werden. Weiterhin zeigt der Überblick über bestehende Data Mining Techniken mit den damit verbundenen Fragestellungen wie den zu wählenden Distanz oder Ähnlichkeitsmaßen, dass es keinen allgemeingültigen Modellierungsansatz gibt und jeweils die passende Auswahl getroffen werden muss.

Wie in der dritten Publikation gezeigt werden konnte, hängt das Vertrauen in jegliche Datenanalysen stark von Transparenz und Möglichkeiten der involvierten Stakeholder ab, die einzelnen Schritte verständlich nachvollziehen zu können. Für das Vertrauen sind weniger spezifische Qualitätskriterien relevant, da diese im Zweifelsfall sehr fachspezifisch sind und nicht gleichmäßig von allen Beteiligten verstanden werden. Sicherlich ist dies für allen Arten von Analysen richtig, betrachtetet man aber die Heterogenität der Stakeholder im Social Media Umfeld ist die Allgemeinverständlichkeit besonders relevant. Die interessierten Parteien, ob akademisch oder wirtschaftlich, teilen sich bereits in Fach- und Managementebenen auf, mit jeweils unterschiedlichem Verständnis der Materie. Die Datenanalysten bedürfen neben Modell- und Analysekenntnissen, ein Verständnis der zugrundeliegenden Daten und der Qualität aus Sicht der Fachabteilungen. Der Nutzer oder die betroffenen Konsumenten der Analyseergebnisse bedürfen, da es sich in vielen Fällen um nicht erfahrbare Vertrauensgüter handelt, besonderes Vertrauen in das Vorgehen. Um den Transfer von bekannten vertrauten Situation (engl. „Institution Based Trust“) oder vertrauten Personen (engl. „Trusting Beliefs“) zu ermöglichen oder zu erleichtern, sollte ein allgemein verständliche Sprache verwendet werden, beginnend bei der Beschreibung der verwendeten Social Media Datenquelle, deren Qualität oder Eignung, über die Modellierung bis zur Evaluierung der Ergebnisse und den Auswirkungen.

Gelingt dies, so zeigt die vierte Publikation, wird Social Media Datenanalyse einen Beitrag zur Entscheidungsfindung leisten können.

5 Ausblick

Im Rahmen der vorliegende Arbeit konnten Chancen und Herausforderungen von Social Media Datenanalysen aufgezeigt werden. Um das Potenzial dieser nutzen zu können, ist eine interdisziplinäre Herangehensweise erforderlich. Während die vorliegenden Publikationen noch durch eine Einzelperson realisiert werden konnten, wird klar das die Einbeziehung der unterschiedlichen Disziplinen eine verstärkte Zusammenarbeit erfordert. Die Erweiterung der Agenten basierten Simulation um die „Mean Field Game Theory“ erfordert z.B. fortgeschrittene Kenntnisse der Physik, Sentiment Analysen erfordert die Zusammenarbeit von Linguisten und Informatikern, Clusteranalysen bedürfen der Zusammenarbeit von Datenanalytikern und Soziologen. Um das Potential der Analyseergebnisse zu heben sollten Wirtschaftswissenschaftler einbezogen bzw. sind diese Treiber und Wertschöpfer. Somit ist zukünftig eine verstärkte Zusammenarbeit zu erwarten, welches zu komplexen Formen der Zusammenarbeit führen wird. Dies wiederum bedingt Konzepte und Frameworks, um die Zusammenarbeit transparent und verständlich gestalten zu können.

6 Literaturverzeichnis

[1] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus, “Knowledge Discovery in Databases: An Overview,” *AI Mag.*, vol. 13, no. 3, p. 57, Sep. 1992, doi: 10.1609/aimag.v13i3.1011.

[2] “Obama Administration Unveils $200M Big Data R&D Initiative » CCC Blog,” Mar. 29, 2012. https://cccblog.org/2012/03/29/obama-administration-unveils-200m-big-data-rd-initiative/ (accessed Jan. 22, 2022).

[3] “social media data | Nature Search Results.” https://www.nature.com/search?q=social%20media%20data&article_type=research&date_range=last_year&order=relevance (accessed Jan. 22, 2022).

[4] “Science ‘social media data,’” *Science.org*. https://www.science.org/action/doSearch?AllField=social+Media+data&Earliest=%5B20210122+TO+20220122%5D&startPage=&ConceptID=505172 (accessed Jan. 22, 2022).

