Universität Rostock
Traditio et Innovatio

Aus der Professur für Geodäsie und Geoinformatik
der Agrar- und Umweltwissenschaftlichen Fakultät

Extraktion und Auswertung von Geodaten aus Sozialen Netzwerken als Element der Bürgerbeteiligung in kommunalen Belangen der Hansestadt Rostock

Dissertation
zur Erlangung des akademischen Grades
Doktor der Ingenieurwissenschaften (Dr.-Ing.)

an der Agrar- und Umweltwissenschaftlichen Fakultät
der Universität Rostock

vorgelegt von
M.Sc. Ferdinand Vettermann
aus
Rostock

Rostock, 01.04.2019

https://doi.org/10.18453/rosdok_id00002539

Dieses Werk ist lizenziert unter einer
Creative Commons Namensnennung - Nicht kommerziell - Keine Bearbeitungen 4.0 International Lizenz.

Gutachter:

Prof. Dr. Ralf Bill, Universität Rostock, AUF, Geodäsie und Geoinformatik, Rostock

Prof. Dr. Jens Tränckner, Universität Rostock, AUF, Wasserwirtschaft, Rostock

Prof. Dr. Jukka Krisp, Universität Augsburg, Institut für Geographie, Prof. für Angewandte Geoinformatik, Augsburg

Prof. Dr. René Westerholt, University of Warwick, Centre for Interdisciplinary Methodologies, Warwick (UK)

Jahr der Einreichung: 2019

Jahr der Verteidigung: 2019

Danksagung

Zuvorderst sei all denen gedankt, die diese Arbeit ermöglicht und in ihrem Verlauf stets mit Rat und Tat unterstützt haben. An erster Stelle steht dabei Herr Prof. Dr. Ralf Bill, der sowohl maßgeblich zur Themenfindung beigetragen als auch bei der Erstellung der Arbeit geholfen hat. Herr Prof. Dr. Ralf Bill hatte für meine Fragen und Probleme immer ein offenes Ohr.

Da sich über 160 Seiten jedoch nicht von alleine schreiben, möchte ich meiner Freundin Anna danken. Sie war mir immer eine Stütze und hat mich stets neu motivieren können, trotz des Verzichts auf viel gemeinsame Zeit. Ohne ihre Hilfe wären die Hindernisse, die immer wieder im Entstehungsprozess einer solchen Arbeit auftreten, wohl kaum zu überwinden gewesen. Und auch ohne das Lachen, das aufmunternde Krietschen und das Strahlen im Gesicht meiner Tochter Tilda, wäre die Arbeit wohl deutlich mühseliger gewesen als sie es war.

Daneben seien zudem die Kollegen der Arbeitsgruppe erwähnt, die bei Fragen stets geholfen haben und das Fluchen ertragen mussten, wenn die Dinge mal nicht so liefen, wie ich es mir eigentlich vorgestellt hatte. Neben meinen Eltern, die durch ihre Korrekturhilfe der Arbeit noch den letzten Schliff gaben, möchte ich Herrn Prof. Dr. Jens Tränckner, Herrn Prof. Dr. Jukka Krisp sowie Herrn Prof. Dr. René Westerholt für die Begutachtung der Arbeit danken. Und selbstverständlich müssen an dieser Stelle noch Enrico und Andreas erwähnt werden, die bei den großen und kleinen Schwierigkeiten mit der IT immer wieder helfend zur Seite standen.

Zusammenfassung

Im Rahmen dieser Arbeit ist eine Methode entwickelt worden, die es sowohl in einer hinsichtlich Twitter nachrichtenarmen Region wie der Hanse- und Universitätsstadt Rostock als auch im deutschsprachigen Raum ermöglicht, Tweets auf einer lokalen Skale zu verorten, sie vordefinierten Themen zuzuordnen und hinsichtlich Stimmungen und Trends zu analysieren. Über die Verbindung zwischen physischer und digitaler Welt wird es möglich, wertvolle Informationen aus den Sozialen Netzen in die physische Umwelt zu überführen.

In einem vergleichsweise kurzen Zeitraum von sechs Wochen (06.08.2018 - 30.09.2018) war die entwickelte Methode in der Lage, 29 771 Tweets mit Bezug zu Rostock zu lokalisieren. Der Algorithmus konnte hierbei 27 % aller Nachrichten mittels eines mehrschichtigen, text-basierten Gazetteers mindestens auf der Ebene von Straßen oder genauer mit einer Gesamtgenauigkeit von 83 % verorten.

Außerdem konnte gezeigt werden, dass die SVM (*Support Vector Machine*) im Rahmen der Sentimentanalyse in dieser Arbeit bessere Ergebnisse liefern als CNNs (*Convolutional Neural Networks*) oder LSTMs (*Long Short Term Memory Networks*). Mittels der Identifikation von Trends und Themen über TF-IDF (*Term Frequency – Inverse Document Frequency*) und LDA (*Latent Dirichlet Allocation*) wird zudem die Möglichkeit geschaffen, Tweets zu kategorisieren und einzelne Events zu identifizieren. Jedoch ließen sich, entgegen des ursprünglichen Ziels, keine konkreten Handlungsempfehlungen hinsichtlich der Stadtplanung ableiten.

Die hochgenaue Geolokalisation ermöglicht es, attraktive sowie unattraktive Orte zu identifizieren und damit eine Datengrundlage für das Stadtmarketing zu bieten, um die Attraktivität für Touristen, Bevölkerung und Wirtschaft zu erhöhen. Des Weiteren ist es möglich, Echtzeit-Entscheidungsunterstützung, beispielsweise für Extremereignisse und Veranstaltungen zu leisten. Die schnelle Reaktionszeit der Twitternutzer konnte am Beispiel der Demonstration der AfD (Alternative für Deutschland) am 22.09.2018 in Rostock verifiziert werden. Aus der spatio-temporalen Kombination der Daten lassen sich so direkt Handlungsempfehlungen für Sicherheitskräfte ableiten.

Abstract

Within this work, a method has been developed which is able to geolocate tweets on a local scale, assign given topics and analyze them for trends and moods. It is special that this is settled in a German speaking area and a region with a low tweet density, the Hanseatic and university city of Rostock. Through the connection to the analogue world, valuable information can be generated and transferred from twitter.

In a relatively short period of six weeks (06.08.2018 - 30.09.2018) the developed algorithm was able to filter and locate 29 771 messages related to the city of Rostock. The procedure was able to place a total of 27 % of the messages with a text based multi-layer gazetteer matching with an overall accuracy of 83 % at least at road level.

Moreover, the work has shown that in this case sentiment classification with SVM (Support Vector Machine), seems more accurate than CNNs (Convolutional Neural Networks) or LSTMs (Long Short Term Memory Networks). In addition, the identification of trends and topics via TF-IDF (Term Frequency – Inverse Document Frequency) and LDA (Latent Dirichlet Allocation) enables the possibility, to categorise the tweets and identify certain events. However, no concrete recommendations for action with regard to urban planning can be derived, as had been hoped for.

The high-precision geolocating gives a wide range of opportunities. These include the identification of particularly popular and attractive as well as unattractive places. From this, potentially recommendations regarding city marketing can be derived to increase the attractiveness for tourism, for the population but also for the economy. Furthermore, it is possible to derive real time recommendations for action for security forces from the spatio-temporal combination of the data. The fast reaction time has been illustrated at the demonstration of the AfD (Alternative für Deutschland) in Rostock on 22.09.2018.

Inhalt

Abkürzungsverzeichnis IX
1 Einleitung, Motivation und Zielsetzung 1
1.1 Einleitung 1
1.2 Motivation 2
1.3 Zielsetzung 3
1.4 Aufbau der Arbeit 5
2 Soziale Medien und GIS 7
2.1 Das Web 2.0 7
2.2 Crowdsourcing 8
2.3 Volunteered Geographic Information 10
2.4 Citizen Science 11
2.5 Open Data 12
2.6 Moderne Verfahren der Bürgerbeteiligung 13
2.7 Soziale Netzwerke 15
2.7.1 Twitter – der Mikroblogging-Dienst 17
2.7.2 Foto-Sharing 18
2.7.3 Soziale Interaktion und Selbstdarstellung 19
2.7.4 Weitere Soziale Netzwerke 19
2.8 Geodateninfrastrukturen 20
2.8.1 Normen 21
2.8.2 OGC-Spezifikationen 22
2.9 Location-based Social Networks 23
2.9.1 Kategorisierung von LBSNs 24
2.9.2 Wirtschaftliches Interesse an LBSNs 25
2.10 Erhebung von Geodaten in Sozialen Netzwerken 25
3 Big Data Analyseverfahren 27
3.1 Big Data 27
3.2 Datensicherheit und Datenschutz 28
3.3 Analyseverfahren 30
3.4 Text Mining in Sozialen Netzwerken 31
3.4.1 Vorprozessierung 32
3.4.2 Maschinelles Lernen 32
3.4.3 Ableitung von Ortsbezügen 40
3.4.4 Sentimentanalyse 43
4 Programmier- und Softwareaspekte 47

4.1 Datenbanken und Programmierung ... 47
4.1.1 PostgreSQL ... 47
4.1.2 Python ... 47
4.2 Programmschnittstellen ... 48
4.2.1 Twitter Streaming API ... 48
4.2.2 Twitter REST API ... 51
4.3 Geodateninfrastrukturen ... 52
4.3.1 OpenLayers 3 ... 52
4.3.2 GeoNetwork ... 53
4.3.3 GeoServer ... 54
4.4 Auswertung und Visualisierung ... 55
4.4.1 GIS-Software ... 55
4.4.2 Gephi ... 55
5 Das Projekt KOGGE und die Hansestadt Rostock als Untersuchungsgebiet ... 57
5.1 Das Projekt KOGGE ... 57
5.2 Die Hansestadt Rostock ... 58
5.3 Geografische, wirtschaftliche und soziokulturelle Hotspots ... 59
5.4 Bevölkerungsentwicklung und -struktur ... 61
5.5 Bürgerbeteiligung und -information in Rostock ... 63
5.5.1 Klarschiff.HRO ... 63
5.5.2 OpenData.HRO ... 64
5.5.3 Zahlungsbereitschaftsanalyse zu den Klein- und Kleinstgewässern ... 65
6 Daten und Methodik ... 69
6.1 Aufbau einer GDI ... 69
6.1.1 GeoNetwork ... 69
6.1.2 GeoServer ... 71
6.2 Social Media Harvesting ... 71
6.2.1 Datenbankerstellung ... 71
6.2.2 Gazetteererstellung ... 73
6.2.3 Filterstrategie und Harvesting ... 74
6.2.4 Textaufbereitung und Normalisierung ... 77
6.2.5 Themenidentifikation ... 79
6.2.6 Trendidentifikation ... 80
6.2.7 Sentimentanalyse ... 80
6.2.8 Expertenfindung ... 87
6.2.9 Gazetteer-Matching ... 88
6.3 Front-End und Visualisierung ... 88

7 Ergebnisse 91
7.1 Allgemeine statistische Eigenschaften 91
7.2 Spatio-Temporale Verteilung 92
7.2.1 Temporale Verteilung 93
7.2.2 Räumliche Verteilung 94
7.2.3 Sentimentanalyse 97
7.2.4 Kombinierte raum-zeitliche Analyse 101
7.3 Trending Topics 109
7.3.1 TF-IDF 109
7.3.2 LDA 112
7.4 Netzwerkdarstellung 115
8 Diskussion 119
8.1 Verortung 119
8.2 Themenklassifikation 120
8.3 Sozio-kulturelle Hotspots 123
8.4 Ereignisrezeption und Soziale Netzwerke 125
9 Zusammenfassung und Ausblick 127
10 Literatur 129
Abbildungsverzeichnis 145
Tabellenverzeichnis 149
Anhang 151
A - 1 Verwendete externe Bibliotheken 151
A - 2 Sentimentanalyse 152
A - 3 Nachrichtenverteilung 153
A - 4 Themenzuordnung 159
A - 5 Themenfindung 161

Abkürzungsverzeichnis

AfD	Alternative für Deutschland
AJAX	Asynchronous JavaScript XML
API	Application Programming Interface
BCP	Best Current Practice
BMBF	Bundesministerium für Bildung und Forschung
BoW	Bag of Words
CAGR	Compound Annual Growth Rate
CBOW	Continuous Bag of Words
CEN	Comité Européen de Normalisation
COBRA	Common Object Broker Request Architecture
COM	Component Object Model
CNN	Convolutional Neural Networks
CSW	Catalogue Service Web
CTS	Coordinate Transformation Service
BMVI	Bundesministerium für Verkehr und Digitale Infrastruktur
DAG	Directed Acyclic Graph
DAI	Distributed Artificial Intelligence
DAU	Daily Active User
DIN	Deutsche Industrie Norm
GDI	Geodateninfrastruktur
GDF	Geographic Data File Format
GML	Geography Markup Language
GNSS	Global Navigation Satellite System
GPS	Global Positioning System
HMM	Hidden Markov Model
IE	Information Extraction
IGA	Internationale Gartenausstellung
IoT	Internet of Things
IR	Information Retrieval
ISO	International Standard Organization
JSON	Java Script Object Notation

KML	Keyhole Markup Language
kNN	k-Nearest Neighbour
KTV	Kröpeliner-Tor-Vorstadt
LBSN	Location Based Social Network
LDA	Latent Dirichlet Allocation
LIW	Location Indicative Word
LVCSR	Large Vocabulary Continuous Speech Recognition
LSTM	Long Short Term Memory Networks
NASA	National Aeronautics and Space Administration
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NN	Neuronale Netzwerke
NSDI	National Spatial Data Infrastructure
MAU	Monthly Active User
MCC	Multiversion Concurrency Control
MGS	Multilingual Sentiment Corpus
OGC	Open Geospatial Consortium
OL	OpenLayers
OL3	OpenLayers 3
OLE	Object Linking and Embedding
ON	Österreichischer Normenausschuss
OpenLS	Location Service Implementation Specification
ORDBMS	Objekt-Relationales Datenbankmanagementsystem
OSM	Open Street Map
POI	Point of Interest
PotTS	Potsdam Twitter Sentiment Corpus
RegEx	Regular Expressions
REST	Representational State Transfer
RDF	Ressource Description Framework
SDI	Spatial Data Infrastructure
SKOS	Simple Knowledge Organization System
SLD	Styled Layer Description

SMOTE	Synthetic Minority Over-sampling Technique
SMS	Short Messaging Service
SNV	Schweizerische Normen-Vereinigung
SOS	Sensor Observation Service
SVD	Singular Value Decomposition
SVM	Support Vector Machines
TC	Technical Committee
TF-IDF	Term Frequency – Inverse Document Frequency
TOPP	The Open Planning Project
t-SNE	t-Distributed Stochastic Neighbour Embedding
UGC	User Generated Content
UTC	Coordinated Universal Time
UTF	Unicode Transformation Format
VGI	Volunteered Geographic Information
WAS	Web Authentication Service
WCS	Web Coverage Service
WFS	Web Feature Service
WFST	Transactional Web Feature Service
WMS	Web Map Service
WPS	Web Processing Service
WPOS	Web Pricing and Ordering Service
WTS	Web Terrain Service
WWW	World Wide Web
W3DS	Web 3D Service
XML	eXtensible Markup Language

1 Einleitung, Motivation und Zielsetzung

„Meine Art ist es, am Anfang zu beginnen. “

Lord George Gordon Noel Byron (1788 - 1824), englischer Dichter der Romantik

1.1 Einleitung

Die Medienlandschaft wurde seit dem Jahr 2000 gehörig durcheinander gewirbelt. Mit dem Aufkommen des Internets hat sie sich radikal verändert. Die herkömmlichen Printmedien, aber auch TV und Radio sehen sich großen Herausforderungen ausgesetzt und müssen sich gegen eine immer stärkere Konkurrenz erwehren (SEVENONE MEDIA GMBH 2018). Dabei scheint ihnen der „Gegner“ in seiner Reichweite und seinem Einfluss immer weiter voraus zu sein: Facebook, Twitter, Google, Netflix und Co (Abbildung 1-1) (FREES & KOCH 2018).

Gerade die Online Social Networks haben seit MySpace, spätestens seit Facebook einen gigantischen Siegeszug begonnen. Allein das Facebook-Imperium mit Facebook selbst, Instagram und WhatsApp besitzt weltweit 2.7 Mrd. Nutzer, von denen etwa 2 Mrd. täglich aktiv sind – Tendenz steigend (FACEBOOK INC. 2019). Dazu kommt Twitter, welches zwar nur 320 Million Monthly Active User (MAU) besitzt und seit einigen Jahren stagniert, dafür aber mit 500 Mio. täglich Nachrichten eine riesige Datenbasis schafft (TWITTER INC. 2019, INTERNET LIVE STATS 2019). Durch das Web 2.0 und das immer mehr in den Vordergrund-treten des User Generated Content (UGC) sind dem Internet mit zahllosen Blogs weitere Konkurrenten entsprungen, die die Kundschaft der alten Medien adressieren und damit deren Geschäftsmodell nach und nach erodieren (SEVENONE MEDIA GMBH 2018).

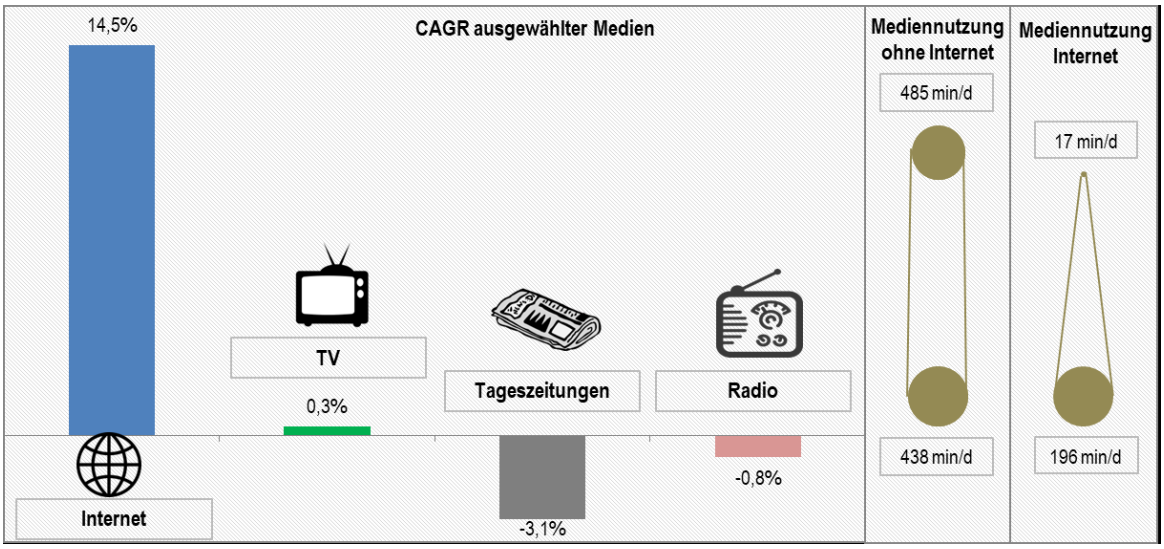


Abbildung 1-1: Compound Annual Growth Rate (CAGR) ausgewählter Medien in Deutschland im Verhältnis zur täglichen Gesamtmediennutzung (eigene Erstellung, Daten kombiniert aus FREES & KOCH 2018, GIERSBERG & LEIBIGER 2019, SEVENONE MEDIA GMBH 2018, EIMEREN & RIDDER 2001, GATTRINGER & TURUCEK 2018, KLINGLER & TURUCEK 2016).