[5] A. M. Kaplan and M. Haenlein, “Users of the world, unite! The challenges and opportunities of Social Media,” *Bus. Horiz.*, vol. 53, no. 1, pp. 59–68, Jan. 2010, doi: 10.1016/j.bushor.2009.09.003.

[6] P. Gundecha and H. Liu, “Mining Social Media: A Brief Introduction,” in *2012 TutORials in Operations Research*, INFORMS, 2012, pp. 1–17. doi: 10.1287/educ.1120.0105.

[7] www.statista.com, “Statistiken,” *Statista*. https://www.statista.com (accessed Nov. 27, 2021).

[8] “Amazon - Umsatz weltweit nach Quartalen 2021,” *Statista*. https://de.statista.com/statistik/daten/studie/197099/umfrage/nettoumsatz-von-amazoncom-quartalszahlen/ (accessed Feb. 25, 2022).

[9] “Twitter by the Numbers (2021): Stats, Demographics & Fun Facts.” https://www.omnicoreagency.com/twitter-statistics/ (accessed Nov. 27, 2021).

[10] H. Allcott and M. Gentzkow, “Social Media and Fake News in the 2016 Election,” *J. Econ. Perspect.*, vol. 31, no. 2, pp. 211–236, May 2017, doi: 10.1257/jep.31.2.211.

[11] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake News Detection on Social Media: A Data Mining Perspective,” vol. 19, no. 1, p. 15.

[12] A. Loktionov, “Ramesses II, victor of Kadesh: a kindred spirit of Trump?,” *The Guardian*, Dec. 05, 2016. Accessed: Nov. 27, 2021. [Online]. Available: https://www.theguardian.com/science/blog/2016/dec/05/ramesses-ii-victor-of-kadesh-a-kindred-spirit-of-trump

[13] D. C. Polage, “Making up History: False Memories of Fake News Stories,” *Eur. J. Psychol.*, vol. 8, no. 2, pp. 245–250, May 2012, doi: 10.5964/ejop.v8i2.456.

[14] J. I. Wong, “Almost all the traffic to fake news sites is from Facebook, new data show,” *Quartz*. https://qz.com/848917/facebook-fb-fake-news-data-from-jumpshot-its-the-biggest-traffic-referrer-to-fake-and-hyperpartisan-news-sites/ (accessed Nov. 27, 2021).

[15] O. Tene, J. Polonetsky, and A.-R. Sadeghi, “Five Freedoms for the Homo Deus,” *IEEE Secur. Priv.*, vol. 16, no. 3, pp. 15–17, May 2018, doi: 10.1109/MSP.2018.2701156.

[16] “Netzwerkdurchsetzungsgesetz,” *Bundesministerium der Justiz und für Verbraucherschutz*. https://www.BMJV.de/DE/Themen/FokusThemen/NetzDG/NetzDG_node.html (accessed Nov. 27, 2021).

[17] T. Khan, A. Michalas, and A. Akhunzada, “Fake news outbreak 2021: Can we stop the viral spread?,” *J. Netw. Comput. Appl.*, vol. 190, p. 103112, Sep. 2021, doi: 10.1016/j.jnca.2021.103112.

[18] S. Ahmed, K. Hinkelmann, and F. Corradini, “Combining Machine Learning with Knowledge Engineering to detect Fake News in Social Networks-a survey,” 2022, doi: 10.48550/ARXIV.2201.08032.
[19] T. Jiang, J. P. Li, A. U. Haq, A. Saboor, and A. Ali, “A Novel Stacking Approach for Accurate Detection of Fake News,” *IEEE Access*, vol. 9, pp. 22626–22639, 2021, doi: 10.1109/ACCESS.2021.3056079.
[20] M. D. Ibrishimova and K. F. Li, “A machine learning approach to fake news detection using knowledge verification and natural language processing,” in *International Conference on Intelligent Networking and Collaborative Systems*, 2019, pp. 223–234.
[21] n-tv NACHRICHTEN, “Hashtags deaktiviert: China unterbindet Shitstorm gegen Eiskunstläuferin,” *n-tv.de*. https://www.n-tv.de/sport/der_sport_tag/Hashtags-deaktiviert-China-unterbindet-Shitstorm-gegen-Eiskunstlaeuferin-article23108788.html (accessed Feb. 25, 2022).
[22] A. Pinto, H. Gonçalo Oliveira, and A. Oliveira Alves, “Comparing the Performance of Different NLP Toolkits in Formal and Social Media Text,” p. 16 pages, 2016, doi: 10.4230/OASICS.SLATE.2016.3.
[23] F. Benenson, *How to speak emoji*. London: Ebury Press, 2015.
[24] “Privacy at XING.” https://privacy.xing.com/en/privacy-policy/general-information (accessed Nov. 28, 2021).
[25] M. Eppler, “A Framework for Information Quality Management,” in *Managing Information Quality*, Springer Berlin Heidelberg, 2006, pp. 57–210. doi: 10.1007/3-540-32225-6_3.
[26] Y. W. Lee, D. M. Strong, B. K. Kahn, and R. Y. Wang, “AIMQ: a methodology for information quality assessment,” *Inf. Manage.*, vol. 40, no. 2, pp. 133–146, 2002.
[27] T. C. Redman, *Data quality for the information age*. Artech House, Inc., 1997.
[28] R. Y. Wang, V. C. Storey, and C. P. Firth, “A framework for analysis of data quality research,” *IEEE Trans. Knowl. Data Eng.*, vol. 7, no. 4, pp. 623–640, Aug. 1995, doi: 10.1109/69.404034.
[29] R. Farnworth, “The Six Dimensions of Data Quality — and how to deal with them,” *Medium*, Jun. 30, 2020. https://towardsdatascience.com/the-six-dimensions-of-data-quality-and-how-to-deal-with-them-bdcf9a3dba71 (accessed Nov. 14, 2021).
[30] D. Loshin, *Master data management*. Amsterdam ; Boston: Elsevier/Morgan Kaufmann, 2009.
[31] Smartbridge, “Data Done Right: 6 Dimensions of Data Quality,” *Smartbridge*, Oct. 14, 2020. https://smartbridge.com/data-done-right-6-dimensions-of-data-quality/ (accessed Nov. 14, 2021).
[32] 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.
[33] Y. Wand and R. Y. Wang, “Anchoring data quality dimensions in ontological foundations,” *Commun. ACM*, vol. 39, no. 11, pp. 86–95, Nov. 1996, doi: 10.1145/240455.240479.
[34] D. H. McKnight and N. L. Chervany, “What is Trust? A Conceptual Analysis and an Interdisciplinary Model,” *AMCIS 2000 Proc. P382*, p. 8, 2000.
[35] “ISO 8000-8:2015(en), Data quality — Part 8: Information and data quality: Concepts and measuring.” https://www.iso.org/obp/ui/#iso:std:60805:en (accessed Dec. 08, 2021).
[36] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From Data Mining to Knowledge Discovery in Databases,” *AI Mag.*, vol. 17, no. 3, p. 37, Mar. 1996, doi: 10.1609/aimag.v17i3.1230.
[37] M. J. Eppler and D. Wittig, “Conceptualizing Information Quality: A Review of Information Quality Frameworks from the Last Ten Years.,” *IQ*, vol. 20, no. 0, p. 0, 2000.
[38] E. Rahm and H. H. Do, “Data cleaning: Problems and current approaches,” *IEEE Data Eng Bull*, vol. 23, no. 4, pp. 3–13, 2000.
[39] N. Laranjeiro, S. N. Soydemir, and J. Bernardino, “A Survey on Data Quality: Classifying Poor Data,” in *2015 IEEE 21st Pacific Rim International Symposium on Dependable*