Dennoch ist gerade in Deutschland das Potential dieses „Datenschatzes“, wohl vor allem bedingt durch Vorbehalte hinsichtlich des Datenschutzes, kaum im öffentlichen Bewusstsein vorhanden oder wird negativ konnotiert (BERGMANN 2009, DOBUSCH 2014). Dies schlägt sich auch in der Forschung nieder. Arbeitet man an diesen Themen, lautet in der Regel der erste Kommentar, dass dies wohl für die Geheimdienste sehr nützlich sei. Dabei kann dieser ungenutzte „Datenschatz“ gerade in einem Bereich, welcher für die Öffentlichkeit hinsichtlich Lebensqualität einen großen, wenn nicht gar den größten Einfluss besitzt, zur Entscheidungsfindung beitragen, wie zum Beispiel bei der Stadtplanung und der dazugehörigen Entscheidungsunterstützung. Hierbei geht es nicht um die Überwachung der Bürger, die durchaus bewusst Informationen über Twitter und Co. frei verfügbar machen und auch wollen, dass diese wahrgenommen werden, sondern darum, deren Meinungen und Verhalten sowohl im Planerischen als auch bei Veranstaltungen und Extremereignissen zu berücksichtigen (FUCHS et al. 2013; LONGLEY & ADNAN 2015; RESCH et al. 2018 u. v. m.). Ziel ist es also letztlich, den Bürger mittels der neuen Technologien des Web 2.0 besser in die Entscheidungsfindungsprozesse einzubinden.

Vor diesem Hintergrund beschäftigt sich die vorliegende Arbeit mit der Frage, wie sich unter Einbeziehung der Datenflut aus den Sozialen Netzen eine Datenbasis erstellen lässt und Algorithmen entwickeln lassen, die notwendige und aktuelle Informationen als Grundlage für Handlungsanweisungen im öffentlichen Raum verfügbar machen. Aus diesen sollen sich sozio-kulturelle Hotspots, aber vor allem auch Einzelereignisse (Sturmfluten, extreme Gewitter etc.) und Großveranstaltungen wie z. B. die Hanse Sail in Rostock, identifizieren lassen. Außerdem soll über eine Webpräsenz ein Überblick über Hotspots und diskutierte Themen gegeben werden. In das Gesamtkonzept sollen dabei neue Technologien des maschinellen Lernens und der neuronalen Netze einbezogen werden, um den aktuellen Forschungsergebnissen in der Computerlinguistik Rechnung zu tragen.

1.2 Motivation

Da Soziale Medien und vor allem geolokalisierte Informationen in Sozialen Medien bedingt durch die „Generation Smartphone“ (auch head down generation) (KNAUß 2014) eine immer größere Verbreitung finden, ist es nur folgerichtig, diese Informationen auch im Rahmen der Bürgerbeteiligung und -information zu nutzen. Gerade im wissenschaftlichen deutschsprachigen Kontext scheint es hier ein Defizit zu geben, da sich nur eine Handvoll Autoren mit derartigen Fragestellungen beschäftigen (FUCHS et al. 2013, SCHEFFLER 2014, GONTRUM & SCHEFFLER 2015, BUSCHBAUM et al. 2017). Da die Hansestadt Rostock bereits einige Anstrengungen hinsichtlich Bürgerbeteiligung und -information unternommen hat, gibt es hier einen guten Ansatzpunkt, diesen Umfang weiter zu vergrößern und die Datenbasis zu verbessern.

Da diese Arbeit im Kontext der Geoinformatik anzusiedeln ist, liegt entsprechend der Fokus auf der Geokodierung der Nachrichten. Die Ortsinformation ist essentiell, um die Daten aus sozialen Netzwerken in einen entsprechenden Kontext zu setzen. Gerade während Extremszenarien, aber auch bei Events und zur Analyse sozialer Fragestellungen ist das „Wo?“ von besonderer, wenn nicht gar entscheidender Wichtigkeit. Im Verlauf des Projektes KOGGE (Kommunale Gewässer Gemeinschaftlich Entwickeln; vgl. Kap. 5.1) ergab sich vor allem die Notwendigkeit der Identifikation von hydrologischen Problemstellen bzw. von einzelnen Ereignissen. Gerade aufgrund der unzureichenden Dichte des hydrodynamischen Messnetzes sind zusätzliche Informationen über den Ort eines Ereignisses von großer Bedeutung. Da diese Ergebnisse i. d. R. sehr publikumswirksam (Abbildung 1-2) sind, ist davon auszugehen, dass hier auch in den Sozialen Medien Informationen darüber geteilt werden.

Neben Extremereignissen sind aus Planungssicht auch Meinungen und Stimmungen der Bevölkerung in einzelnen Stadtteilen zu spezifischen Themen von Bedeutung. Beispielhaft sei hier die geplante Erweiterung Rostocks um den Stadtteil Neu-Biestow für etwa 25 000 Einwohner genannt oder aber auch die Großdemonstration der AfD (Alternative für Deutschland) im Stadtzentrum Rostocks (KÖRNERF 2016, NDR 2018). Es ist davon auszugehen, dass in den Sozialen Medien darüber entsprechend diskutiert wird und die Meinung dazu zum Ausdruck kommt. U. U. kommen hierbei sogar Ideen zum Ausdruck, die vorher so in der Planung keine Berücksichtigung fanden.

Abbildung 1-2: Überschwemmungen während des Sturmtiefs Axel am 04.01.2017 (OSTSEEZEITUNG 2017).

Schlussendlich ist auch von Feedbackeffekten auszugehen. Existiert ein Portal, in dem die Einwohner der Hansestadt ihre Diskussionen und Meinungen zu dezidierten Themen einsehen können, steigt die Bereitschaft, hierzu Informationen, gerade hinsichtlich Extremereignissen, zu teilen, um so die Behörden auf Problemstellen hinzuweisen. Für die Entscheidungsunterstützung ist vor allem das echtzeitnahe Auflaufen der Nachrichten und deren Visualisierung von großer Bedeutung. So kann ohne eine aufwendige Analyse von Tabellen direkt eingesehen werden, wo sich Hotspots befinden, ob bestimmte Schlagwörter trenden oder welche Bedeutung die in den Sozialen Netzwerken aktive Bevölkerung diesen Themen zuweist – Stichwort der Betroffenheit.

1.3 Zielsetzung

Aus den genannten Vorüberlegungen lassen sich nun konkrete Ziele in Form von Thesen formulieren, die im Rahmen dieser Arbeit erreicht bzw. diskutiert werden sollen.

These 1 Die ubiquitäre Nutzung von Sozialen Netzwerken führt zu einer ausreichend hohen Nachrichtendichte um Ortsbezüge in einer mittelgroßen Stadt wie Rostock herzustellen.

Ein wesentlicher Punkt der Arbeit ist es, eine ausreichend hohe sowie repräsentative Nachrichtendichte für eine Stadt mit einer verhältnismäßig geringen Einwohnerzahl zu erhalten. Dies stellt eine Art Grundvoraussetzung dar, weshalb eine möglichst große Anzahl potentieller Nachrichten mit dem Bezug zur Hansestadt Rostock abgefangen werden sollen und müssen.

These 2 Die Genauigkeit von geolokalisierten Informationen aus Sozialen Netzwerken lässt sich bis auf die Ebene *Straße* herunterbrechen.

Ein häufiges Problem bei vergleichbaren Analysen ist, dass sie sich nur auf geokodierte Nachrichten beziehen. Mit Berücksichtigung von These 1 ist davon auszugehen, dass der Anteil dieser Nachrichten für die Hansestadt für eine ausreichend hohe Nachrichtendichte zu gering sein könnte. Aus diesem Grund sollen andere Methoden entwickelt werden, deren Tauglichkeit zur hochgenauen Verortung schließlich analysiert werden soll. Im Vordergrund steht dabei die Analyse des Nachrichtentextes.

These 3 Mit einem Gazetteer-Dienst, welcher auch lokale Ortsbezeichnungen umfasst und damit das Vorortwissen über Plätze in einer Stadt beschreibt, lässt sich die Genauigkeit von geolokalisierten Nachrichten in Sozialen Netzwerken signifikant erhöhen.

Grundidee ist es, über eine Datenbank, welche möglichst viele Ortsbezeichnungen der Stadt Rostock enthält, eine Verortung der Nachrichten durchzuführen. Im Rahmen der Arbeit soll die Fähigkeit dieser Methode zur hochgenauen Verortung aufgezeigt werden.

These 4 Über die Textanalyse von Nachrichten in Sozialen Netzwerken lassen sich Bedürfnisse und Wünsche der Bevölkerung ableiten.

Ein vieldiskutiertes und aktuelles Thema bezüglich der Analyse von Sozialen Netzwerken ist die Analyse von Stimmungen, auch Sentimentanalyse. Es soll geprüft werden, ob sich ebendiese Meinungen zu bestimmten, räumlich verortbaren Themen für die Hansestadt Rostock ableiten lassen.

These 5 Mit Hilfe von Sozialen Netzwerken lassen sich konkrete Handlungsempfehlungen im Bereich der kommunalen Planung aussprechen.

Ein wesentlicher Aspekt der Arbeit ist es, zu prüfen ob sich tatsächlich auch Handlungsempfehlungen aus den sozialen Medien ableiten lassen. Im Kern geht es darum, zu prüfen ob die Bürger z. T. konkrete Wünsche hinsichtlich der Stadtplanung äußern oder aber die Ableitung letztlich über die statistische Analyse der Daten und der visuellen Aufbereitung durchgeführt werden kann und muss.

These 6 Geolokalisierte Informationen aus Sozialen Netzwerken können dazu genutzt werden, touristisch oder kulturell-sozial relevante Orte in Rostock hervorzuheben.

Ein neben den Extremereignissen interessantes Ergebnis wären Heatmaps, um zu identifizieren, welche Gebiete der Stadt Rostock zu welchen Zeiten für bestimmte Bevölkerungsgruppen besonders attraktiv sind.

These 7 Über die Analyse von Sozialen Netzwerken lassen sich einzelne Events und das Interesse der Bevölkerung an selbigen visualisieren.

Neben dem normalen Tages- bzw. Wochenverlauf soll geprüft werden, ob einzelne Ereignisse wie Silvester, die Hanse Sail oder beispielsweise die Warnemünder Woche sowie sportliche oder kulturelle Ereignisse, deren Häufigkeit größer als die von Extremereignissen ist, aus den Nachrichten der Sozialen Netzwerke ersichtlich werden.

1.4 Aufbau der Arbeit

Um die in der Einleitung (Kap. 1.3) gesetzten Ziele und aufgestellten Thesen strukturiert abzuarbeiten, zu bestätigen oder zu widerlegen ist eine strukturelle Herangehensweise erforderlich. Aus diesem Grund soll in diesem Kapitel der Aufbau der Arbeit kurz umrissen werden.

Nach der Einleitung folgt in Kapitel 2 ein Überblick darüber, wie die Technologien des Web 2.0 die Strukturen des Internets verändert haben und welche Möglichkeiten sich daraus ergeben. Im Fokus steht auch, wie aktuell die Bürger in die Planungsprozesse eingebunden werden. Dazu zählt neben der Aufarbeitung der Schlagwörter Web 2.0 (Kap. 2.1), Crowdsourcing (Kap. 2.2), Volunteered Geographic Information (VGI, Kap. 2.3), Citizen Science (Kap. 2.4) und Open Data (Kap. 2.5) auch, wie aktuell Bürgerbeteiligung gestaltet werden kann und wird (Kap. 2.6). Des Weiteren wird auf die bedeutendsten Sozialen Netze eingegangen (Kap. 2.7). Anschließend werden moderne Methoden und Standards des Geodatenmanagements, der Bereitstellung von Web Services sowie Webmaps aufgezeigt (Kap. 2.8). Abschließend wird es darum gehen, wie all diese Technologien zu den Location Based Social Networks (LBSNs) geführt haben, welche letztlich die Schnittstelle zwischen Raum, Zeit und Nutzer darstellen (Kap. 2.9). Zuletzt findet eine Analyse statt, in welcher Art Geodaten überhaupt in Sozialen Netzen vorkommen (Kap. 2.10).

In Kapitel 3 wird auf konkrete Datenverarbeitungsmethoden im Bereich von Big Data eingegangen. Neben der Beschreibung des Begriffes selbst (Kap. 3.1) stehen hier die aktuelle Diskussion zum Thema Datenschutz und Privatsphäre (Kap. 3.2), aktuelle Analyseverfahren (Kap. 3.3) sowie ganz konkret die Textanalyse und Textverarbeitung im Vordergrund (Kap. 3.4).

Einleitend in die methodische Vorgehensweise werden zuerst die verwendeten Softwareumgebungen beschrieben (Kap. 4). Eingangs wird auf die Schnittstellen und die hauptsächlich verwendeten Programmiersprachen eingegangen (Kap. 4.1), folgend werden die Programmschnittstellen vorgestellt (Kap. 4.2), dann die Geodateninfrastrukturen und deren Eigenschaften erläutert (Kap. 4.3) und abschließend die Programmbibliotheken zur Auswertung und Visualisierung aufgezeigt (Kap. 4.4).

Kapitel 5 geht auf die Besonderheiten der Hansestadt Rostock als Untersuchungsgebiet näher ein. Eingangs wird der Bezug zum Projekt KOGGE hergestellt (Kap. 5.1). In Kapitel 5.2 liegt der Fokus auf der Hanse- und Universitätsstadt Rostock, um anschließend geografische, wirtschaftliche und kulturelle Hotspots aufzuzeigen (Kap. 5.3). Folgend steht die aktuelle Bevölkerungsstruktur und Entwicklung im Vordergrund (Kap. 5.4) als auch Verfahren, wie derzeit in Rostock Bürgerbeteiligung und Information realisiert wird (Kap. 5.5).

Nach dem Abschluss des theoretischen Grundgerüsts wird in Kapitel 6 auf die konkreten, verwendeten Datensätze und die Methodik bei deren Erhebung und Verarbeitung eingegangen. Zuerst ist der Blick auf den Aufbau einer Geodateninfrastruktur (GDI) gerichtet (Kap. 6.1), anschließend liegt der Fokus auf dem Harvesting von Daten aus den Sozialen Netzwerken (Kap. 6.2) gefolgt von der Visualisierung der Daten in naher Echtzeit (Kap. 6.3).

Der Ergebnisteil beginnt mit allgemeinen statistischen Eigenschaften der erhobenen Daten im spezifizierten Untersuchungszeitraum (Kap. 7.1). Ausgehend davon steht dann die spatio-temporale Verteilung der Daten (Kap. 7.2), die Analyse und Klassifizierung der Daten hinsichtlich Trends und Themen (Kap. 7.3) und schließlich die Analyse des Sozialen Netzes Rostocks im Vordergrund (Kap. 7.4).

Gefolgt werden die Ergebnisse von der kritischen Betrachtung und Diskussion der selbigen. Im Fokus stehen hier zunächst der Algorithmus und die Genauigkeit der Verortung (Kap. 8.1). Nachfolgend wird die Zuordnung und Identifikation von vieldiskutierten Themen in Rostock aufgegriffen und noch einmal näher betrachtet (Kap. 8.2). Eng damit verbunden ist die Analyse der identifizierten sozio-kulturellen Hotspots (Kap. 8.3) und die Ereignisrezeption (Kap. 8.4).

Beendet wird der Hauptteil mit der abschließenden Zusammenfassung (Kap. 9), welcher noch das Literatur-, das Abbildungs- sowie das Tabellenverzeichnis und schließlich der Anhang folgen.

2 Soziale Medien und GIS

„Das Web ist mehr eine soziale Erfindung als eine technische."

Tim Berners-Lee (1955), britischer Physiker und Informatiker

2.1 Das Web 2.0

Soziale Netzwerke und Medien stellen einen essentiellen Bestandteil des Web 2.0 dar. Aus diesem Grund ist zuallererst zu klären, um was es sich bei diesem Begriff eigentlich handelt und was das Web 2.0 umfasst.

Die am weitesten verbreitete Beschreibung stammt von O'REILLY (2005), weshalb hier insbesondere auf diese Bezug genommen werden soll. Das Web 2.0 ist aus den Trümmern der Dotcom Blase zu Beginn der 2000er Jahre geboren worden. Am deutlichsten werden die Eigenschaften an einigen Beispielen, die O'REILLY (2005) aufführt, um zu veranschaulichen, wie sich die Inhalte des Internets von Web 1.0 zu Web 2.0 verändert haben (Tabelle 2-1). So wird ersichtlich, dass sich das ursprüngliche Internet von seiner Funktion, Informationen zentral zu liefern, verabschiedet hat und vielmehr zu einem Instrument öffentlicher Teilhabe geworden ist. Besonders deutlich wird dies sowohl an Blogs aber auch an Wikis und letztlich den verschiedenen Sozialen Netzwerken, die hier ansetzen. Das „neue" Internet kann somit vielmehr als Plattform diverser Dienste und Funktionen verstanden werden und ist dezentral organisiert. Es stellt also die Verdrahtung einzelner zu einer kollektiven Intelligenz dar (O'REILLY & BATTELLE 2009).

Folglich stellen die Entwicklungen des Web 2.0 die Basis für die Sozialen Netzwerke (vgl. Kap. 2.7), wie sie heute existieren, dar. Häufig wird das Web 2.0 mit Sozialen Netzwerken gleichgesetzt. Die Methoden und Ideen beziehen sich jedoch auf ein viel breiteres Spektrum der Online-Interaktion. So sind Ideen wie VGI (vgl. Kap. 2.3), das Crowdsourcing (vgl. Kap. 2.2) sowie die vielfältigen Aspekte der Citizen Science (vgl. Kap. 2.4) in ihrer aktuellen Ausprägung erst durch das Web 2.0 ermöglicht worden. Der Kern der Services, die über das Web angeboten und großteils durch seine Nutzer selbst befüllt werden, sind schließlich die Daten. O'REILLY (2005) bezeichnet daher die Daten auch als den neuen „*Intel Inside*" der aktuellen Generation von Computeranwendungen.

Dieser Entwicklung Vortrieb leisteten insbesondere die Smartphones, welche durch ihre vielfältigen Sensoren (Kameras, Global Navigation Satellite System - GNSS, Schrittzähler, …) und der räumlich weitestgehend unabhängigen Verfügbarkeit eines Internetzugangs vielfältige Informationen (auch standortunabhängig) abrufbar und teilbar machten und machen. Hinzu kommen Metadaten, wie z. B. Freundschaftsbeziehungen in Sozialen Netzwerken, welche den Wert der Informationen weiter steigern (O'REILLY & BATTELLE 2009). Hieraus wurden Konzepte wie Humans as Sensors sowie Internet of Things (IoT) geboren. Aufgrund des großen Datenaufkommens werden schlussendlich neue Herangehensweisen erforderlich, wie sich dieses verarbeiten lässt. Dazu zählen die vielfältigen Algorithmen im Bereich Big Data (vgl. Kap. 3).