Computing (PRDC), Zhangjiajie, China, Nov. 2015, pp. 179–188. doi: 10.1109/PRDC.2015.41.
[40] “Review: Databases built for social networks | IT World Canada News,” Feb. 18, 2010. https://www.itworldcanada.com/article/review-databases-built-for-social-networks/40980 (accessed Dec. 07, 2021).
[41] Richthammer, Christian, Netter, Michael, Riesner, Moritz, Sänger, Johannes, and Pernul, Günther, “Taxonomy of Social Network Data Types,” 2014, doi: 10.5283/EPUB.30742.
[42] S. Wilson, “Big data held to privacy laws, too,” *Nature*, vol. 519, no. 7544, pp. 414–414, Mar. 2015, doi: 10.1038/519414a.
[43] Z. Ghahramani, “Unsupervised Learning,” in *Advanced Lectures on Machine Learning*, vol. 3176, O. Bousquet, U. von Luxburg, and G. Rätsch, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 72–112. doi: 10.1007/978-3-540-28650-9_5.
[44] J. Rocca, “Ensemble methods: bagging, boosting and stacking,” *Medium*, Mar. 21, 2021. https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205 (accessed Dec. 02, 2021).
[45] J. Kuha, “AIC and BIC: Comparisons of Assumptions and Performance,” *Sociol. Methods Res.*, vol. 33, no. 2, pp. 188–229, Nov. 2004, doi: 10.1177/0049124103262065.
[46] L. Torgo and R. Ribeiro, “Precision and recall for regression,” in *International Conference on Discovery Science*, 2009, pp. 332–346.
[47] N. R. Draper and H. Smith, *Applied regression analysis*, 3rd ed. New York: Wiley, 1998.
[48] J. Han and M. Kamber, *Data mining: concepts and techniques*, 3rd ed. Burlington, MA: Elsevier, 2012.
[49] D. Arthur and S. Vassilvitskii, “k-means++: The advantages of careful seeding,” Stanford, 2006.
[50] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *J. R. Stat. Soc. Ser. B Methodol.*, vol. 39, no. 1, pp. 1–22, 1977.
[51] D. Dueck, *Affinity propagation: clustering data by passing messages*. Citeseer, 2009.
[52] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise.,” in *kdd*, 1996, vol. 96, no. 34, pp. 226–231.
[53] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, “OPTICS: ordering points to identify the clustering structure,” *ACM SIGMOD Rec.*, vol. 28, no. 2, pp. 49–60, Jun. 1999, doi: 10.1145/304181.304187.
[54] E. Hartuv and R. Shamir, “A clustering algorithm based on graph connectivity,” *Inf. Process. Lett.*, vol. 76, no. 4–6, pp. 175–181, Dec. 2000, doi: 10.1016/S0020-0190(00)00142-3.
[55] M. S. Handcock, A. E. Raftery, and J. M. Tantrum, “Model-based clustering for social networks,” *J. R. Stat. Soc. Ser. A Stat. Soc.*, vol. 170, no. 2, pp. 301–354, Mar. 2007, doi: 10.1111/j.1467-985X.2007.00471.x.
[56] H. Steinicke and Deutsche Akademie der Naturforscher Leopoldina, Eds., *Herausforderungen und Chancen der integrativen Taxonomie für Forschung und Gesellschaft: taxonomische Forschung im Zeitalter der OMICS-Technologien: Juni 2014 ; Stellungnahme]*, 1. Aufl. Berlin: Mediabogen, 2014.
[57] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987, doi: 10.1016/0377-0427(87)90125-7.
[58] T. Caliński and J. Harabasz, “A dendrite method for cluster analysis,” *Commun. Stat.-Theory Methods*, vol. 3, no. 1, pp. 1–27, 1974.
[59] W. M. Rand, “Objective criteria for the evaluation of clustering methods,” *J. Am. Stat. Assoc.*, vol. 66, no. 336, pp. 846–850, 1971.
[60] M. Belkin, P. Niyogi, and V. Sindhwani, “Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples,” p. 36.