All diese Elemente können und sollen schließlich Anwendung im Bereich urbaner Planungsprozesse finden, da hier eine große Datenbasis erhoben werden kann. So lassen sich durch Online-Informationen Rückschlüsse und Handlungsempfehlungen geben, wie urbane Räume gestaltet werden können und worauf die Menschen Wert legen. Dies ist insbesondere dadurch möglich, dass die in der physischen Welt existierenden Objekte

einen Schatten im Cyberspace werfen (KUNIAVSKY & CREECH 2009). Daneben bieten sich Möglichkeiten durch die Informationsbereitstellung in Echtzeit. Durch das Vorhandensein von Sensoren und Diensten an allen Orten ist es auf diesen Wegen möglich, Informationen deutlich schneller zu erfassen und zu teilen als dies herkömmlich (Polizei, Reporter) stattgefunden hat (O'REILLY & BATTELLE 2009). Gerade in Notsituationen können hier Soziale Medien eine wichtige Ergänzung zur Identifikation von Problempunkten und Einsatzplanung darstellen (TERPSTRA et al. 2012).

Tabelle 2-1: Beispiele der Veränderung der Inhalte von Web 1.0 zu Web 2.0 (O'REILLY 2005).

Web 1.0	**Web 2.0**
Ofoto	Flickr
Akamai	BitTorrent
Britannica Online	Wikipedia
Persönliche Internetseiten	Blogs
Veröffentlichungen	Partizipation
Content Management Systeme	Wikis
Screen Scraping (Bildschirm schürfen)	Web Services

2.2 Crowdsourcing

Häufig werden unter dem Begriff Crowdsourcing alle Arten einer Onlinebeteiligung bezeichnet. D. h. sowohl Soziale Netzwerke, Onlineenzyklopädien wie *Wikipedia*[1] oder auch die Amazon Plattform *Mechanical Turk*[2] zählen dazu. Betrachtet man jedoch diverse Definitionen von Crowdsourcing, so wird deutlich, dass ein Großteil dieser Plattformen fälschlicherweise diesem zugeordnet wird. Entstanden ist das Konzept in den frühen 2000er Jahren. Der Begriff wurde durch HOWE (2006) geprägt. Die Grundlagen wurden allerdings vor dem Beginn des Internetzeitalters geschaffen (BRABHAM 2013). BRABHAM (2013) definiert Crowdsourcing wie folgt:

> *„Crowdsourcing is a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge, and/or experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage what the user has brought to the venture, whose form will depend on the type of activity undertaken."*

Dies bedeutet, dass eine Crowdsourcing-Plattform immer ein bestimmtes Ziel im Auge hat (Top-Down), welches durch eine Vielzahl von Personen online bearbeitet wird (Bottom-Up). Der Begriff setzt sich aus Outsourcing, also Externalisieren von Arbeit, sowie der Crowd, also einer Onlinegemeinschaft, zusammen. Somit zählen Soziale Netzwerke, Plattformen wie Wikipedia oder auch historische Beispiele wie das Oxford English Dictionary in dieser strengen Definition nicht zum Crowdsourcing (BRABHAM 2013, KLEEMANN et al. 2008). Dies liegt daran, dass Projekte wie Wikipedia nur dem Bottom-Up Prinzip entsprechen, die Crowd sich selbst organisiert sowie die Artikel keinem bestimmten Ziel ent-

[1] https://www.wikipedia.de/
[2] https://www.mturk.com/mturk/welcome

sprechen. Andere Autoren hingegen führen Wikipedia häufig als das Beispiel für Crowdsourcing an. Bei ihren Definitionen wird jedoch in der Regel die Top-Down-Struktur, die kommerzielle Ausrichtung oder die explizite Ausweisung eines Ziels vernachlässigt (HAMMON & HIPPNER 2012, STREICH 2014, HOWE 2008). Allerdings werden von diesen Projekte wie Wikipedia oder auch Open Street Map (OSM) ebenfalls gesondert aufgeführt. So besitzt die Crowd nach HAMMON & HIPPNER (2012) in beiden Fällen keinen besonderen Innovationscharakter.

Eine Klassifizierung der Systeme haben DOAN et al. (2011) versucht durchzuführen. Sie trennen dabei nach expliziten und impliziten Crowdsourcing (Tabelle 2-2). Einfacher, aber auch ungenauer, ist eine Einteilung nach aktiven (die Nutzer werden aktiv aufgefordert zu kollaborieren) und passiven (die Daten werden ohne Anregung bereitgestellt) Systemen, wobei es auch zu einer Mischung kommen kann. So lassen sich beispielsweise passiv aus Sozialen Netzen Informationen gewinnen und die Ergebnisse hieraus dazu nutzen, die Menschen aktiv zu Beiträgen anzuregen (CHARALABIDIS et al. 2014).

Neben Crowdsourcing ist auch Crowdmining zu nennen. Dieses ist unmittelbar mit Crowdsourcing verbunden, da es sich dieselben Grundlagen zu Nutze macht. „Der Unterschied von Crowdsourcing und Crowdmining besteht vor allem darin, dass Crowdsourcing mit dem bewussten, aktiven und offenen Handeln von Crowd-„Mitgliedern“ einhergeht, beim Crowdmining dagegen Akteure handeln, die sich nicht offen zeigen, und der Vorgang deshalb dem einzelnen Crowd-„Subjekt“ nicht bewusst ist (STREICH 2014).“ Dass dies nicht zwingend gilt, zeigen CHARALABIDIS et al. (2014). Dabei wird deutlich, dass beide Begriffe oft vermischt und nicht klar zu trennen sind. Besonders prominente Beispiele, in denen Crowdsourcing im wissenschaftlichen Bereich zur Anwendung kommt, sind *SETI@home*[3] (Searching for Extraterrestrial Intelligence) oder *Folding@home*[4]. Bei diesen Verfahren wird die Rechenleistung der heimischen Computer durch die Crowd den jeweiligen Projekten bereitgestellt.

Tabelle 2-2: Kategorisierung von Crowdsourcingsystemen (eigene Bearbeitung, nach DOAN et al. 2011).

Typ	Architektur	Aktive Nutzersuche	Tätigkeit	Beispiele
Explizit	Standalone	Ja	Bewertungen	Amazon Produktbewertung
			Teilen von Informationen	YouTube, Flickr, Mailing Lists
			Netzwerken	Facebook, LinkedIn, Twitter
			Teilhabe an Projekten	Linux, Wikipedia
			Ausführen von Aufgaben	Cryptocurrency mining
Implizit	Standalone	Ja	Bestimmte Onlinespiele, Wetten, Captchas lösen	World of Warcraft, recaptcha.net
	Nutzung eines anderen Systems	Nein	Onlineshopping, -suche und browsing	Google, Microsoft

[3] https://setiathome.berkeley.edu/
[4] http://folding.stanford.edu/

2.3 Volunteered Geographic Information

VGI stellt in jüngerer Zeit ein immer breiteres Forschungsfeld in der Geografie dar. Geprägt wurde der Begriff durch GOODCHILD (2007). GOODCHILD (2007) definiert hierbei VGI wie folgt:

> *„[…] the widespread engagement of large numbers of private citizens, often with little in the way of formal qualifications, in the creation of geographic information, a function that for centuries has been reserved to official agencies. They are largely untrained and their actions are almost always voluntary, and the results may or may not be accurate. But collectively, they represent a dramatic innovation that will certainly have profound impacts on geographic information systems (GIS) and more generally on the discipline of geography and ist relationship to the general public. I term this volunteered geographic information (VGI), a special case of the more general Web phenomenon of usergenerated content […].“*

Er bezeichnet damit räumlich verortbare Information, die auf verschiedensten Wegen erhoben und bereitgestellt wird. Im Kern steht dabei jedoch immer der Nutzer, der diese Daten entweder freiwillig verfügbar macht oder diese ohne sein explizites Einverständnis durch diverse Dienste (bspw. Mobilfunkdaten) erhoben werden (RESCH 2017). Ursprünglich hat sich VGI aus Projekten wie OSM, Wikimapia und Google Earth entwickelt. Erst in jüngerer Zeit sind auch die Sozialen Netzwerke dazu gekommen. Allen diesen Plattformen ist gemein, dass sie frei zugänglich sind und die Inhalte erst durch freiwillige Nutzer hinzugefügt bzw. erweitert und ergänzt werden können. Dies kann unter dem Begriff UGC zusammengefasst werden. UGC wird dabei durch die OECD (2007) als öffentlich zugänglicher, selbst erstellter, kreativer Inhalt bezeichnet, der außerhalb professioneller Praktiken und Routinen kreiert worden ist. Eingebettet in den UGC sind schließlich auch die VGI.

Basis für diese Entwicklung stellen die Technologien des Web 2.0 dar. Erst mit den technologischen Möglichkeiten, die sich in den frühen 2000er Jahren herausgebildet haben, ist VGI ermöglicht worden. So wurden Protokolle entwickelt, die den Nutzer Erweiterungen von Server-Datenbanken durchführen ließen. Die immer selbstverständlichere Verfügbarkeit von Geo-Informationssystemen (GIS) wie QGIS, ArcGIS o. ä. vereinfachte zudem die geografische Verortung von Informationen immer weiter. So ist es nicht nötig, die exakten Koordinaten eines Objektes zu kennen, solange es sich bildhaft auf einer Karte verorten lässt. Dazu kommt die ubiquitäre Verfügbarkeit von Globalen Satellitennavigationssystemen wie dem US-amerikanischen Global Positioning System (GPS) oder dem europäischen Galileo, wodurch sich Bilder, Nachrichten o. ä. einfach verorten lassen. Damit letztlich die Information hoch- und runtergeladen werden können, ist die Breitbandanbindung unerlässlich (GOODCHILD 2007).

Wesentlich für VGI sind die Konzepte, die zu deren Entstehung beigetragen haben. Dazu zählt v. a. das Konzept der Geodateninfrastrukturen (auch Spatial Data Infrastructure - SDI), das Human as Sensors Konzept sowie Citizen Science. Als Ursprung der VGI bezeichnet GOODCHILD (2007) im Wesentlichen die 1994 entworfene National Spatial Data Infrastructure (NSDI), welche mit der Einführung von entsprechenden Standards den Datenaustausch, die Metainformationen sowie die Darstellung standardisiert haben.

Da die Datenmengen in jüngerer Zeit immer größer werden, spielt die Big Data Analyse eine gewichtige Rolle. Daher wird auch von der Data-driven Geography gesprochen, weil immer mehr Sensoren geografische Auswertungen, räumliche Analysen und Verortungen ermöglichen. Dies bedeutet, dass dank Sozialer Netzwerke inzwischen Bezüge zu ganzen

Bevölkerungsteilen hergestellt werden können und nicht mehr nur einzelne Stichproben analysiert werden (MILLER & GOODCHILD 2015). Damit einher gehen allerdings klassische Probleme. So sind Daten häufig unstrukturiert, besitzen Tipp- und Schreibfehler und können zusätzlich falsche Informationen enthalten. Zudem ist die Kategorisierung in diesem Bereich immer wieder ein vieldiskutiertes Thema, da so schnell Verallgemeinerungen und somit auch Vorverurteilungen stattfinden können (ZEDNER 2010). Zudem spielt der Datenschutz, insbesondere in Sozialen Netzwerken, eine wesentliche Rolle (VICENTE et al. 2011).

2.4 Citizen Science

Citizen Science ist ein englischer Begriff für Wissenschaftler und Forscher die nicht direkt als solche arbeiten. Handelt es sich hierbei um geografische Fragestellungen, können sie als ein Teil der VGI (vgl. Kap. 0) betrachtet werden (HAKLAY 2013). Populäre Beispiele dafür sind Benjamin Franklin oder auch Charles Darwin. Hauptberufliche Wissenschaftler sind eher ein Phänomen des 20. Jahrhunderts und waren vorher vielmehr die Ausnahme. Folglich sind Citizen Sciences nicht erst im Internetzeitalter geboren worden, vielmehr erleben sie heute eine Renaissance. Der Unterschied besteht darin, dass die wissenschaftliche Betätigung der breiten Masse und nicht nur wenigen Privilegierten zugänglich ist (SILVERTOWN 2009). COHN (2008) definiert Citizen Science wie folgt:

> *„The term "citizen scientists" refers to volunteers who participate as field assistants in scientific studies. Citizen scientists help monitor wild animals and plants or other environmental markers, but they are not paid for their assistance, nor are they necessarily even scientists. Most are amateurs who volunteer to assist ecological research because they love the outdoors or are concerned about environmental trends and problems and want to do something about them. Typically, volunteers do not analyze data or write scientific papers, but they are essential to gathering the information on which studies are based. Citizen scientists represent 'a partnership between volunteers and scientists to answer real-world questions'."*

In jüngerer Zeit gibt es eine große Zahl von Projekten, die sich die Intelligenz der Masse zu Nutze machen. Die Projekte sind dabei sowohl der Wissenschaft als auch dem Crowdsourcing zuzuordnen. Eine ganze Reihe an Projekten entstammt dem Bereich der Ökologie (Auffinden oder Zählen diverser Arten) oder auch der Astronomie (COHN 2008, BRABHAM 2013). Für eine bessere Verbindung zwischen Crowdsourcing-Programmen im Bereich Citizen Science und Freizeit-Wissenschaftlern gibt es Plattformen wie z. B. *InnoCentive*[5]. Folglich kommen in den Citizen Sciences sehr häufig die Methoden des Crowdsourcing zur Anwendung, obgleich Citizen Sciences auch offline durchgeführt werden können.

Die Methoden der Citizen Science empfehlen sich hierbei vor allem dann, wenn Gelder knapp sind und eine große Anzahl an Forschern benötigt wird, die Daten erheben. Ist die Methode klar umrissen, sind in der Regel auch die erhobenen Daten entsprechend belastbar (COHN 2008).

Angetrieben werden die Entwicklungen von denselben Faktoren wie die VGI, d. h. Internetstandards wie der eXtensible Markup Language (XML), der ubiquitären Verfügbarkeit von Kameras und Positionssensoren etc. Neben den technischen Treibern ist vor allem der Anstieg an gut ausgebildeten Personen sowie Personen, die die grundsätzlichen wissenschaftlichen Methoden beherrschen und Interesse an Forschung haben, zu nennen. Hinzu kommt ein Anstieg der Freizeit in den meisten Staaten der ersten Welt über die

[5] https://www.innocentive.com/

letzten Jahre. Daraus resultiert ein großer Pool an potentiellen, freiwilligen Wissenschaftlern (HAKLAY 2013).

Somit stellen die Citizen Sciences ein wachsendes Feld dar. Die Einbeziehung der einzelnen Bürger kann gerade im Bereich der Stadtplanung durchaus noch Potential entfalten; insbesondere wenn sich hier die Bürger selbst als Vorort-Experten einbringen mit ihrem persönlichen Bezug zu spezifischen Fragestellungen.

2.5 Open Data

Eng verbunden mit UGC, VGI, Citizen Sciences als auch mit Web 2.0 ist das Konzept der frei verfügbaren Daten, kurz Open Data. Häufig sind diese ohne Einschränkungen kostenfrei nutzbar. Teilweise existieren aber entsprechende Regelungen und Lizenzen, dass die Daten nicht kommerziell weiterverwendet werden dürfen. Zu den offenen Daten zählen zum Beispiel die *mCLOUD*[6] des Bundesministeriums für Verkehr und Digitale Infrastruktur (BMVI), Daten wie das bereits angesprochene OSM, aber eben auch Daten, die öffentlich in Sozialen Netzwerken geteilt und veröffentlicht werden. Des Weiteren werden durch die Datenportale der Länder als auch der GDI-DE freie Daten bereitgestellt. Zudem bieten diverse Unternehmen ebenfalls eingeschränkt freie Daten an, wie z. B. Google Maps (Abbildung 2-1) (BILL et al. 2018).

Zurückgeführt werden kann das Konzept von Open Data bereits auf die 50er Jahre des 19. Jahrhunderts. In der Debatte um den Freedom of Information Act fiel der Begriff dabei zum ersten Mal (YU & ROBINSON 2012). 1957 wurden im Rahmen des Internationalen Geophysikalischen Jahres Datenaustauschzentren geschaffen und der Versuch unternommen, Metadaten zu standardisieren (JIFFREY 2006). Das Konzept wurde in den 70er Jahren durch die National Aeronautics and Space Administration (NASA) weiter vorangetrieben. In den 90er Jahren kamen schließlich weitere, eng verbundene Konzepte wie Open Access und Open Government hinzu, welche ebenfalls darauf abzielen, Daten einer breiten Öffentlichkeit verfügbar zu machen (YU & ROBINSON 2012).

Damit stellt Open Data einen wichtigen Bestandteil sowohl für die kommerzielle Aufbereitung als auch für Forschung und Wissenschaft dar. Allerdings stehen der umfassenden Nutzung und intelligenten Vernetzung häufig sowohl organisatorische als auch technische Hemmnisse entgegen (BILL et al. 2018). Vor allem fehlt in der Regel eine systematische Erfassung der angebotenen Daten. D. h. es ist dem Nutzer häufig unbekannt, auf welche Daten er wo zugreifen kann. Um dem entgegenzuwirken, entstand im Rahmen des Projektes *OpenGeoEdu*[7] eine Plattform, auf der eine Vielzahl der in Deutschland frei zugänglichen Datenportale gelistet worden ist (HINZ & BILL 2018).

[6] https://www.mcloud.de/
[7] http://portal.opengeoedu.de/

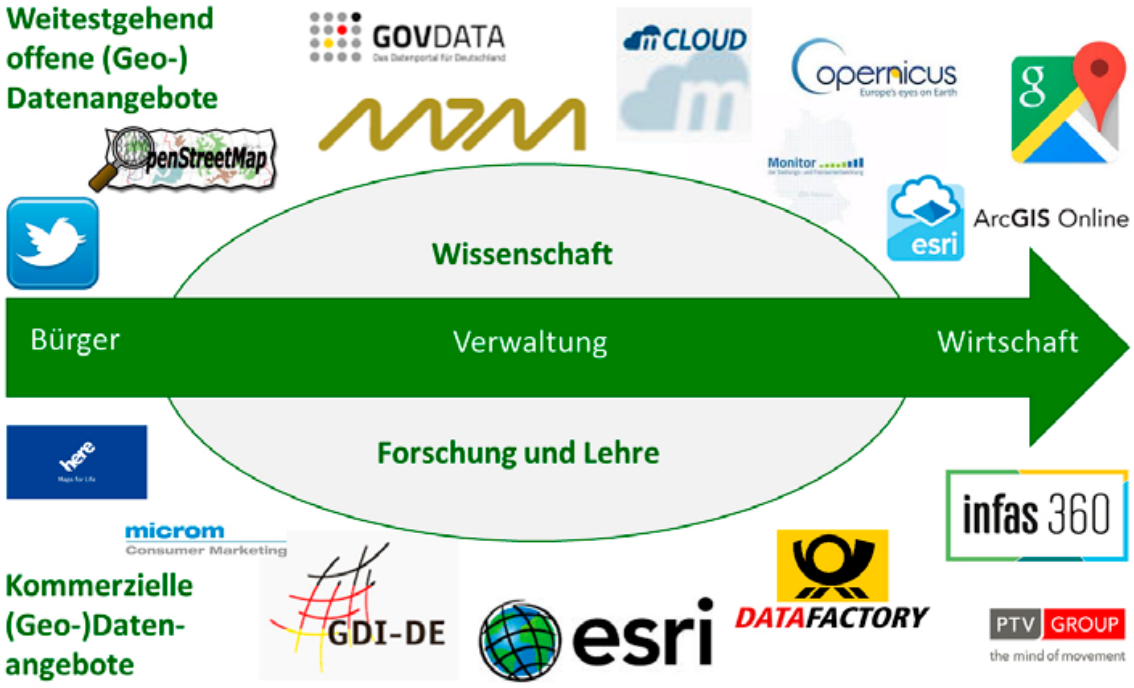


Abbildung 2-1: Offene und kommerzielle Angebote von Geodaten (BILL et al. 2018).