[61] O. Chapelle, J. Weston, and B. Schölkopf, “Cluster kernels for semi-supervised learning,” 2002.
[62] O. Chapelle, B. Schölkopf, and A. Zien, Eds., *Semi-supervised learning*. Cambridge, Mass: MIT Press, 2006.
[63] M. F. A. Hady and F. Schwenker, “Semi-supervised Learning,” in *Handbook on Neural Information Processing*, M. Bianchini, M. Maggini, and L. C. Jain, Eds. Berlin, Heidelberg: Springer, 2013, pp. 215–239. doi: 10.1007/978-3-642-36657-4_7.
[64] “An overview of proxy-label approaches for semi-supervised learning,” *Sebastian Ruder*, Apr. 26, 2018. https://ruder.io/semi-supervised/ (accessed Feb. 07, 2022).
[65] T. Joachims, *Transductive Support Vector Machines*. Cambridge, Mass: MIT Press, 2006.
[66] Zhi-Hua Zhou and Ming Li, “Tri-training: exploiting unlabeled data using three classifiers,” *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 11, pp. 1529–1541, Nov. 2005, doi: 10.1109/TKDE.2005.186.
[67] A. Sogaard, “Simple Semi-Supervised Training of Part-Of-Speech Taggers,” p. 4.
[68] J. Kügelgen, A. Mey, M. Loog, and B. Schölkopf, “Semi-supervised learning, causality, and the conditional cluster assumption,” in *Conference on Uncertainty in Artificial Intelligence*, 2020, pp. 1–10.
[69] M. Deza and E. Deza, *Encyclopedia of distances*, Second edition. Heidelberg ; New York: Springer, 2013.
[70] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, “On the surprising behavior of distance metrics in high dimensional space,” in *International conference on database theory*, 2001, pp. 420–434.
[71] P. Domingos, “A few useful things to know about machine learning,” *Commun. ACM*, vol. 55, no. 10, pp. 78–87, 2012.
[72] T. González-Arteaga, J. C. R. Alcantud, and R. de Andrés Calle, “A cardinal dissensus measure based on the Mahalanobis distance,” *Eur. J. Oper. Res.*, vol. 251, no. 2, pp. 575–585, 2016.
[73] K.-H. Yuan and P. M. Bentler, “Effect of outliers on estimators and tests in covariance structure analysis,” *Br. J. Math. Stat. Psychol.*, vol. 54, no. 1, pp. 161–175, 2001, doi: 10.1348/000711001159366.
[74] S.-S. Choi, S.-H. Cha, and C. C. Tappert, “A Survey of Binary Similarity and Distance Measures,” vol. 8, no. 1, p. 6, 2010.
[75] C.-M. Hwang, M.-S. Yang, and W.-L. Hung, “New similarity measures of intuitionistic fuzzy sets based on the Jaccard index with its application to clustering,” *Int. J. Intell. Syst.*, vol. 33, no. 8, pp. 1672–1688, 2018, doi: 10.1002/int.21990.
[76] L. Leydesdorff, “On the normalization and visualization of author co-citation data: Salton’s Cosine versus the Jaccard index,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 59, no. 1, pp. 77–85, Jan. 2008, doi: 10.1002/asi.20732.
[77] A. K. Gupta and N. Sardana, “Significance of Clustering Coefficient over Jaccard Index,” in *2015 Eighth International Conference on Contemporary Computing (IC3)*, Aug. 2015, pp. 463–466. doi: 10.1109/IC3.2015.7346726.
[78] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, “Recommender systems survey,” *Knowl.-Based Syst.*, vol. 46, pp. 109–132, Jul. 2013, doi: 10.1016/j.knosys.2013.03.012.
[79] J. Bobadilla, F. Serradilla, and J. Bernal, “A new collaborative filtering metric that improves the behavior of recommender systems,” *Knowl.-Based Syst.*, vol. 23, no. 6, pp. 520–528, Aug. 2010, doi: 10.1016/j.knosys.2010.03.009.
[80] Baeldung, “Euclidean Distance vs Cosine Similarity | Baeldung on Computer Science,” Jun. 23, 2020. https://www.baeldung.com/cs/euclidean-distance-vs-cosine-similarity (accessed Mar. 06, 2022).

[81] N. Alnajran, K. Crockett, D. McLean, and A. Latham, “Cluster Analysis of Twitter Data: A Review of Algorithms:,” in *Proceedings of the 9th International Conference on Agents and Artificial Intelligence*, Porto, Portugal, 2017, pp. 239–249. doi: 10.5220/0006202802390249.
[82] Z. Atashgahi, J. Pieterse, S. Liu, D. C. Mocanu, R. Veldhuis, and M. Pechenizkiy, “A Brain-inspired Algorithm for Training Highly Sparse Neural Networks,” *ArXiv190307138 Cs*, Oct. 2021, Accessed: Jan. 10, 2022. [Online]. Available: http://arxiv.org/abs/1903.07138
[83] “COSINE DISTANCE, COSINE SIMILARITY, ANGULAR COSINE DISTANCE, ANGULAR COSINE SIMILARITY.” https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/cosdist.htm (accessed Jan. 10, 2022).
[84] “vector spaces - Cosine similarity vs angular distance,” *Mathematics Stack Exchange*. https://math.stackexchange.com/questions/2874940/cosine-similarity-vs-angular-distance (accessed Jan. 10, 2022).
[85] S. Kullback, *Information theory and statistics*, Reprint. Gloucester, Mass: Smith, 1978.
[86] I. H. Witten, E. Frank, and M. A. Hall, *Data mining: practical machine learning tools and techniques*, 3rd ed. Burlington, MA: Morgan Kaufmann, 2011.
[87] L. Van der Maaten and G. Hinton, “Visualizing data using t-SNE.,” *J. Mach. Learn. Res.*, vol. 9, no. 11, 2008.
[88] U. Brandes and T. Erlebach, *Network Analysis: Methodological Foundations*. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg, 2005.
[89] G. Nandi and A. Das, “A Survey on Using Data Mining Techniques for Online Social Network Analysis,” vol. 10, no. 6, p. 6, 2013.
[90] M. Adedoyin-Olowe, M. M. Gaber, and F. Stahl, “A Survey of Data Mining Techniques for Social Network Analysis,” p. 25.
[91] S. Gole and B. Tidke, “A survey of big data in social media using data mining techniques,” in *2015 International Conference on Advanced Computing and Communication Systems*, Jan. 2015, pp. 1–6. doi: 10.1109/ICACCS.2015.7324059.
[92] B. Batrinca and P. C. Treleaven, “Social media analytics: a survey of techniques, tools and platforms,” *AI Soc.*, vol. 30, no. 1, pp. 89–116, Feb. 2015, doi: 10.1007/s00146-014-0549-4.
[93] D. Elangovan, V. Subedha, R. Sathishkumar, and V. D. Ambeth kumar, “A Survey: Data Mining Techniques for Social Media Analysis,” presented at the International Conference for Phoenixes on Emerging Current Trends in Engineering and Management (PECTEAM 2018), Chennai, India, 2018. doi: 10.2991/pecteam-18.2018.19.
[94] S. Das and A. Biswas, “Deployment of Information Diffusion for Community Detection in Online Social Networks: A Comprehensive Review,” *IEEE Trans. Comput. Soc. Syst.*, vol. 8, no. 5, pp. 1083–1107, Oct. 2021, doi: 10.1109/TCSS.2021.3076930.
[95] A. Lancichinetti and S. Fortunato, “Community detection algorithms: A comparative analysis,” *Phys. Rev. E*, vol. 80, no. 5, p. 056117, Nov. 2009, doi: 10.1103/PhysRevE.80.056117.
[96] J. Leskovec, K. J. Lang, and M. Mahoney, “Empirical comparison of algorithms for network community detection,” in *Proceedings of the 19th international conference on World wide web*, New York, NY, USA, Apr. 2010, pp. 631–640. doi: 10.1145/1772690.1772755.
[97] A. McCallum, A. Corrada-Emmanuel, and X. Wang, “Topic and role discovery in social networks,” 2005.
[98] Z. Zhao, S. Feng, Q. Wang, J. Z. Huang, G. J. Williams, and J. Fan, “Topic oriented community detection through social objects and link analysis in social networks,” *Knowl.-Based Syst.*, vol. 26, pp. 164–173, Feb. 2012, doi: 10.1016/j.knosys.2011.07.017.
[99] L. Parsons, E. Haque, and H. Liu, “Subspace clustering for high dimensional data: a review,” *ACM SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 90–105, Jun. 2004, doi: 10.1145/1007730.1007731.