2.6 Moderne Verfahren der Bürgerbeteiligung

Als Bürgerbeteiligung wird in der Regel der organisierte Prozess für die Einbeziehung der Öffentlichkeit bezeichnet. Dies umfasst nach CREIGHTON (2005) in der Regel die folgenden vier Kategorien:

1. Die Öffentlichkeit informieren
2. Der Öffentlichkeit zuhören
3. Beiträge zur Problemlösung diskutieren
4. Übereinkommen entwickeln

Das heißt, Bürgerbeteiligung lässt sich mit den Schlagworten Information, Kommunikation und Partizipation beschreiben.

Häufig ist ein Diskussionsthema, ob eine Bürgerbeteiligung überhaupt sinnvoll ist. So bringt sie zusätzlich Risiken mit sich – diese reichen von Machtverlust über finanzielle Fragestellungen, über mangelndes Wissen der teilnehmenden Personen und über die Rahmenbedingungen bis hin zum Vorurteil, dass sich hier immer nur die gleichen Personen engagieren, die von vornherein gegen das jeweilige Projekt sind (BIRZER 2015). Es ist davon auszugehen, dass sich diese Probleme auch auf VGI basierenden Daten wiederfinden.

In der Stadtplanung und zum Einbeziehen der Bevölkerung spielt Crowdsourcing sowie Crowdmining bisher in Deutschland eine untergeordnete Rolle, obgleich diese Methoden ein großes Potential besitzen und es viele Projekte diesbezüglich gibt. Dies zeigen diverse Beispiele wie der *Leerstandsmelder*[8], *Klarschiff.HRO*[9], *Streetbump*[10] u. v. m. Vor allem

[8] https://www.leerstandsmelder.de
[9] https://www.klarschiff-hro.de/
[10] http://www.streetbump.org

international gibt es eine Vielzahl an Ansätzen. Als Beispiel sei hier das 2009 ins Leben gerufene Projekt *Next Stop Design*[11], bei dem die Crowd Entwürfe für eine Busstation sowie für eine Kreuzung in Salt Lake City erstellen sollte, genannt. Das Projekt ist dabei auf eine enorme Resonanz gestoßen und konnte zeigen, dass die Stadtplanung diese Methode durchaus in Erwägung ziehen sollte (BRABHAM 2012). Eine jüngere Idee, die mittels des Spiels Minecraft versucht, die Öffentlichkeit in die Stadtplanung einzubeziehen, ist *Block By Block*[12]. Seit 2012 können die Teilnehmer hier spielerisch öffentliche Plätze und Anlagen gestalten. Bisher kam die Idee in über 30 Orten weltweit zur Anwendung (BLOCK BY BLOCK 2018). Häufig kommen aber auch die Sensoren von Smartphones zur Anwendung, etwa zur Berechnung des verbrauchten CO_2 oder zur Identifikation von Verkehrsknotenpunkten (SHIN 2016).

Noch jünger sind erste Methoden, die sich auf die Sozialen Netze stützen. Allerdings steht hier i. d. R. nur die Analyse oder aber die Bevölkerungsinformation über Kanäle in Sozialen Netzwerken, nicht aber die aktive Beteiligung der Bevölkerung im Vordergrund. Vor allem Bevölkerungsbewegungen aber auch detaillierte, sozioökonomische Daten lassen sich gut über Twitter abbilden bzw. erheben (SHELTON et al. 2015, LONGLEY & ADNAN 2015). Werden sie darüber hinaus als Diskussionsplattformen genutzt, findet dies häufig in der Form von Gruppen via Facebook statt (CIVITAS 2015). Daneben finden häufig Mikroblogs wie Twitter vor allem als Element im Notfallmanagement Anwendung, da hier die Echtzeitkommunikation von entscheidender Bedeutung ist (TERPSTRA et al. 2012). Interessant ist bei Nachrichten in Sozialen Medien nicht nur der Fokus auf die räumliche Verortung. Vielmehr sind vor allem Inhalte und Metainformationen von großer Bedeutung (ZOOK 2017, CRAMPTON et al. 2013).

Dies zeigt, dass UGC und VGI eine Möglichkeit darstellen, gerade die jüngere, in sozialen Medien engagierte Generation in urbane Planungsprozesse einzubeziehen (LI 2016). Besonders hebt LI (2016) dabei hervor, dass Soziale Medien es ermöglichen, eine unbegrenzte Anzahl an Bürgern teilhaben zu lassen, da die Diskussionen und Beiträge weder zeitlich noch räumlich begrenzt sind. Dabei ist auch von Feedbackeffekten auszugehen. D. h., wenn die Leute wissen, dass unter bestimmten Kategorien in den Sozialen Netzwerken ihre Probleme und Sorgen hinsichtlich der Stadtplanung erfasst werden, sind sie u. U. mehr dazu motiviert, zu diesen Themen Nachrichten zu verfassen. Solche Effekte sind bereits durch BOND et al. (2012) nachgewiesen worden, wobei hier vor allem die politische Beeinflussung über Facebook im Vordergrund stand.

Soziale Medien sind dabei jedoch nicht in der Lage, die herkömmlichen Verfahren wie Begehungen, Bürgerräte o. ä. zu ersetzen, die ebenfalls auf eine direkte Einbeziehung der Bürger abzielen (NANZ & FRITSCHE 2012). Nutzt man soziale Netzwerke zur Bestimmung von Trends und Meinungen, wird schnell der Punkt erreicht, an dem deutlich wird, dass die Nutzer Sozialer Netzwerke nicht der Grundgesamtheit der Bevölkerung entsprechen. Dadurch werden bestimmte Bevölkerungsgruppen ausgeschlossen oder über- bzw. unterschätzt. Daher lassen sich die Daten nicht generalisieren (MILLER & GOODCHILD 2015). Eine häufige Kritik an Diskursen in Sozialen Medien ist zudem, dass hier häufig Diskussionen schnell in einem „*Flame War*“, d. h. üblen Beschimpfungen und damit in Unsachlichkeit enden. Außerdem besteht die Gefahr, dass die Themen durch Trolle oder andere Personenkreise gekapert werden (KREIL 2017). Somit bieten sie vielmehr in der Ergänzung einen entscheidenden Mehrwert im Bereich der Bürgerbeteiligung und stellen somit eine gute Erweiterung zu den herkömmlichen Verfahren dar.

[11] http://nextstopdesign.com/
[12] https://blockbyblock.org/

2.7 Soziale Netzwerke

Die Entwicklung des Web 2.0 und die Verbreitung von durch Nutzer generierten Inhalten stellen die Basis für die beispiellose Expansion der Sozialen Netzwerke im World Wide Web (WWW) dar. Soziale Netzwerke selbst stellen dabei ein relativ junges Phänomen dar, wobei das Konzept selbst eigentlich schon recht alt ist und aus der Anfangszeit des Internets stammt. So wurde der Begriff Virtual Community bereits durch LICKLIDER & TAYLOR (1968) erstmals verwendet. Auch Mailinglisten und Newsgroups bildeten eine frühe Basis der Sozialen Netzwerke. Die Kommerzialisierung und die Etablierung der Sozialen Netzwerke, wie wir sie heute kennen, fand jedoch erst Ende der 1990iger Jahre statt (HEIDEMANN 2010). Seither hat sich eine Vielzahl von Netzwerken entwickelt, die alle versuchen verschiedene Interessen und Bedürfnisse zu bedienen (Abbildung 2-2).

BOYD & ELLISON (2007) definieren Online Social Networks dabei nach drei grundsätzlichen Kriterien:

1. Erstellung eines öffentlichen oder halböffentlichen Profils innerhalb eines begrenzten Systems
2. Darstellung einer Liste von Nutzern, mit denen die Person in Beziehung steht
3. Betrachten und Durchsuchen der Beziehungsliste durch andere Nutzer des Netzwerkes

Es stehen also der Aufbau und das Pflegen der Verbindungen zwischen den Akteuren im Vordergrund, deren Kommunikation und Interaktion durch eine Online Software unterstützt wird.

Um die verschiedenen Sozialen Netzwerke einordnen zu können haben sich diverse Klassifikationssysteme etabliert. Sie hängen dabei stark von der jeweiligen Fragestellung ab. So spielt es eine entscheidende Rolle, ob die Kategorien sich beispielsweise an den Wegen des Marketings, die ein Netzwerk bietet, an der Funktionsweise des Netzwerkes als solches oder der Art der über das Netzwerk zur Verfügung gestellten und verbreiteten Information orientieren. Aus dem Marketing-Bereich heraus empfiehlt sich die in Tabelle 2-3 ersichtliche Unterteilung.

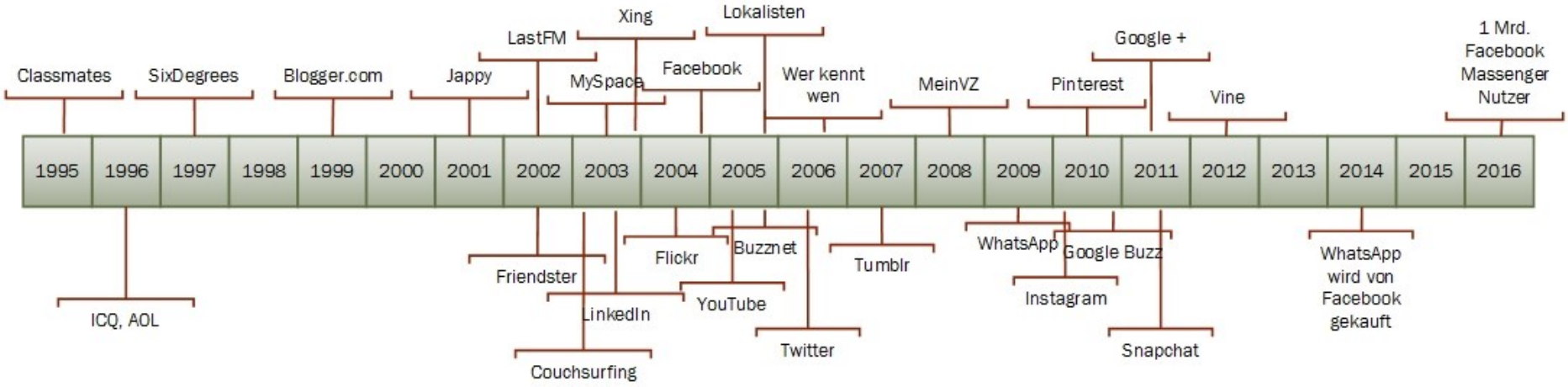


Abbildung 2-2: Entwicklung Sozialer Netzwerke (eigene Erstellung, nach HEIDEMANN 2010, STEINBRENNER 2017).

Tabelle 2-3: Klassifizierung Sozialer Netzwerke (eigene Bearbeitung, nach GUNDECHA & LIU 2012, ZARELLA 2010).

Typ	Beschreibung	Beispiele
Online Social Networking	Web basierte Services, die einzelne Individuen miteinander verbinden und deren Kommunikation ermöglichen.	Facebook, LinkedIn, ...
Blogging	Website-ähnliches Teilen von Texten, Bildern und anderen Informationen und Anzeige nach zeitlicher Abfolge (neueste zuerst).	Huffington Post, Stefans Börsenblog, ...
Mikroblogs	Ähnlich wie Blogs, wobei die Länge der Posts hier stark eingeschränkt ist.	Tumblr, Twitter, Instagram, …
Wikis	Wikis stellen eine Umgebung dar, in der sich Webseiten kollaborativ erstellen lassen.	Wikipedia, Wikihow, ...
Social News	Teilen und Verlinken von aktuellen Neuigkeiten durch die Webgemeinschaft.	Reddit, Slashdot, Seeking Alpha, ...
Social Bookmarking	Lesezeichen, die von mehreren Nutzern gleichzeitig verwendet werden können.	Delicious, StumbleUpon, ...
Media Sharing	Teilen von Medien aller Art, wobei Videos und Bilder im Vordergrund stehen.	Youtube, Flickr, Twitch, ...
Meinungen, Rezensionen und Bewertungen	Die Seiten dienen hauptsächlich dazu, Bewertungen von verschiedenen Medien, Lokalitäten oder Produkten zu veröffentlichen.	Rotten Tomatoes, Yelp, Foursquare, ...
Hilfeseiten	Plattformen für Nutzer, die Hilfe, Rat oder Anleitungen suchen bzw. teilen. Die Antworten werden i. d. R. entsprechend bewertet.	WikiAnswers, Stackoverflow, ...
Virtuelle Welten	Soziale Netzwerke, die eine zweite Identität schaffen und in denen man mit anderen Identitäten agiert.	World of Warcraft, Second Life, ...
Foren	Seiten, in denen sich verschiedene Nutzer über diverse Themen austauschen können.	Wertpapier-Forum, Meistertrainerforum, ...

2.7.1 Twitter – der Mikroblogging-Dienst

Twitter ist ein Mikroblogging-Dienst, welcher im Jahr 2006 von Jack Dorsey gegründet wurde. Twitter unterscheidet sich dabei wesentlich von den Sozialen Netzwerken Google+ oder Facebook (TWITTER INC. 2015). So lautet das Motto des Unternehmens: „*To give everyone the power to create and share ideas and information instantly, without barriers* (TWITTER INC. 2016a)."

Dies beinhaltet, dass die derzeit 320 Mio. MAUs auf maximal 140 Zeichen (seit November 2017 280 Zeichen) begrenzte Nachrichten oder einzelne Bilder posten sowie anderen Nutzern folgen können, um deren Posts mitzulesen. Daneben besteht auch die Möglichkeit, private Nachrichten direkt an einen spezifischen Nutzer zu versenden. Twitter ist insbesondere aus dem Grund interessant, da das Gros der Nutzer (ca. 80 %) vom Smartphone auf die App zugreift (TWITTER INC. 2015). Damit verbunden ist ein gewisser Anteil an geolokalisierten Nachrichten (in Deutschland ca. 1 %) (FUCHS et al. 2013). So lässt sich in

jeder Nachricht die GNSS-Koordinate des Smartphones hinterlegen, wodurch Twitter-Nachrichten vor allem für räumliche Analysen interessant sind.

Allerdings ist festzuhalten, dass Twitter aktuell nur einen geringen Nutzerzuwachs verzeichnet. Nichtsdestotrotz wurden 2013 ca. 500 Mio. Nachrichten pro Tag versendet (TWITTER INC. 2015). Diese Zahl wird seither jedoch nicht mehr aktualisiert, daher ist davon auszugehen, dass das tägliche Nachrichtenaufkommen etwas unter dieser Anzahl liegt (ORESKOVIC 2015).

Neben Twitter bieten auch andere Soziale Netzwerke wie Facebook Mikroblogging-Funktionen. Diese sind jedoch nicht vordergründig dem Geschäftsmodell der Unternehmen zuzurechnen. Damit besitzt Twitter ein klares Alleinstellungsmerkmal. Zudem lassen sich die Tweets durch Wissenschaft und Forschung sehr gut auswerten, da diese über deren API (Application Programming Interface) gut zugänglich sind (vgl. Kap. 4.2.1 und Kap. 4.2.2). Dies zeigen über 737 sozialwissenschaftliche Forschungsarbeiten, die zwischen 2007 und 2015 auf Scopus gelistet werden (PFAFFENBERGER 2016). Nicht zuletzt hat der Dienst durch zahlreiche Tweets des 45. US-Präsidenten Donald Trump einen neuerlichen Bekanntheitsschub erhalten (JAUERNIG 2017, KOLB 2016).

2.7.2 Foto-Sharing

2.7.2.1 Instagram

Instagram ist ein im Jahre 2010 gegründetes Unternehmen, welches seit 2012 Teil von Facebook ist. Dabei ist die Plattform weiterhin unabhängig von Facebook, obgleich sich Verknüpfungen herstellen lassen. Instagram ist dabei, ähnlich wie Twitter, ein Mikroblogging-Dienst, welcher sich jedoch auf Fotos spezialisiert hat. Instagram hat dabei im Juni 2018 1 Mrd. MAU erreicht und es werden über 100 Mio. Bilder pro Tag geteilt (ASLAM 2018).

Instagram ermöglicht es, ein Foto direkt via Smartphone-App auszuwählen, zu bearbeiten und abschließend zu teilen. Der Nutzer kann dabei verschiedenen anderen Nutzern folgen und sich deren Fotos betrachten. Instagram bietet dabei ebenfalls die Möglichkeit, Geoinformationen zu teilen. So kann entweder direkt ein GNSS getaggtes Foto hochgeladen werden, oder aber manuell der aktuelle Standort bzw. der Standort des Smartphones in der Nachricht angegeben werden. Daneben besitzen die Bilder in der Regel eine Beschreibung. In dieser können ebenfalls Lokalitäten vorkommen (INSTAGRAM 2016).

2.7.2.2 Flickr

Flickr ist, ähnlich wie Instagram, ebenfalls eine Plattform, bei der der Schwerpunkt auf Fotos gelegt ist. Allerdings ist Flickr, anders als Instagram, insbesondere an professionelle Fotografen gerichtet. Das Unternehmen wurde 2004 gegründet und ist seit 2005 Teil von Yahoo (DELANEY 2005, FAKE 2008).

Flickr bietet verschiedene Möglichkeiten, Zusatzinformationen wie Beleuchtungsdauer, verwendetes Objektiv etc. anzugeben. Dies mündet darin, dass Flickr eine deutlich geringere Nutzerbasis aufweist, insgesamt jedoch von einer höheren Qualität der Fotos als solches auszugehen ist. Insgesamt bietet Flickr Zugriff auf über 6 Mrd. Fotos, 2017 sind etwa 1.6 Millionen pro Tag dazu gekommen (MICHEL 2018).

Der Aspekt der Vernetzung ist auch hier durch die Möglichkeit gegeben, bestimmten Nutzern zu folgen oder Bilder zu bewerten und zu kommentieren. Außerdem will Flickr eine Organisationsplattform für Bilder bieten und hat somit Gemeinsamkeiten mit Diensten wie beispielsweise Picasa (FLICKR 2016).

2.7.3 Soziale Interaktion und Selbstdarstellung

2.7.3.1 Facebook

Facebook stellt wohl das bekannteste aller sozialen Netzwerke dar. Das im Jahr 2004 unter anderem von Marc Zuckerberg gegründete Netzwerk besitzt etwa 1.5 Mrd. Daily Active User (DAU) (FACEBOOK INC. 2018). Dabei ist das Netzwerk heute aus dem Alltag kaum noch wegzudenken. So sinken die Umsätze der Mobilfunkunternehmen mit SMS (Short Messaging Service) bedingt durch das Facebook zugehörige Unternehmen WhatsApp deutlich, und mit der immer weiteren Verbreitung von Smartphones ist dieser Trend zukünftig kaum aufzuhalten (LICHTENBERG 2016). Auch Veranstaltungsinformationen werden über Facebook geteilt und die neuesten Urlaubsbilder auf der Pinnwand der Freunde angesehen, wodurch die Kommunikation sich immer stärker auf Facebook konzentriert hat. Seit 2010 ist es zudem möglich, über Facebook Places Geoinformationen zu teilen. So lässt sich beispielsweise identifizieren, welcher Freund sich wann wo befindet oder welche Geschäfte in näherer Umgebung gerade geöffnet haben (FACEBOOK INC. 2016).

2.7.3.2 Google+

Da Google erkannt hat, dass mit sozialen Netzwerken und den dort von den Nutzern bereitgestellten Informationen viel Geld zu verdienen ist, entschloss sich der Internetkonzern 2011 dazu, ebenfalls ein Soziales Netzwerk an den Start zu bringen: Google+ (MILLER 2011). Es ist allerdings festzuhalten, dass Google+ ein besonderes Konzept aufweist. So dient Google+ dazu, verschiedene Teile des Technologiekonzerns miteinander zu verknüpfen. So sind über Google+ bereits vorhandene Dienste, wie Gmail oder Youtube, integriert. Allerdings wird dies aktuell nicht weitergeführt, da Google+ zwar einmal das am schnellsten wachsende soziale Netzwerk war, jedoch nie an die Popularität von Facebook heran gereicht hat (FAZ 2014). Vielmehr soll Google+ als ein internes Rückgrat fortgeführt werden (FAZ 2014, 2014, TSOTSIS & PANZARINO 2014).

2.7.4 Weitere Soziale Netzwerke

Neben den genannten Netzwerken gibt es zahllose weitere, welche jedoch im Rahmen dieser Arbeit keine weitere Erwähnung finden. Darunter fallen Foursquare, Youtube, SnapChat und viele andere mehr. Allerdings gibt es auch zahlreiche Soziale Netzwerke, welche heute nicht mehr existieren oder kaum noch verwendet werden. Dazu zählt zum Beispiel MySpace oder das hierzulande ehemals sehr beliebte StudiVZ (DITTMANN 2016). Jedoch steht und fällt ein Netzwerk mit der Anzahl seiner Benutzer. Wenn das Gros des Bekanntenkreises bei Facebook ist, wird man am ehesten ebenfalls dorthin wechseln, um am sozialen Leben teilhaben zu können bzw. den Kontakt zu alten Freunden beizubehalten (ELLISON et al. 2007). Daher scheitern häufig nationale Netzwerke, internationale haben dagegen eine deutlich größere Erfolgsquote (BERNET 2010). Man denke hierbei an die zahlreichen deutschsprachen Videoportale, die letztlich neben Youtube, was einen Marktanteil von knapp 81 % aufweist (Stand 2016), kaum noch eine Rolle spielen (STATISTA 2016).

Jedes Netzwerk besitzt ein spezifisches Ziel. Häufig gibt es, aufgrund der Nutzerkonzentration, nur eine geringe Konkurrenz in einem Bereich. So teilen sich beispielsweise LinkedIn und Xing den Markt der Jobsuche im Internet in Deutschland oder Facebook ist weltweit als führendes Netzwerk zum Pflegen von Freundschaften bekannt (BERNET 2010).

2.8 Geodateninfrastrukturen

Definiert werden kann eine Geodateninfrastruktur als „[…] *eine aus technischen, organisatorischen und rechtlichen Regelungen bestehende Bündelung von Geoinformationsressourcen, in der Anbieter von Geodaten und Geodiensten mit Nachfragern solcher Dienste kooperieren.*“ (BILL 2016). Dabei werden raumbezogenen Daten mit fachlichen Thematiken kombiniert, wobei der Anwender hier eigene Daten hinzufügen und seinen Datenbestand mit dem der Infrastruktur synchronisieren kann. Wichtig sind dabei die verschiedenen Bestandteile (Abbildung 2-3). Dazu zählen:

- die Geodatenbasis
- die Metadaten
- das Geoinformationsnetzwerk
- Dienste
- Standards
- politische Rahmenbedingungen
- und interorganisatorische Vereinbarungen

Somit schafft die GDI die Basis zur Wertschöpfung im kommerziellen und nicht kommerziellen Bereich, auf der sich neue Services, Workflows sowie Produktionsketten entwickeln und etablieren lassen. Ein wesentliches Charakteristikum ist, dass Informationsanbieter und -nutzer nicht mehr direkt miteinander in Verbindung treten, sondern über entsprechende Identifikations- und Aufbereitungsservices interagieren (BILL 2016). Eingerahmt sind die GDIs in zahlreiche Gesetze und Richtlinien. Für eine detaillierte Beschreibung sei an dieser Stelle auf THIEL (2015) verwiesen.

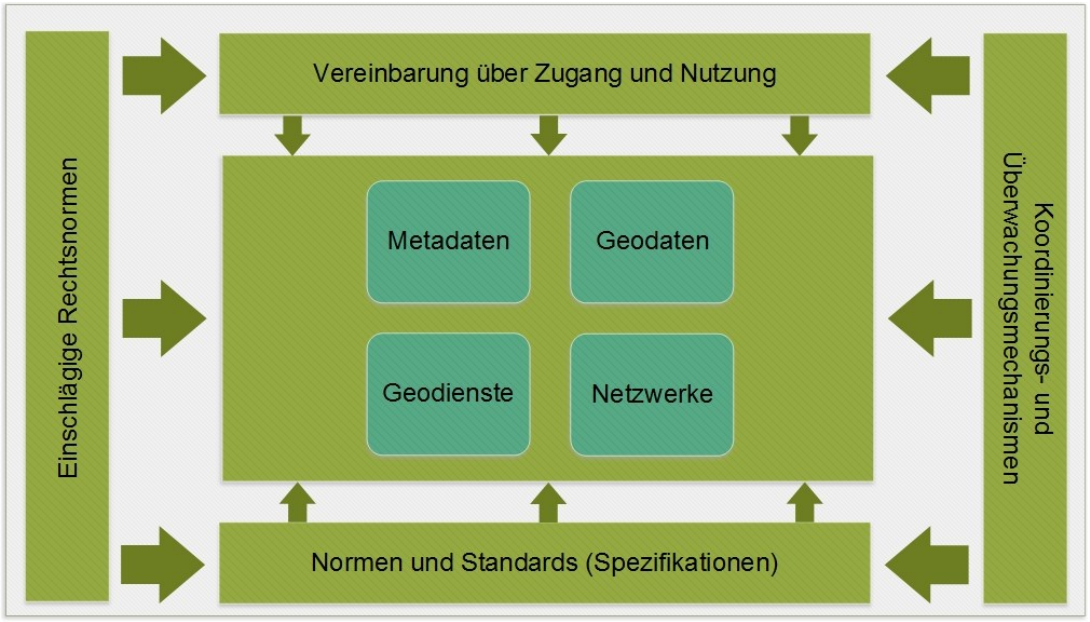


Abbildung 2-3: Komponenten und Rahmenbedingungen einer GDI (eigene Erstellung, nach BILL 2016).

Neben der hauptsächlichen Verwendung als Verwaltungsinstrument sind Geodateninfrastrukturen aus Projekten und Arbeiten im Fachbereich der Geografie, Landschaftsplanung, Umweltwissenschaften aber auch des Ingenieurbereichs kaum noch wegzudenken.

Sobald Geodaten vorhanden sind, stellt sich immer die Frage nach Standards zu Datenaustausch, -zugriff und -prozessierung. Aus diesem Grund beinhalten zahlreiche Forschungsprojekte als ein wesentliches Kernelement den Entwurf und die Entwicklung von Geodateninfrastrukturen (KOLDRACK et al. 2017).

Der Beginn der GDIs wurde durch die Aufforderung Al Gores zum Aufbau eines weltweiten, hochaufgelösten Datenbestandes und den Start der NSDI 1994 in den USA initiiert. Seither sind zahlreiche nationale und internationale Lösungen entworfen worden (BILL 2016). Dazu zählt sowohl die deutsche Geodateninfrastruktur (*GDI-DE*[13]) als auch *INSPIRE*[14] auf europäischer Ebene (BKG 2016, EUROPEAN COMISSION 2017). Zudem gibt es zahlreiche kommunale und regionale Lösungen. Hier wären beispielsweise *OpenData.HRO*[15] oder die Infrastrukturen der deutschen Bundesländer (z. B. *GeoPortal.MV*[16]) zu nennen (HANSESTADT ROSTOCK 2015a, BKG 2016). Letztlich sind also die einzelnen Infrastrukturen immer in eine übergeordnete Ebene eingebettet, wobei die höchste Ebene in Europa INSPIRE darstellt (Abbildung 2-4). Aus diesem Grund ist eine INSPIRE-Konformität der jeweiligen Infrastrukturen ein wesentliches Kriterium (GDI-DE 2015).

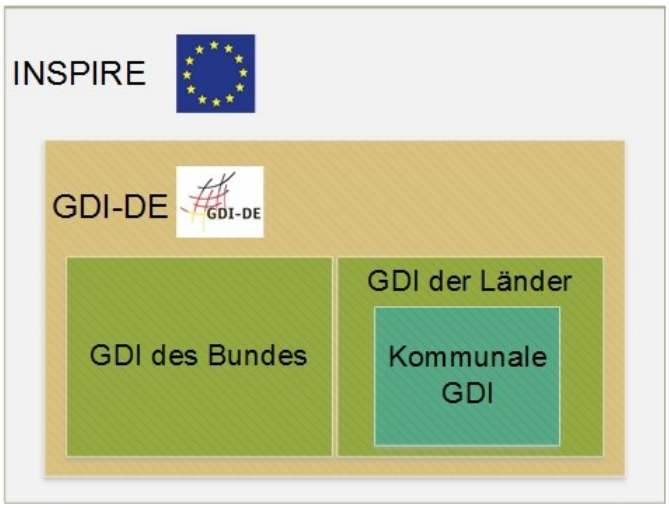


Abbildung 2-4: Hierarchie der GDIs in Europa (eigene Erstellung, nach ARBEITSKREIS ARCHITEKTUR 2013).

2.8.1 Normen

Die Normen und Standards werden durch die International Standard Organization (ISO) bzw. durch das Comité Européen de Normalisation (CEN) und nationale Normungseinrichtungen wie die Deutsche Industrie Norm (DIN) definiert. Für Geodaten bzw. raumbezogene Daten ist das Technical Committee (TC) 211 „Geoinformatik/Geomatik“ sowie bezüglich Fahrzeugnavigation TC 204 „Intelligent Transport Systems“ zuständig. Diese TCs haben eine ganze Reihe an Normen publiziert[17] - der bekannteste im Geoinformatikbereich dürfte der Metadatenstandard ISO 19115 sein (KRESSE & FADAIE 2004).

Auf europäischer Ebene werden durch das CEN, welches die Normungsorganisationen der EU und EFTA vereint, Normierungen und Standards verabschiedet. Es orientiert sich

[13] http://www.geoportal.de/DE/GDI-DE/gdi-de.html?lang=de
[14] http://inspire.ec.europa.eu/
[15] https://www.opendata-hro.de/
[16] http://www.geoportal-mv.de/portal/
[17] www.isotc211.org

dabei ebenfalls an der ISO. Mit Bezug zu Geodaten sind insbesondere die Normen des TC 278 „*Road Traffic and Transport Telematics*“ sowie das 1996 durch die Arbeitsgruppe 7 verabschiedete Geographic Data File Format (GDF) zu nennen (BILL 2016).

In Deutschland bzw. im deutschsprachigen Raum ist vor allem der Ausschuss Bauwesen im Fachbereich 03 „*Vermessungswesen, Geoinformation*“ der Deutschen Industrienorm (DIN) mit der Festlegung von Normen beschäftigt. Dabei werden in der Regel die Normen der ISO durch die DIN übernommen und hierzulande verabschiedet. Im deutschsprachigen Bereich gibt es neben der DIN zusätzlich den österreichischen Normenausschuss (ON) sowie die Schweizerische Normen-Vereinigung (SNV) (BILL 2016).

2.8.2 OGC-Spezifikationen

Das Open Geospatial Consortium (OGC) ist eine internationale non-profit-Organisation, welche für Geodaten und Geoservices Spezifikationen erstellt und festlegt. Die Organisation wurde 1994 gegründet und umfasst inzwischen etwa 500 Mitglieder (GIS-Hersteller, Dienstanbieter, Hochschulen, …). Wesentliches Ziel ist es, Interoperabilität herzustellen. Dies bedeutet, „[…] *die Definition einer Technologie, welche es einem Anwendungsentwickler und Anwender ermöglicht, jede Art von geocodierten Daten und Geo-Funktionalität oder -Prozess zu nutzen, welche im Netz verfügbar ist, innerhalb seiner Umgebung und seines jeweiligen individuellen und einzelnen Arbeitsablaufs.*“ (BILL 2016).

Tabelle 2-4: Auswahl von OGC Web Services (eigene Erstellung, nach BILL 2016).

Bezeichnung	Abkürzung	Beschreibung
Web Map Service	WMS	Visualisierung eines Rasterbildes (jpg, png, …) aus vorhandenen Daten
Web Feature Service	WFS	Zugriff auf Objekte in Vektorform
Transactional WFS	WFST	Schreibender Zugriff möglich
Web Processing Service	WPS	Datenprozessierungswerkzeuge
Web Coverage Service	WCS	Zugriff auf Objekte in Rasterform
Catalogue Service Web	CSW	Katalogdienste nach ISO 19119
Coordinate Transformation Service	CTS	Koordinatenumrechnung
Web Terrain Service	WTS	Geländemodelle
Web 3D Service	W3DS	3D-Darstellung von Daten
Grid Coverage Service Implementation Specification		Methoden zur Interoperabilität von Rasteranalysen
Location Service Implementtation Specification	OpenLS	Offene Plattform für Location Based Services
Web Authentication Services	WAS	Nutzerverwaltung und Authentifizierung
Web Pricing and Ordering Service	WPOS	Bepreisung von Geodaten und -diensten

In den Spezifikationen werden die Schnittstellen und Codierungsregeln für Softwareentwickler beschrieben. Zum problemlosen Zusammenspiel definieren die Simple Features Implementation Specifications 1 und 2, die Common Object Broker Request Architecture (COBRA) und Object Linking and Embedding (OLE) bzw. das Component Object Model (COM) die Schnittstellen für den Geodatenzugriff in heterogenen Systemen. Des Weiteren setzt das OGC auf das Open Geodata Model. Hierbei sollen die diversen Geodatenmo-

delle der verschiedenen GIS-Anbieter harmonisiert werden. Dafür müssen die entsprechenden Schnittstellen bereitgestellt werden (BILL 2016). Alle OGC-Spezifikationen können online nachgeschlagen werden[18].

Seit den 2000er Jahren liegt der Fokus vor allem auf internetbasierten Lösungen (Tabelle 2-4). In der Regel besitzt jeder Dienst die Abfrage nach seinem Leistungsumfang (GetCapabilities). Darüber hinaus lassen sich über diverse Get-Abfragen weitere Informationen vom Dienst extrahieren. So dient die GetMap-Abfrage beim Web Map Service (WMS) dazu, die Karte als Bild zurück zu liefern. Beim Web Feature Service (WFS) dient die GetFeature-Abfrage dazu, die Daten im Geography Markup Language (GML) Format zurück zu geben. Daneben bietet jeder Dienst weitere Anfragen auf deren Auflistung hier jedoch verzichtet werden soll (BILL 2016).

2.9 Location-based Social Networks

Soziale Netzwerke bedienen den Bedarf, dass Menschen wissen wollen, was andere Nutzer gerade machen und wo diese sich befinden. Ersteres wurde bereits durch die herkömmlichen Sozialen Netzwerke geliefert, letzteres bieten erst die räumlichen Sozialen Netzwerke – auch LBSNs. So wird schließlich die Information, die eine Person teilt, deutlich interessanter für andere Personen (TRAYNOR & CURRAN 2013). Möglich wurde dies durch die ubiquitäre Verwendung von Smartphones mit GNSS-Empfängern und Internetzugang (GOODCHILD 2007). Das Teilen einer Lokation wird hierbei als check in bezeichnet und kann dann von einer bestimmten Auswahl an Personen (z. B. Freunde) oder auch öffentlich betrachtet werden. Daraus folgen nach TRAYNOR & CURRAN 2013 verschiedene Bedürfnisse und Vorteile, die durch die LBSNs gedeckt werden.

1. Es lassen sich Ortsinformationen eines Check-Ins abrufen und man ist so z. B. vor einer teuren Hotelbar vorgewarnt.
2. Es besteht die Möglichkeit zu prüfen, wer gerade in der Nähe ist, um sich kurzfristig mit der Person zu treffen. Auch Gruppentreffen lassen sich so sehr gut organisieren, da lediglich ein Punkt auf einer Karte im Smartphone angesteuert werden muss.
3. Es lassen sich neue Orte entdecken, von denen man vorher nichts wusste. Dies ist insbesondere im touristischen Bereich interessant.
4. Oft bieten LBSNs auch einen spielerischen Aspekt. So lassen sich häufig Punkte oder Rabatte verdienen, wenn man zu bestimmten Zeiten und Orten eincheckt. Durch eine Rangliste wird hier der Anreiz verstärkt, ähnlich wie beim Geocaching.
5. Mit Hilfe von LBSNs lässt sich eine Historie von check ins erstellen. So kann man in einem virtuellen Tagebuch z. B. die Orte einer Reise nachverfolgen.

Die Funktionsweise der LBSNs ist in Abbildung 2-5 dargestellt. Der geografische Layer beinhaltet hierbei den check in Verlauf, der soziale Layer die Beziehungen und Freunde und der Content Layer schließlich die Angaben und Hinweise des Nutzers über verschiedene Orte. Alle Ereignisse beziehen sich dabei letztlich auf eine Zeitschiene (MORSTATTER et al. 2015).