[100] S. Ghasemi and A. Zarei, “Improving link prediction in social networks using local and global features: a clustering-based approach,” *Prog. Artif. Intell.*, Sep. 2021, doi: 10.1007/s13748-021-00261-3.
[101] O. Husain, N. Salim, R. A. Alias, S. Abdelsalam, and A. Hassan, “Expert Finding Systems: A Systematic Review,” *Appl. Sci.*, vol. 9, no. 20, p. 4250, Oct. 2019, doi: 10.3390/app9204250.
[102] E. Gaussier, C. Goutte, K. Popat, and F. Chen, “A hierarchical model for clustering and categorising documents,” in *European Conference on Information Retrieval*, 2002, pp. 229–247.
[103] A. Vinokourov and M. Girolami, “A Probabilistic Framework for the Hierarchic Organisation and Classification of Document Collections,” *J. Intell. Inf. Syst.*, vol. 18, no. 2/3, pp. 153–172, 2002, doi: 10.1023/A:1013677411002.
[104] A. Legout, G. Urvoy-Keller, and P. Michiardi, “Rarest first and choke algorithms are enough,” in *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*, New York, NY, USA, Oct. 2006, pp. 203–216. doi: 10.1145/1177080.1177106.
[105] P. Wang, B. Xu, Y. Wu, and X. Zhou, “Link prediction in social networks: the state-of-the-art,” *Sci. China Inf. Sci.*, vol. 58, no. 1, pp. 1–38, 2015.
[106] S. Aslan and B. Kaya, “Time-aware link prediction based on strengthened projection in bipartite networks,” *Inf. Sci.*, vol. 506, pp. 217–233, Jan. 2020, doi: 10.1016/j.ins.2019.08.025.
[107] S. Daminelli, J. M. Thomas, C. Durán, and C. V. Cannistraci, “Common neighbours and the local-community-paradigm for link prediction in bipartite networks,” *New J. Phys.*, vol. 17, no. 11, p. 113037, Nov. 2015, doi: 10.1088/1367-2630/17/11/113037.
[108] I. Warwas, M. Dzimińska, and A. Krzewinska, “The Frequency of Using Websites and Social Media by Various Age Groups to Form Opinions about Scientific Topics: Findings from the European Context,” presented at the Hawaii International Conference on System Sciences, 2021. doi: 10.24251/HICSS.2021.324.
[109] F. Kamoun, S. Gharbi, and A. A. Ghazeli, “Reconnecting with the past: a framework to better serve the information needs of older people on social networking sites,” *Work. Older People*, 2018.
[110] R. Van Meteren and M. Van Someren, “Using content-based filtering for recommendation,” in *Proceedings of the machine learning in the new information age: MLnet/ECML2000 workshop*, 2000, vol. 30, pp. 47–56.
[111] G. Antolić and L. Brkić, “Recommender system based on the analysis of publicly available data,” in *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, May 2017, pp. 1379–1384. doi: 10.23919/MIPRO.2017.7973637.
[112] D. H. McKnight and N. L. Chervany, “What Trust Means in E-Commerce Customer Relationships: An Interdisciplinary Conceptual Typology,” *Int. J. Electron. Commer.*, vol. 6, no. 2, Art. no. 2, Dec. 2001, doi: 10.1080/10864415.2001.11044235.
[113] M. Opuszko, T. Wöhner, R. Peters, and J. Ruhland, “Qualitätsmessung in der Wikipedia.,” in *MKWI*, 2010, pp. 705–716.
[114] H. Fang, X. Li, and J. Zhang, “Integrating social influence modeling and user modeling for trust prediction in signed networks,” *Artif. Intell.*, vol. 302, p. 103628, Jan. 2022, doi: 10.1016/j.artint.2021.103628.
[115] S. Kumar, F. Spezzano, V. S. Subrahmanian, and C. Faloutsos, “Edge Weight Prediction in Weighted Signed Networks,” in *2016 IEEE 16th International Conference on Data Mining (ICDM)*, Dec. 2016, pp. 221–230. doi: 10.1109/ICDM.2016.0033.
[116] N. Girdhar, S. Minz, and K. K. Bharadwaj, “Link prediction in signed social networks based on fuzzy computational model of trust and distrust,” *Soft Comput.*, vol. 23, no. 22, pp. 12123–12138, Nov. 2019, doi: 10.1007/s00500-019-03768-z.

[117] T. DuBois, J. Golbeck, and A. Srinivasan, “Predicting Trust and Distrust in Social Networks,” in *2011 IEEE Third Int’l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int’l Conference on Social Computing*, Boston, MA, USA, Oct. 2011, pp. 418–424. doi: 10.1109/PASSAT/SocialCom.2011.56.

[118] J. Leskovec, D. Huttenlocher, and J. Kleinberg, “Predicting positive and negative links in online social networks,” in *Proceedings of the 19th international conference on World wide web - WWW ’10*, Raleigh, North Carolina, USA, 2010, p. 641. doi: 10.1145/1772690.1772756.

[119] C.-J. Hsieh, K.-Y. Chiang, and I. S. Dhillon, “Low rank modeling of signed networks,” in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’12*, Beijing, China, 2012, p. 507. doi: 10.1145/2339530.2339612.

[120] K.-Y. Chiang, N. Natarajan, A. Tewari, and I. S. Dhillon, “Exploiting longer cycles for link prediction in signed networks,” in *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2011, pp. 1157–1162.

[121] Y. Jing, H. Wang, K. Shao, and X. Huo, “Relation Representation Learning via Signed Graph Mutual Information Maximization for Trust Prediction,” *Symmetry*, vol. 13, no. 1, p. 115, Jan. 2021, doi: 10.3390/sym13010115.

[122] Z. Wu, C. C. Aggarwal, and J. Sun, “The Troll-Trust Model for Ranking in Signed Networks,” in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, New York, NY, USA, Feb. 2016, pp. 447–456. doi: 10.1145/2835776.2835816.