[18] http://www.opengeospatial.org/docs/is

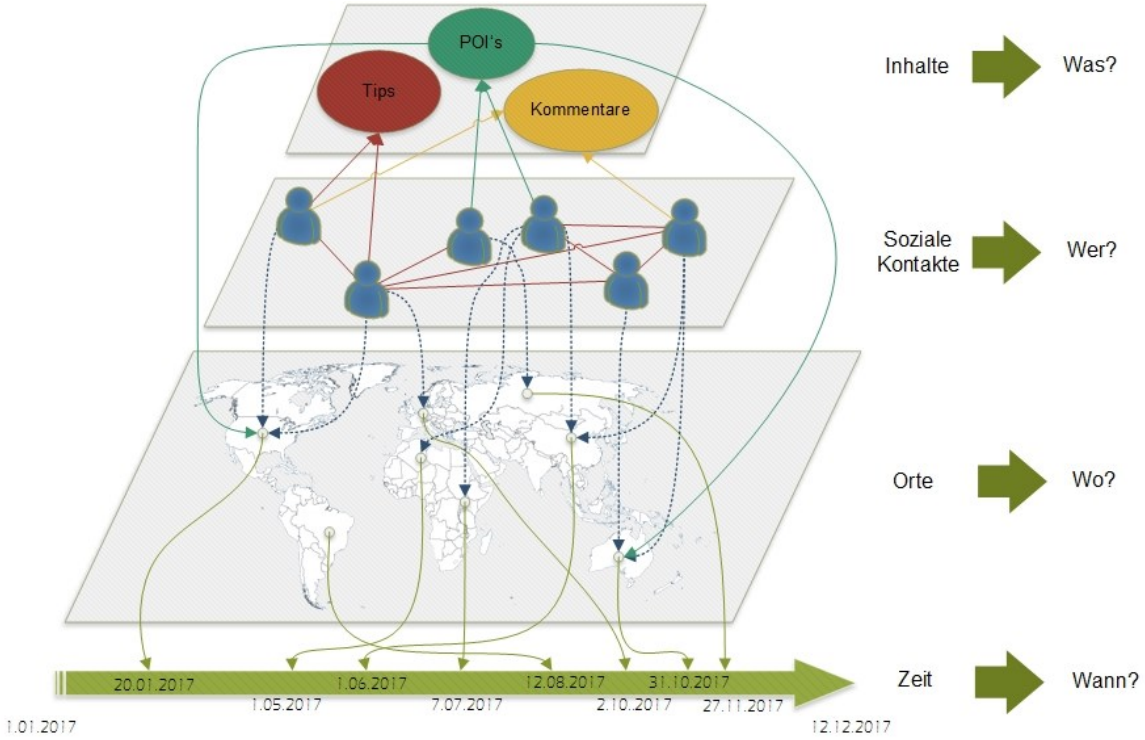


Abbildung 2-5: Funktionsweise von LBSNs (MORSTATTER et al. 2015).

2.9.1 Kategorisierung von LBSNs

LBSNs lassen sich in verschiedene Kategorien einteilen, da die verschiedenen Netzwerke unterschiedliche Ziele und Funktionen aufweisen. Eine Einteilung der Netzwerke hat ZHENG (2011) durchgeführt (Tabelle 2-5). Prinzipiell lassen sie sich in die drei Gruppen *Geo-tagged*, *Punktbasiert* sowie *Wegbasiert* einteilen. Die verschiedenen Typen bieten hierbei einen sich unterscheidenden Informationsgehalt. So liegt bei den punktbasierten Netzwerken der Informationsgehalt unter dem, den ein wegbasiertes Netzwerk bietet. Dieses ist auch in der Lage die Verweildauer oder den Weg zum Punkt mit aufzunehmen, woraus ein besseres Verständnis für das individuelle Verhalten resultiert. Gleichzeitig liefern wegbasierte Netzwerke meist keine Echtzeit-Daten, da hier oft Urlaubstouren oder ähnliches, die in der Vergangenheit liegen, hochgeladen werden. Die Daten der ersten beiden Netzwerkarten können zudem in einen Weg umgewandelt werden, indem die einzelnen Punkte verkettet werden (ZHENG 2011).

Tabelle 2-5: Klassifizierung von LBSNs (eigene Erstellung, nach ZHENG 2011).

Art	Beschreibung	Echtzeit	Informationsgehalt	Beispiel
Medien mit Geo-TAG	Fotos, Videos, Texte etc. welche Lokalisation besitzen	Normal	Gering	Flickr, Panoramio, Twitter, Instagram
Punktlokationen	Restaurants, Sehenswürdigkeiten, …→ Wer checkt wo ein?	Sofort	Normal	Foursquare, Facebook
Wegbeschreibungen	Erfassung von Joggingwegen und Erfahrungen dazu	Langsam	Hoch	Bikely, SportsDo

2.9.2 Wirtschaftliches Interesse an LBSNs

Alle Sozialen Netzwerke arbeiten profitorientiert. Sie finanzieren sich zum großen Teil über Werbung, die in personalisierten Angeboten verbreitet wird. Für diese Angebote haben sich sogar eigene Netzwerke wie Foursquare oder Yelp gegründet und auch Facebook spielt hier eine Rolle (MORSTATTER et al. 2015, vgl. Kap. 2.7.1). So gehen die Netzwerke Partnerschaften mit Unternehmen ein, damit für diese explizit auf den Plattformen geworben werden kann. Von besonderer Bedeutung für die Unternehmen sind hierbei die Interessen und Ziele einer Person, welche sich aus ihrer Positionshistorie und damit auch ihrem Kaufverhalten ableiten lassen (MORSTATTER et al. 2015, TRAYNOR & CURRAN 2013). Zu nennen ist hier ein Fall aus den USA, wo die Supermarktkette Target lediglich aus dem Kaufverhalten einer jungen Frau ableiten konnte, dass diese wohl schwanger ist (HILL 2012). Bei diesem Data Mining spielen insbesondere Twitter und Facebook eine entscheidende Rolle. Letztlich gibt die zusätzliche Geoinformation die Möglichkeit, Werbung deutlich besser auf eine Zielgruppe zuzuschneiden. Es werden aber auch virtuelle Einzugsgebiete von Läden erstellt. Wenn eine Person in die Nähe eines Ladens kommt, bekommt sie eine entsprechende Werbung auf dem Smartphone angezeigt (TRAYNOR & CURRAN 2013). Die LBSNs bieten zusätzlich kleineren Läden die Möglichkeit, auf sich aufmerksam zu machen. Zudem ist die Kundenbindung von Bedeutung. Durch das Belohnungssystem, welches beispielsweise Foursquare in Zusammenarbeit mit Domino's Pizza anbietet, ist der Kunde bedingt durch eine wöchentliche Gratispizza eher geneigt, erneut zu Domino's Pizza zu gehen. Zudem posten die Nutzer häufig selber Bilder oder Nachrichten aus bestimmten Läden, wodurch sie letztlich kostenfrei Werbung für das jeweilige Unternehmen machen (TRAYNOR & CURRAN 2013).

Außerdem spielt für die Unternehmen der soziale Aspekt eine große Rolle. Man ist i. d. R. eher dazu geneigt in einen Laden zu gehen, in welchem die Freunde bereits einkaufen waren, da häufig ähnliche Vorlieben oder Interessen existieren. Dieses Verhalten wird unter dem Begriff sozio-räumliche Eigenschaften (socio-spatial properties) zusammengefasst (SCELLATO et al. 2011). Des Weiteren ist die zeitliche Abfolge von großer Bedeutung. Wenn eine Person immer samstags einkaufen geht, wird hier zielgerichtete Werbung einen deutlich größeren Erfolg haben, als Montagmorgen, wo die Person auf Arbeit ist (MORSTATTER et al. 2015).

2.10 Erhebung von Geodaten in Sozialen Netzwerken

Geodaten sind bei Sozialen Netzwerken über vier verschiedene Methoden ableitbar:

1. Aus dem Textelement einer Nachricht
2. Aus der GNSS-Koordinaten in einem Geo-TAG der Nachricht selbst oder in einem Bild
3. Aus der IP-Adresse
4. Aus dem WLAN-Hotspot

Die verschiedenen Möglichkeiten weisen dabei sowohl unterschiedliche Genauigkeiten als auch unterschiedliche Anwendungsszenarien auf. So ist es über Geo-Tags möglich, eine Nachricht bis auf Metergenauigkeit zu lokalisieren, da hier durch das Gerät direkt GNSS-Koordinaten ermittelt und gespeichert werden (GRAHAM et al. 2014). Im Bereich von mehreren hunderten Metern liegt die Ortszuweisung über die IP-Adresse oder WLAN-Hotspots. Dennoch lässt sich über sie eine Region eingrenzen (SPLITTERBERGER 2014). Diese Informationen sind jedoch nicht für Entwickler abgreifbar – nur das Soziale Netzwerk selbst kennt diese Daten. Zudem ist die Ortsangabe durch die IP nicht immer korrekt. Ein internationales Unternehmen mit Sitz in Hamburg könnte z. B. denselben Adressenbereich

nutzen, die Außenstellen aber sind global verteilt. Dennoch würde eine IP-Verortung immer Hamburg ergeben (HAN et al. 2014). Anders als bei Textelementen oder Ortsangaben im Benutzerprofil lassen sich diese maschinell erhobenen Daten schwer durch den Nutzer manipulieren. So werden im Benutzerprofil häufig Wohnorte wie „Wunderland“ oder „Universe“ angegeben, wodurch letzten Endes nur ein Teil der Angaben einer verwertbaren Ortsangabe entspricht (GRAHAM et al. 2014). GNSS-getaggte Nachrichten haben allerdings eine entscheidende Einschränkung: Sie können nur von Geräten mit dieser Funktionalität gesendet worden sein. Zwar sind GNSS-fähige Smartphones ubiquitär in Verwendung, dennoch wird die verwendete Hardware auf eben diese eingeschränkt. Daraus folgt, dass die Daten mit GNSS-Tag andere Eigenschaften als die ohne GNSS-Tag aufweisen, denn letztere umfassen eine deutlich größere Bandbreite an Geräten, die zum Senden der Nachricht verwendet werden können (GOUWS et al. 2011).

Um den Ort einer gesendeten Nachricht zu bestimmen, muss häufig auf die Textanalyse zurückgegriffen werden, da der Anteil an Nachrichten in sozialen Netzwerken mit GNSS-Koordinaten sehr gering ist. FUCHS et al. (2013) kamen bei ihren Untersuchungen auf einen Anteil von etwa 1 % geolokalisierter Tweets in Deutschland. STEIGER et al. (2015) nehmen global einen Anteil von 2-4 % an geolokalisierten Tweets an. Mit Hilfe der Analyse des Textes kann auf eine deutlich größere Basis zurückgegriffen werden. Dies ist vor allem für ländliche Regionen und kleinere Städte interessant, da hier die Dichte an geolokalisierten Tweets nicht ausreicht, um eine repräsentative Aussage treffen zu können. Dies spielt insbesondere bei der Hazard-Forschung eine übergeordnete Rolle (CRESCI et al. 2015). Auf die Verfahren zur Textanalyse wird in Kapitel 3.3 genauer eingegangen.

3 Big Data Analyseverfahren

„Information is the oil of the 21st century, and analytics is the combustion engine."

Peter Sondergaard (1965), Vice President der Gartner Inc.

3.1 Big Data

Der Begriff Big Data stammt ursprünglich nicht aus dem Bereich der Sozialen Medien. Vielmehr fallen große Datenmengen insbesondere in der Astronomie, Physik, Meteorologie, Genetik oder den Umweltwissenschaften an. Es handelt sich oft um Datenmengen im Bereich der Peta- (1 000 Terabytes) oder Exabytes (1 000 000 Terabyte). Man kann davon auszugehen, dass sich alle zwei Jahre die Menge der Daten verdoppelt. Der Anteil der Daten, die überhaupt analysiert werden, liegt jedoch bei lediglich 0.5 % (BROWNING 2015). Dabei steht das IoT als einer der größten, neuen Datenlieferanten gerade erst am Beginn seiner Entwitcklung.

Vor allem die Speicherung und die Aufnahme dieser Datenmengen sind von großer Bedeutung. Im Bereich der Sozialen Medien hingegen ist der Datenumfang deutlich geringer. Hier steht vielmehr die Komplexität der Daten im Vordergrund, wodurch sich die Methoden von Big Data ebenfalls als Analyseverfahren empfehlen (MARR 2014). Zusätzlich zur Datenmenge (*Volume*) lässt sich Big Data mittels verschiedener Faktoren charakterisieren (Abbildung 3-1). Ein Faktor umfasst die unterschiedlichen Formate und Strukturen der Daten (*Variety*). Neben den erwähnten physikalischen Daten fallen hierunter beispielsweise Finanzdaten der Börsen. Bei diesen spielt insbesondere die Geschwindigkeit eine große Rolle, mit der die Daten generiert werden (*Velocity*). Es ist hier beispielsweise an den Hochfrequenzhandel der Börsen zu denken, oder aber an die Zahl der Tweets, die pro Sekunde generiert werden. Zudem ist Korrektheit der Daten ein wesentlicher Faktor (*Veracity*). Hier stehen vor allem Nachrichten in Sozialen Netzwerken im Vordergrund. Diese sind in der Regel unstrukturiert und enthalten zahlreiche Abkürzungen, Fehler und Dialekte. Abschließend ist noch der wichtigste Faktor von Big Data zu nennen: der Wert (*Value*). So nützen die schieren Datenmengen allein aufgrund der oben genannten Eigenschaften wenig, insofern aus diesen nicht eine Struktur oder ein Produkt generiert werden kann, welches einen Mehrwert schafft.

An den fünf V's (*Volume*, *Variety, Velocity, Veracity, Value) als* wesentliches Charakteristikum von Big Data wird sehr deutlich, dass die hier vorliegende Arbeit im Bereich von Big Data angesiedelt ist, auch wenn es sich bei den zu erwartenden Datenmengen nicht um Exabytes handelt. Dennoch ist von einem relativ großen Volumen an Nachrichten auszugehen. VETTERMANN et al. (2017a) haben gezeigt, dass innerhalb von sechs Tagen ca. 4 000 Nachrichten, welche sich Rostock zuordnen lassen, auflaufen. Von großer Bedeutung ist außerdem die Geschwindigkeit, mit der die Daten verarbeitet werden, da eine schnelle Bereitstellung insbesondere für die Entscheidungsunterstützung wichtig ist. Außerdem ist die Abschätzung der Korrektheit des Inhalts der einzelnen Nachrichten wesentlich um feststellen zu können, welche Verortungen vertrauenswürdig sind und welche nicht.

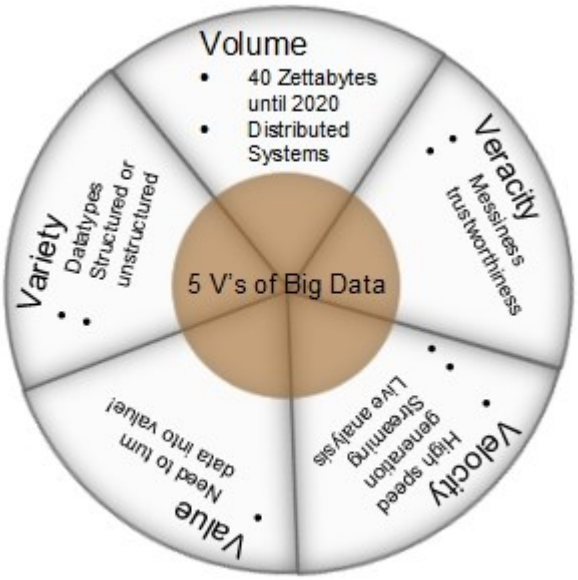


Abbildung 3-1: Die fünf V's von Big Data (eigene Erstellung, nach MARR 2014).

3.2 Datensicherheit und Datenschutz

Ein vieldiskutiertes Problem bei der Analyse großer Datenmengen ist die Wahrung der Privatsphäre. Aus diesem Grund gibt es zahlreiche rechtliche Rahmenbedingungen, die zum Schutz dieser eingeführt worden sind. Außerdem müssen entsprechende Sicherungshürden geschaffen werden, damit nicht jeder auf die Produkte, die durch die Verknüpfung der Daten erzeugt werden, zugreifen kann. Da hier die herkömmlichen Sicherungsmethoden wie Firewalls nicht greifen, sind neue Methoden zur Sicherung entwickelt worden. Allerdings ist festzuhalten, dass ein komplettes Sicherungssystem für Big Data aktuell nicht existiert. Im Wesentlichen muss ein Sicherungssystem die folgenden Bereiche berücksichtigen (MOURA & SERRÃO 2015, CLOUD SECURE ALLIANCE 2013):

1. Sicherheit der Infrastruktur
    a. Sicherheit der verteilten Datenprozessierung
    b. Sicherheit für nicht-relationale Datenbanken
2. Privatsphäre
    a. Datenanalyse durch Data Mining mit Einschränkungen des Datenschutzes
    b. Verschlüsselung
    c. Zugriffsregelungen
3. Datenmanagement
    a. Sicherer Datenspeicher und Zugriff-Logs
    b. Präzise Zugriffsüberwachung
    c. Datenherkunft
4. Reaktive Sicherheitsmaßnahmen
    a. Ende-zu-Ende Filterung und Validierung
    b. Echtzeitüberwachung der Sicherheitsstufen

Die Herausforderungen für Privatsphäre sowie sicherheitstechnische Aspekte umfassen schließlich den gesamten Verwertungszyklus bei Big Data (Abbildung 3-2).

In der vorliegenden Arbeit ist vor allem der Aspekt der Privatsphäre von besonderer Bedeutung, da die Sicherheit des Servers zumeist durch den Provider wie z. B. die Universität Rostock gewährleistet wird. Somit ist bereits ein entsprechendes Konzept zum Schutz des Servers, auf dem die jeweiligen Prozesse laufen sowie Datenbanken zu verorten sind, vorhanden (UNIVERSITÄT ROSTOCK 2017).

Das Konzept der Privatsphäre kann auch als „*right to be let alone*“ bezeichnet werden und ist bereits Ende des 19. Jahrhunderts entstanden (WARREN & BRANDEIS 1891). Das Thema Privatsphäre wird seit dem Auftreten der online basierten Sozialen Netzwerke sehr kontrovers diskutiert. Autoren wie HELLER (2011) sprechen vom Zeitalter der Post-Privacy, das heißt, dass durch sie das Ende der bisherigen Privatsphäre ausgerufen wird. Demgegenüber stehen zahlreiche abweichende Meinungen. So gibt es verschiedene Ansätze, die Privatsphäre in Mikroblogs wie Twitter zu schützen, die entsprechend von den Anbietern umgesetzt werden. Auf diese Weise lässt sich die Vernetzung bestimmter Information durch den Autor der Nachricht entsprechend beschränken (SCHMIDT 2012, ZIEGELE & QUIRING 2012). Dennoch haben verschiedene Studien gezeigt, dass die Nutzer den Umgang mit ihren privaten Informationen bei Twitter und Co. eher sorglos handhaben. HERRING et al. (2007) fanden heraus, dass 66 - 79 % in ihren Blogs Vor- oder den gesamten Namen des Autors nennen. Ähnliches beschreiben HUFFAKER & CALVER (2005), nämlich, dass 67 % von jugendlichen Bloggern ihr Alter oder weitere Informationen wie ihre E-Mail-Adresse (61 %) veröffentlicht haben.

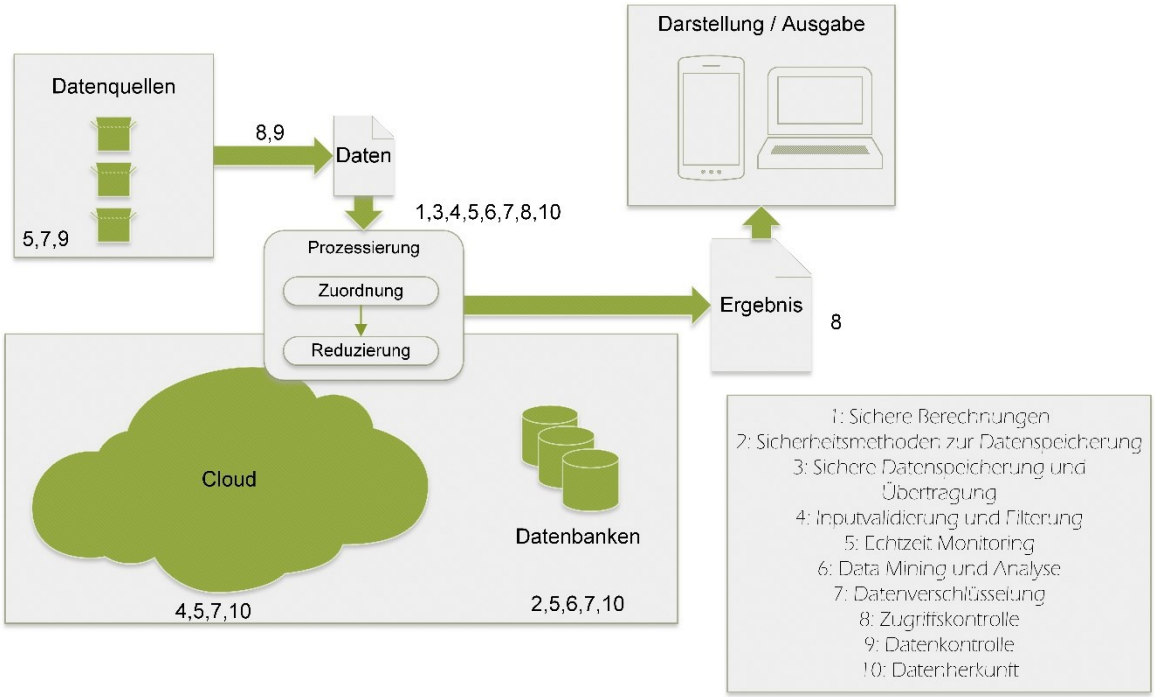


Abbildung 3-2: Grundlegende Struktur bei dem Verarbeitungszyklus von Big Data (eigene Erstellung, nach CLOUD SECURE ALLIANCE 2013).

Um eine uneingeschränkte Nutzung der Daten zu verhindern, besitzt jedes Unternehmen im Bereich der Sozialen Medien eine eigene Datenschutzrichtlinie sowie Regelungen, wie mit den zugänglichen Daten umzugehen ist[19,20,21,22] (SPLITTERBERGER 2014). In der Regel behalten sich die Unternehmen vor, Nachrichten zu löschen, welche gegen ihre Nutzungsbedingungen verstoßen.

Wesentlich ist, dass sich alle Analysemethoden in dieser Arbeit auf die frei zugänglichen Daten der Nutzer der Sozialen Netzwerke beziehen. Das heißt, die Nutzer willigen explizit ein, dass diese Daten durch Dritte betrachtet werden können (SPLITTERBERGER 2014). Nichtdestotrotz ist ein Schutz der Daten als auch deren Sicherheit wichtig und notwendig (TWITTER INC. 2018).

3.3 Analyseverfahren

Für Big Data gibt es eine große Zahl von Analyseverfahren, die sich in der Regel auf die Sozialen Netzwerke übertragen lassen. Sie lassen sich hinsichtlich ihrer Ziele gliedern. Im Wesentlichen stehen die folgenden Schwerpunkte im Vordergrund (GANDOMI & HAIDER 2015):

1. Textanalyse / Text Mining
    a. Textzusammenfassung
    b. Fragebeantwortung
    c. Sentimentanalyse
2. Audioanalyse
    a. Phonetisch basierte Auswertung
    b. Large Vocabulary Continuous Speech Recognition (LVCSR)
3. Videoanalyse
    a. Serverbasierte Analyse
    b. Kantenbasierte Analyse
4. Analyse Sozialer Medien
    a. Inhaltsbasierte Analyse
    b. Strukturbasierte Analyse
        i. Detektion von Gemeinschaften
        ii. Detektion des Sozialen Einflusses
        iii. Verbindungsvorhersage und Analyse
5. Vorhersage
    a. Historische Muster in die Zukunft fortschreiben
    b. Variablenanalyse um Zukunft anhand der Werte vorherzusagen

Für die vorliegende Arbeit sind die Bausteine des Text Minings sowie die der Analyse der Sozialen Medien von besonderem Interesse. Allerdings ist anzumerken, dass durchaus die Videoanalyse ebenfalls im Bereich der Sozialen Netzwerke von Nutzen sein kann. Ge-

[19] Twitter: https://twitter.com/tos?lang=de
[20] https://help.instagram.com/478745558852511
[21] https://www.flickr.com/atos/pro/
[22] https://de-de.facebook.com/legal/terms

rade wenn es um kurze oder längere Videos aber auch um Bilder geht, die in den Nachrichten und Netzen enthalten sind, kann das viele Vorteile mit sich bringen. Neben dem Forschungsziel lässt sich zusätzlich die Verfahrensweise selbst unterscheiden. Hierbei ist vor allem der zugrunde liegende Algorithmus von Interesse. Prinzipiell lässt sich hier zwischen klassischen, entscheidungsbasierten Methoden sowie der Anwendung von maschinellen Lernmethoden, das heißt Neuronalen Netzwerken (NN), unterscheiden.

3.4 Text Mining in Sozialen Netzwerken

Der Begriff Text Mining wurde erstmals durch FELDMAN & DAGAN (1995) geprägt und bezeichnet die computergestützte Textanalyse. Text Mining setzt sich grundsätzlich aus vier Schritten zusammen (Abbildung 3-3): Schritt eins bezieht sich auf die Beschaffung der Information (Information Retrieval - IR), Schritt zwei auf die Analyse des Textes (Normalisierung, Löschen von Stoppwörtern, Lemmatisierung, Stemmatisierung o. ä.; Natural Language Processing - NLP), Schritt drei schließlich auf die Extraktion von Information (Information Extraction - IE) und Schritt vier auf die Wissensgenerierung aus der gewonnenen Information (MCDONALD & KELLY 2012, HOTHO et al. 2005). Es gibt für das Text Mining in Sozialen Netzwerken grundsätzlich zahlreiche Ansätze, von denen hier vier Methoden beschrieben werden.

Eine erste Methode ist, nach Schlüsselwörtern zu suchen. Hierfür werden einige Schlüsselwörter definiert, z. B. „hoch“, „wasser“, sowie „waser“, um alle Nachrichten mit dem Inhalt Hochwasser filtern zu können (vgl. FUCHS et al. 2013). Von besonderer Bedeutung ist hierbei die Suche nach Experten zu einem bestimmten Thema. So ist ein Post eines Experten deutlich relevanter als der eines normalen Nutzers (ZHANG et al. 2007). In der Theorie können Soziale Netzwerke als ein Graph angenommen werden, bei dem jeder Knoten Textinformationen enthält. Diese ist meist öffentlich zugänglich, weshalb sie eine gute Basis für eine Suche darstellt. Hier setzt die Schlüsselwortsuche an (AGGARWAL & WANG 2011).

Zweitens besteht die Möglichkeit, die Social-Media-Elemente mit Beschreibungen zu verbinden. Diese beschriebenen Elemente können anschließend klassifiziert werden, wobei Verlinkungen hier die Klassifikation deutlich verbessern können. Für diese Verfahrensweise stehen zahlreiche Methoden zur Verfügung (AGGARWAL & WANG 2011). Die Klassifikation ist dabei ein überwachtes Verfahren, wobei zuerst dem Algorithmus Trainingsdaten zur Verfügung gestellt werden müssen (BARBIER & LIU 2011). Problem ist dabei die große sprachliche Diversität der Nachrichten. Zudem sind die Texte meist sehr kurz und eine genaue Beschreibung fehlt (AGGARWAL & WANG 2011).

Die dritte Verfahrensweise besteht darin, Cluster aus den Daten zu erstellen. Dies ist ein unüberwachtes Verfahren. Zudem müssen die Daten keine Beschreibung haben (BARBIER & LIU 2011). Es werden dabei Cluster aus Nachrichten mit ähnlichen Eigenschaften erstellt. Verlinkungen können dabei eine wichtige Rolle spielen. Durch die Kombination von Inhalt und Verlinkung kann eine deutlich höhere Qualität der Cluster erreicht werden (AGGARWAL & WANG 2011).

Eine vierte Verfahrensweise bezieht sich auf die Analyse der Verlinkungen. Da Objekte in Sozialen Netzwerken vielfältig verlinkt sind (Fotos, TAGs, Texte, u. v. m.), kann diese Information zur Analyse und Einordnung der selbigen genutzt werden. Diese Verfahrensweise wird auch als transfer learning bezeichnet. Für alle genannten Verfahren bieten Verlinkungen meist einen Mehrwert, da so das Ergebnis, bedingt durch die Struktur der Netzwerke, verbessert werden kann (AGGARWAL & WANG 2011).

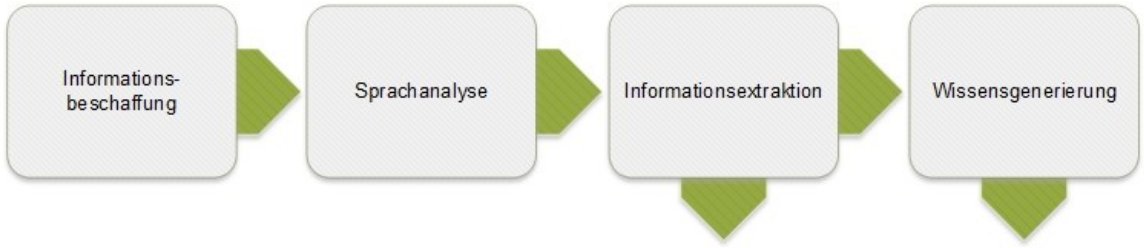


Abbildung 3-3: Die vier Schritte beim Text Mining (eigene Erstellung, nach McDONALD & KELLY 2012).

3.4.1 Vorprozessierung

Entscheidend für alle Verfahren ist die Vorprozessierung der Texte. Ein wesentlicher Schritt ist die Zerlegung des Textes in Einzelwörter (Tokenization). Dabei werden auch alle Satzzeichen, Tabstopps, Sonderzeichen etc. entfernt. Bei der Tokenbildung kommt oft auch das Part-of-Speech tagging (POS) zur Anwendung, welches den Einzelwörtern die Wortart zuweist. Zusätzlich kommen häufig die Gruppierung zu Texteinheiten (Chunks), die Ableitung des Textsinns (Word Sense Disambiguation - WSD) sowie das Parsing, d. h. die Erstellung eines Verhältnisbaums aller Wörter im Text, zur Anwendung. Zur Vorprozessierung zählen ebenfalls die bereits oben genannte Lemmatisierung und Wortstammbildung sowie das Filtern von Stoppwörtern (HOTHO et al. 2005).

3.4.2 Maschinelles Lernen

Maschinelles Lernen stellt einen wichtigen Ansatz zur Analyse von Sozialen Medien dar. Es beschreibt Methoden, die in der Lage sind, dem Computer automatisiert die Extraktion von Informationen beizubringen (WITTEN & FRANK 2005). Das maschinelle Lernen lässt sich grundsätzlich in zwei methodische Bereiche unterteilen: überwacht und unüberwacht. Zudem gibt es noch semiüberwachte Verfahren, die einen Mittelweg darstellen (KOTSIANTIS 2007).

Überwachtes maschinelles Lernen umfasst ein ganzes Paket an verschiedenen Verfahren. Es bezieht sich immer auf einen Trainingsdatensatz, der dazu verwendet wird, dem Algorithmus beizubringen, wie das Ergebnis auszusehen hat. Bei unüberwachten Verfahren ist dieses Training nicht notwendig. Ziel ist die Bildung von Clustern, die vorher noch nicht bekannt sind (Abbildung 3-4). Die Textanalyse in Sozialen Netzwerken ist ein komplexes Aufgabenfeld, bei dem die Verfahren des maschinellen Lernens oft zur Anwendung kommen. Hierfür lassen sich verschiedene Methoden finden, welche alle in einer mehr oder weniger guten Qualität die Ortsbezüge aus Nachrichten in Sozialen Netzwerken extrahieren können (ZHANG & GELERNTER 2014). Aus diesem Grund soll auf einige Verfahren in den folgenden Kapiteln näher eingegangen werden. Für die Programmiersprache Python stellt die Bibliothek *Scikit-Learn*[23] eine gute Basis dar, mit der viele der gängigen Algorithmen in Python integriert werden (PEDREGOSA et al. 2011).

[23] http://scikit-learn.org/stable/index.html

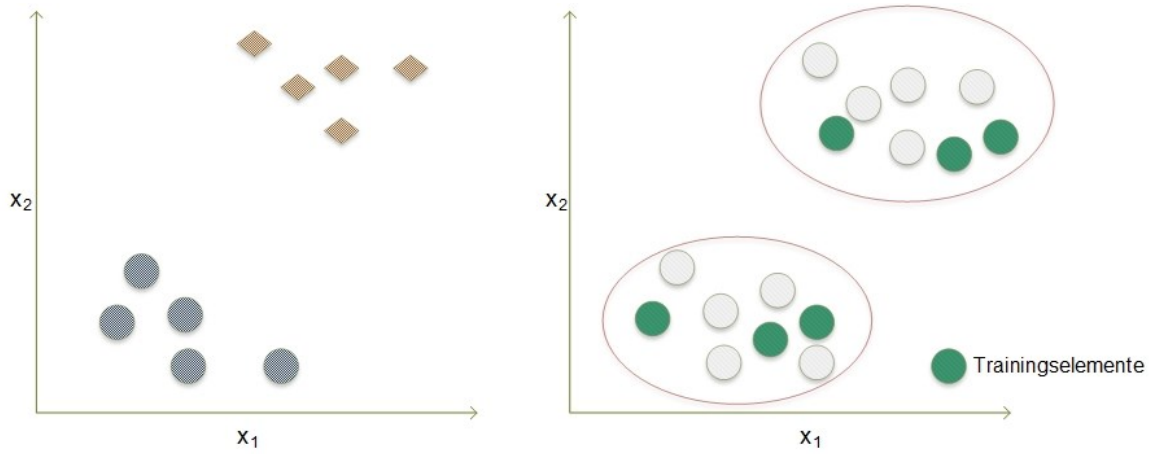


Abbildung 3-4: Überwachte (links) und unüberwachte (rechts) Klassifikation.

Eine Herausforderung ist die Überführung von Text in maschinenlesbare Information. Hierfür gibt es verschiedene Verfahren. Die simpelste Möglichkeit ist, die Wörter mittels des Bag of Words (BoW) Prinzip in Zahlen zu überführen (MANNING et al. 2008). Dabei wird das Auftreten der einzelnen Wörter gezählt und anschließend in eine Matrix geschrieben. Der Satz „The cat sat on the mat“ würde dabei den Vektor [2,1,1,1,1] ergeben. Zudem kann eine Einbindung weiterer Verfahren für die Bestimmung der einzelnen Vektoren durchgeführt werden (MANNING et al. 2008).

Eine weitere Variante sind sogenannte One Hot Vektoren (WANG et al. 2017). Dabei werden alle Wörter des Wörterbuches einer bestimmten Stelle im Vektor zugewiesen. Für den Satz „The cat sat on the mat“ würde dies folgende Vektoren zur Folge haben:

The	[1, 0, 0, 0, 0]
Cat	[0, 1, 0, 0, 0]
Sat	[0, 0, 1, 0, 0]
On	[0, 0, 0, 1, 0]
The	[1, 0, 0, 0, 0]
Mat	[0, 0, 0, 0, 1]

Der Vorteil dieser beiden Methoden ist, dass sie sehr einfach umzusetzen sind. Der entscheidende Nachteil jedoch besteht darin, dass aus der bloßen Position oder aber der Anzahl keine Bedeutung des Wortes abgeleitet und auch keine Beziehung zu den umliegenden Wörtern hergestellt werden kann. Eine Lösung hierfür stellen Word Embeddings dar. Dabei wird jedem Wort ein n-dimensionaler Vektor zugeordnet, über welchen sich die Beziehungen untereinander ableiten lassen (MIKOLOV et al. 2013a). Eine genauere Beschreibung dieser Vektoren enthält Kapitel 3.4.4, da diese Verfahren im Wesentlichen im Bereich der Sentimentanalyse eine entscheidende Rolle spielen.

3.4.2.1 Überwachte Verfahren

In den folgenden Kapiteln sollen einige wichtige überwachte Verfahren vorgestellt werden, die das Potential haben, in dieser Arbeit Anwendung zu finden. In Abbildung 3-5 ist eine Übersicht für einige der hier genannten Verfahren und ihr Zusammenhang dargestellt. Diese kann als grundlegende Einordnung der Verfahren verstanden werden.

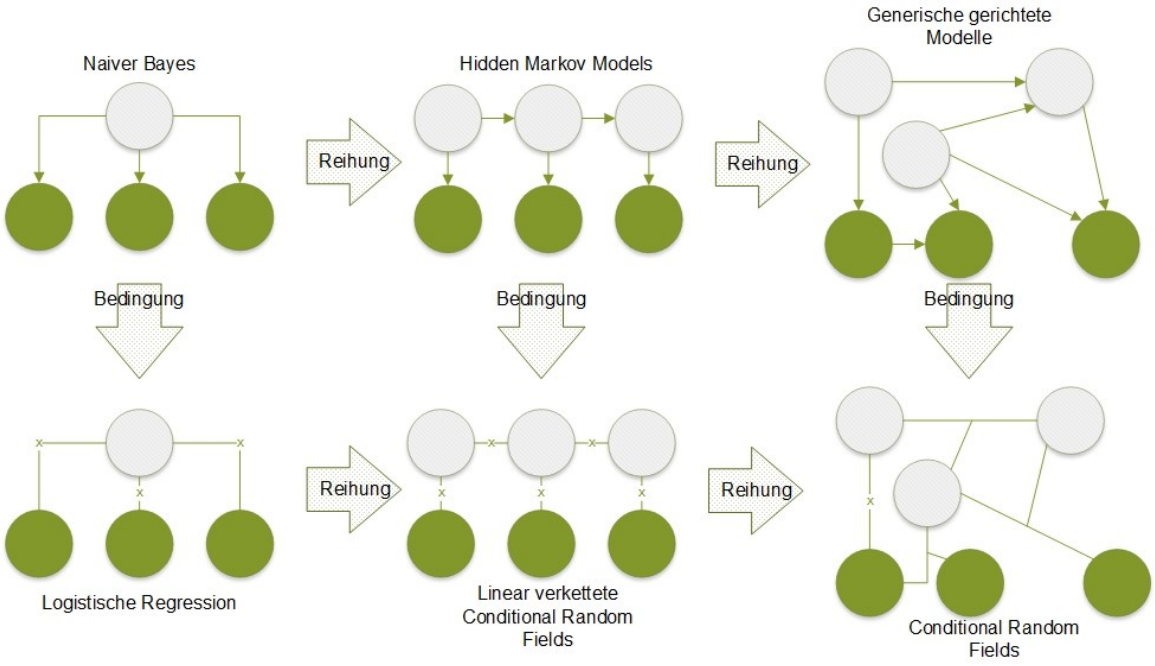


Abbildung 3-5: Überblick über die Beziehungen verschiedener Klassifikationsverfahren zueinander (eigene Erstellung, nach SUTTON & MCCALLUM 2012).