[123] “Propagation of trust and distrust for the detection of trolls in a social network - ScienceDirect.” https://www.sciencedirect.com/science/article/pii/S138912861200179X (accessed Dec. 20, 2021).

[124] “Epinions,” *Wikipedia*. Nov. 07, 2021. Accessed: Dec. 20, 2021. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Epinions&oldid=1054055544

[125] S. M. Ghafari *et al.*, “A Survey on Trust Prediction in Online Social Networks,” *IEEE Access*, vol. 8, pp. 144292–144309, 2020, doi: 10.1109/ACCESS.2020.3009445.

[126] B. Mathew, R. Dutt, P. Goyal, and A. Mukherjee, “Spread of Hate Speech in Online Social Media,” in *Proceedings of the 10th ACM Conference on Web Science*, Boston Massachusetts USA, Jun. 2019, pp. 173–182. doi: 10.1145/3292522.3326034.

[127] C. Chamley, A. Scaglione, and L. Li, “Models for the Diffusion of Beliefs in Social Networks: An Overview,” *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 16–29, May 2013, doi: 10.1109/MSP.2012.2234508.

[128] “Scientific realism,” *Wikipedia*. Aug. 09, 2021. Accessed: Dec. 20, 2021. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Scientific_realism&oldid=1037989005

[129] D. Bauso, H. Tembine, and T. Başar, “Opinion Dynamics in Social Networks through Mean-Field Games,” *SIAM J. Control Optim.*, vol. 54, no. 6, pp. 3225–3257, Jan. 2016, doi: 10.1137/140985676.

[130] R. A. Banez, H. Gao, L. Li, C. Yang, Z. Han, and H. V. Poor, “Belief and Opinion Evolution in Social Networks Based on a Multi-Population Mean Field Game Approach,” in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, Jun. 2020, pp. 1–6. doi: 10.1109/ICC40277.2020.9148985.

[131] A. Nordio, A. Tarable, C.-F. Chiasserini, and E. Leonardi, “Belief dynamics in social networks: A fluid-based analysis,” *IEEE Trans. Netw. Sci. Eng.*, vol. 5, no. 4, pp. 276–287, 2017.

[132] 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 - EMNLP ’02*, Not Known, 2002, vol. 10, pp. 79–86. doi: 10.3115/1118693.1118704.

[133] B. Liu and L. Zhang, “A Survey of Opinion Mining and Sentiment Analysis,” in *Mining Text Data*, C. C. Aggarwal and C. Zhai, Eds. Boston, MA: Springer US, 2012, pp. 415–463. doi: 10.1007/978-1-4614-3223-4_13.
[134] “GAY (adjective) definition and synonyms | Macmillan Dictionary.” https://www.macmillandictionary.com/dictionary/british/gay_1 (accessed Jan. 23, 2022).
[135] D. Maynard, K. Bontcheva, and D. Rout, “Challenges in developing opinion mining tools for social media,” *Proc. NLP Can U Tag Usergeneratedcontent*, pp. 15–22, 2012.
[136] “ECSM Future and Past • Academic Conferences International,” *Academic Conferences International*. https://www.academic-conferences.org/conferences/ecsm/ecsm-future-and-past/ (accessed Jan. 22, 2022).

Eigenanteil an den Publikationen

Publikation	Idee	Konzeption	Theorie	Datensammlung	Untersuchung	Überarbeitung
Gehrke, S., Wenige, L. and Ruhland, J., 2019, June. Feasibility Study of Analysis of Senior IT Management Skills/Qualifications in Social Networks. In *ECSM 2019 6th European Conference on Social Media* (p. 324). Academic Conferences and publishing limited.	***	***	***	***	***	***
Gehrke, S., Opuszko, M., Niemz, S. and Ruhland, J., 2021, July. Comparison of Social Media and Panel Data Analyses. In *ECSM 2021 8th European Conference on Social Media* (p. 276). Academic Conferences Inter.	***	***	***	***	***	***
Kirchner, K., Opuszko, M. and Gehrke, S., 2021. Decision Support in the Era of Social Media and User-Generated Content. In *EURO Working Group on DSS* (pp. 79-95). Springer, Cham.	**	**	**	**	**	**
Gehrke, S. and Ruhland, J., 2022, Mai. Trusting the Big Data Analytics Process from the Perspective of Different Stakeholders, in DATA ANALYTICS 2022, The Eleventh International Conference on Data Analytics	***	***	***	***	***	***

*** federführend, ** proportional, * gering, x kein Anteil