Logikbasierte Verfahren - Entscheidungsbäume

Entscheidungsbäume, auch Decision Trees, stellen ein Netzwerk aus verschiedenen Knoten dar, bei dem jeder Knoten mit einem Label sowie mit Attributen, welche geordnet oder ungeordnet sein können, versehen ist (Abbildung 3-6). Ein Entscheidungsbaum wird aus einem Trainingsdatensatz erstellt und enthält sogenannte interne Knoten (internal nodes) sowie Endknoten (leaf nodes). Jeder Entscheidungsbaum benötigt mindestens einen Endknoten, die Zahl der internen Knoten ist variabel. Jeder interne Knoten besitzt mindestens zwei Kinder (child nodes). Somit wird bei jedem internen Knoten entschieden, wie sich die Attribute an die Kinder weitergeben (splits). Die Kante vom Elternknoten zum Kindknoten ist mit der jeweiligen Merkmalsausprägung, die vererbt wird, gekennzeichnet. Die Konstruktion als solches wird als Induktion (induction), Baumbildung (tree building) oder Baumwachstum (tree growing) beschrieben (MURTHY 1998).

Um ein Objekt klassifizieren zu können, ist es notwendig, bei der Wurzel des Baumes zu beginnen und durch die einzelnen internen Knoten zu iterieren. Je nachdem wie die Entscheidung ausfällt, kommt man schließlich an einem Endknoten an, der die Klasse beschreibt. Sollte die Beschreibung des Objektes nicht zu dem des Endknotens passen, war die Klassifikation fehlerhaft. Der Anteil der korrekt klassifizierten Objekte beschreibt letztlich die Genauigkeit des Entscheidungsbaumes, der Anteil der falsch klassifizierten Objekte den Fehler (MURTHY 1998).

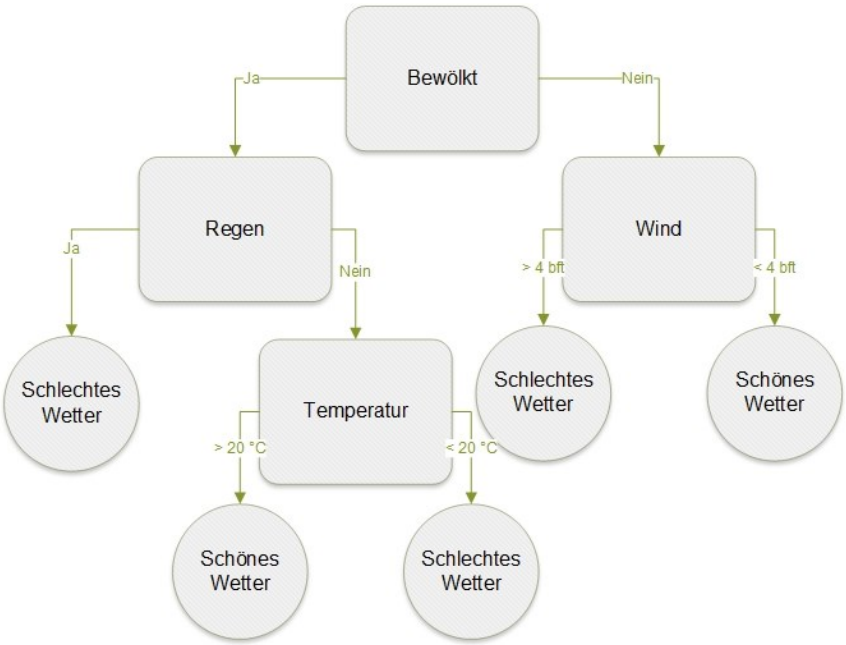


Abbildung 3-6: Beispiel eines Entscheidungsbaums bezüglich der Wettereinschätzung.

Logikbasierte Verfahren – Regelbasierte Zuordnung
Entscheidungsbäume finden sich auch in der regelbasierten Zuordnung wieder, jedoch in abgewandelter Form. In diesem Verfahren wird jedem Endknoten ein bestimmter Regelsatz zugewiesen. Ziel ist es dabei, ein möglichst kleines Regelset zu konstruieren, welches den gesamten Trainingsdatensatz umfasst (QUINLAN 1993). Je größer dabei das Regelset ausfällt, desto eher ist dies ein Indikator dafür, dass der Algorithmus mehr den Trainingsdatensatz abbildet, anstatt zu erschließen, welche Zusammenhänge sich in ihm tatsächlich verbergen (KOTSIANTIS 2007).

Neuronale Netzwerke
Neuronale Netzwerke (NN) sind der Versuch, das menschliche Denken in eine maschinelle Verarbeitung zu überführen. Hierbei entspricht jeder Knoten einem Neuron. Jedes Neuron lässt sich über die folgende Annahme beschreiben: Sind x_1 bis x_n die Eingangswerte und w_1 bis w_n die Gewichtungen, dann berechnet sich der Wert des Neurons als Summe der gewichteten Eingangswerte (Gleichung 1)

$$\sum_i x_i w_i \qquad \text{i: Nummer; x: Eingangswerte; w: Wichtungen} \qquad (1)$$

Die Ausgabe wird über einen Grenzwert bestimmt: Ist die Summe größer als der Grenzwert, ist das Ergebnis 1, ist sie geringer, ist es 0 (KOTSIANTIS 2007).

NN besitzen den großen Vorteil, dass sie in der Lage sind, sich selbst an die Datenbasis anzupassen. Außerdem ist es ihnen möglich, sich an jedwede Funktion anzunähern, was vor allem dann von Bedeutung ist, wenn die Funktion vorher unbekannt ist (HORNIK 1991, CYBENKO 1989). Zudem sind neuronale Netze nicht-lineare Modelle, wodurch sie sich an komplexere Probleme sehr gut anpassen können. Des Weiteren schätzen sie die Wahrscheinlichkeiten ab, welche die Basis für die Erstellung von Klassifikationsregeln und statistische Analysen darstellen (RICHARD & LIPPMANN 1991). Dies macht sie sehr vielseitig anwendbar und nahezu jede Problematik lässt sich mit ihrer Hilfe lösen (ZHANG 2000).

In Abbildung 3-7 ist ein sogenanntes Multi-Layer-Netzwerk dargestellt. Dieses besteht aus einer Vielzahl von Neuronen, die miteinander verbunden sind. Dabei gibt es drei Klassen an Neuronen: Input Nodes, welche die Information erhalten, Output Nodes, welche das Ergebnis liefern sowie Hidden Nodes, die die eigentlichen Neuronen zwischen Input- und Output-Nodes darstellen. Trainiert wird das Netzwerk dann, indem mittels eines Trainingsdatensatzes, in dem In- und Output bekannt sind, die Gewichtung der Verbindungen definiert werden. Das größte Problem dabei ist die Festlegung der Größe des Hidden Layers, da zu wenige Hidden Nodes sich dem gewünschten Ergebnis nur unzureichend annähern. Zu viele würden hingegen eine Überanpassung auf den Trainingsdatensatz verursachen und dadurch könnte das Problem nicht mehr generalisiert werden (KOTSIANTIS 2007).

In jüngerer Zeit finden die neuronalen Netze in der Form von Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) oder Long Short Term Memory Networks (LSTM) immer wieder im Zusammenhang mit der Textanalyse und Sentimentklassifikation, aber auch mit Bezug zur Kategorisierung nach Themen o. ä. Erwähnung (CIELIEBAK et al. 2017, GOODFELLOW et al. 2016).

Bei den CNNs findet dabei die einfache oder mehrfache Faltung der einzelnen Layer des Netzwerks statt. In der Regel geschieht dies mittels einer Multiplikation der einzelnen Werte. Anschließend werden sogenannte Feature Maps erzeugt, die über einen Max-Pooling Layer generalisiert und schließlich ausgegeben werden. Eine genaue Beschreibung der CNNs und ihrer Funktionsweise ist in GOODFELLOW et al. (2016) oder LECUN et al. (1998) zu finden.

Sind die CNNs vor allem auf die Verarbeitung von gerasterten Daten ausgerichtet, beziehen sich die RNNs insbesondere auf sequentielle Daten. Dabei teilen sich die einzelnen Teile des Modells die gleichen Parameter. Dies ist insbesondere dann von Bedeutung, wenn Informationen innerhalb einer Sequenz an verschiedenen Stellen zu finden sind. Vor allem bei Texten ist dies von besonderer Bedeutung (GOODFELLOW et al. 2016).

Neuronale Netze sind allerdings auch im Bereich der unüberwachten Klassifikationsmethoden anwendbar. Diese werden als sich selbst organisierende Netze bezeichnet. Als eine der fortschrittlichsten Technologien in diesem Bereich wird das sogenannte Deep Learning betrachtet, bei dem keinerlei Eingriff mehr in den Lernprozess des neuronalen Netzes stattfindet (LECUN et al. 2015).

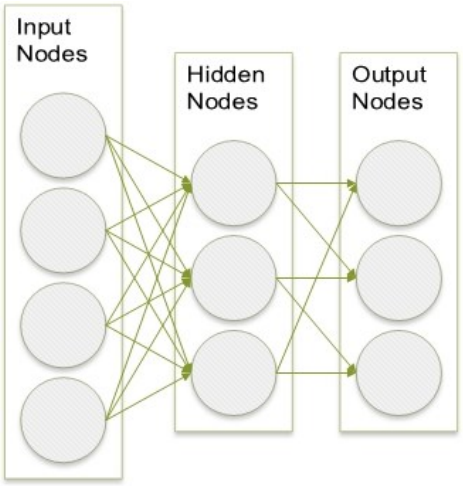


Abbildung 3-7: Beispiel eines einfachen feedforward neuronalen Netzwerks (eigene Erstellung, nach KOTSIANTIS 2007).

Statistische Verfahren – Bayes Netze

Bayes‘sche Netze beschreiben die Beziehungen zwischen einzelnen Variablen mit bestimmten Eigenschaften. Es handelt sich im konkreten Fall um einen gerichteten azyklischen Graphen (Directed Acyclic Graph - DAG). Die einzelnen Knoten sind wiederum mit Kanten verbunden, d. h. es existiert eine Eltern-Kind-Beziehung, ähnlich wie bei den Entscheidungsbäumen. Bei der Erstellung des Bayes‘schen Netzes stehen die Aufgaben der Strukturerstellung sowie die Festlegung der Parameter im Vordergrund. Die Struktur kann hierbei auch bereits vorgegeben sein, lediglich die Parameterwerte müssen gelernt werden (KOTSIANTIS 2007).

Der größte Vorteil der Bayes‘schen Netze gegenüber den Entscheidungsbäumen und neuronalen Netzen ist, dass durch die Struktur des DAG bereits eine große Menge an Vorinformation in das Netzwerk gespeist werden kann. Nachteilig hingegen ist, dass die Methodik bei einer großen Anzahl an Elementen versagt (YANG & WEBB 2003).

Es gibt unterschiedliche Bayes‘sche Netze. So ist das naive Bayes‘sche Netz bzw. der dazugehörige Klassifikator eine besonders leicht abzubildende Form, der häufig in der Analyse von Texten zur Anwendung kommt (BIRD et al. 2009). Er stellt einen DAG dar, der lediglich einen Elternknoten mit mehreren Kindern besitzt, wobei die Kinder unabhängig vom Elternknoten sind. Der große Vorteil ist die geringe Trainingszeit (KOTSIANTIS 2007). Zum Training anhand von Texten wird die Wahrscheinlichkeit jeder Beschreibung durch die Häufigkeit im Trainingsdatensatz bestimmt. Das zu klassifizierende Element wird schließlich der Beschreibung zugewiesen, für die die Wahrscheinlichkeit am höchsten ist (BIRD et al. 2009). Dies ist in Abbildung 3-8 dargestellt. Aus dieser geht hervor, dass initial das Thema des Textes im Bereich der Automobile anzusiedeln ist. Durch die Verwendung der verschiedenen Schlagwörter wird die angenommene Thematik des Textes jeweils in eine bestimmte Richtung gezogen. „Dunkel“ gibt eher eine Tendenz in Richtung des Themas Mord, „Fußball“ hingegen liefert ein sehr deutliches Indiz, dass der Text im Bereich Sport einzuordnen ist.

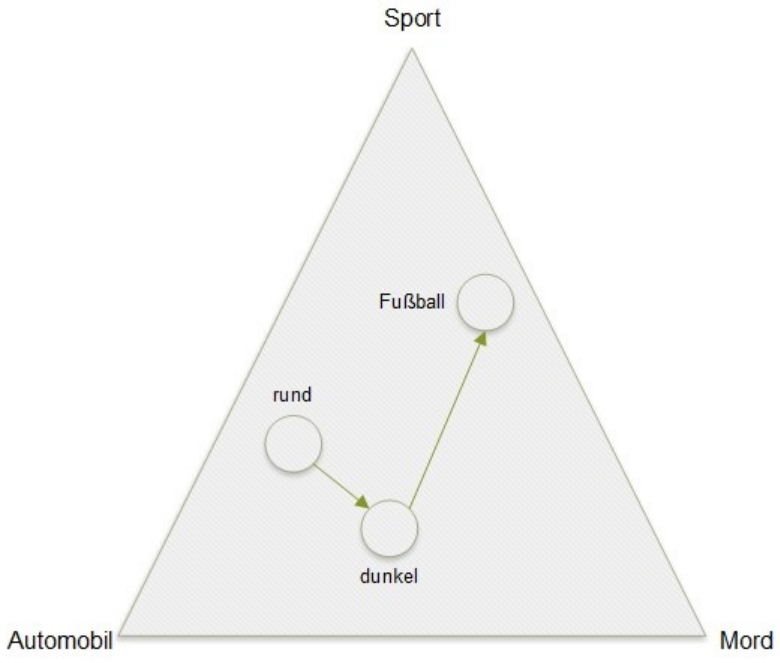


Abbildung 3-8: Zuweisung eines Themas zu einem Textelement (eigene Erstellung, nach BIRD et al. 2009).

Ein für die Textanalyse und Spracherkennung interessantes Werkzeug aus dem Bereich der Bayes'schen Netze sind die Hidden Markov Models (HMM). Sie stellen ein Werkzeug dar, Wahrscheinlichkeitsverteilungen über eine Anzahl von Werten darzustellen. Sie gehen davon aus, dass der Zustand des Prozesses S_t, der zur Erzeugung eines Elementes zum Zeitpunkt *t* geführt hat, vor dem Beobachter versteckt (hidden) ist. Die zweite Annahme ist, dass dieser versteckte Prozess das Markov-Kriterium erfüllt – d. h., dass der Zustand S_t unabhängig von allen Zuständen vorher bis *t-1* ist. Dies bedeutet, dass der aktuelle Zustand alles Wissenswerte über die Vergangenheit beinhaltet, wodurch sich der zukünftige Wert vorhersagen lässt. Daraus folgt, dass S_t und Y_t vollständig unabhängig voneinander sind, wodurch sich die Gleichung (2) ergibt:

$$P(S_{1:T}, Y_{1:T}) = P(S_1)P(Y_1|S_1)\prod_{t=2}^{T} P(S_t|S_{t-1})P(Y_t|S_t) \quad (2)$$

P: Wahrscheinlichkeit; Y: Wert; S: Prozesszustand; t: Zeitpunkt; T: Produkt der Zeitschirtte

Diese beschreibt den Graph eines Bayes'schen Netzes unter Annahme der Markov-Kriterien (GHAHRAMANI 2001).

Statistische Verfahren – Instanzbasierte Verfahren
Instanzbasierte Verfahren benötigen einen relativ geringen Trainingsaufwand, dafür ist ihre Berechnung aufwendiger. Zu den instanzbasierten Verfahren gehört beispielsweise das bekannte k-Nearest Neighbour Verfahren (kNN). Grundlage ist die Entfernung der Trainingselemente vom zu klassifizierenden Element. Auf die Klasse eines Elements wird durch die Klassen der Nachbarelemente geschlossen, wobei hier nur die Klasse mit der größten Ähnlichkeit Verwendung findet. Das k steht hierbei für die Anzahl der Nachbarn (KOTSIANTIS 2007).

Statistische Verfahren – Conditional Random Fields
Conditional Random Fields (CRF) sind ein weit verbreitetes Verfahren für die Klassifikation, beispielsweise von Texten. Sie stellen ebenfalls ein wahrscheinlichkeitsbasiertes Modell dar und das Verfahren findet in der Native Entity Recognition (NER) Anwendung (RITTER 2012, SUTTON & MCCALLUM 2012). Sie wurden entwickelt, um die Schwachstellen der diversen, bereits genannten graphenbasierten Modelle zu lösen. Die Zuordnung der Wahrscheinlichkeitsfunktionen weist bei diesen häufig eine sehr komplexe Struktur auf. Diese lässt sich dann nur noch unzureichend abbilden, wodurch das Modell fehlerhaft bzw. ungenau wird. CRFs versuchen die Wahrscheinlichkeitsverteilung bedingt und nicht probabilistisch *(y|x)* direkt abzubilden.

Der Vorteil besteht darin, dass das Modell eine deutlich einfachere Struktur besitzen kann. Die Graphen sind bei den CRF ungerichtet (Abbildung 3-5). Eine ausführlicher Überblick über die CRF ist in SUTTON & MCCALLUM (2012) gegeben.

Support Vector Machines
Support Vector Machines (SVM) stellen ein Verfahren zur Mustererkennung dar, welches durch VAPNIK (1998) entwickelt wurde. Es handelt sich dabei um ein dem kNN ähnliches Verfahren. Im Kern geht es darum, eine Hypothese zu finden, welche den geringsten Fehler im Ergebnis liefert. Es wird jedes Objekt durch einen Vektor im Raum repräsentiert. Dieser Raum soll durch eine sogenannte Hyperebene unterteilt werden, welche die Trainingsobjekte in zwei Klassen aufteilt. Der Abstand der der Ebene am nächsten liegenden Vektoren wird dabei maximiert. Alle Punkte, die außerhalb dieser Hyperebene liegen, sind

vernachlässigbar, wodurch die Komplexität der SVM nicht von der Anzahl der Objekte abhängt. Aus diesem Grund ist das Verfahren sehr gut für eine große Menge an Objekten geeignet. Eine saubere Trennung ist dabei nur dann möglich, wenn die Objekte linear trennbar sind, was in der Regel nicht erfüllt ist. Daher wird eine weitere Hyperebene eingefügt, welche den höher-dimensionalen Objektraum definiert. Diese Hyperebene wird durch die sogenannte Kernel-Funktion aufgespannt. Wird diese nun zurück transformiert in den 2D-Raum, können sich Kurven ergeben, wodurch eine saubere Trennung auch bei nicht-linear zusammenhängenden Objekten möglich ist (KOTSIANTIS 2007).

SVM können für nahezu jedes Problem verwendet werden, worunter auch die Textanalyse fällt. Sie scheinen geeignet hierfür, da es sich in der Regel um eine große Menge an verschiedenen Elementen handelt. Zudem sind nur wenige Elemente eines Textes für dessen Klassifikation unwichtig, weshalb sich hier ebenfalls die SVM empfehlen. Zudem sind die Elemente meist linear trennbar. Genau diese Elemente lassen sich nun mit SVM identifizieren (JOACHIMS 1998). Für die Analyse in Twitter kann daher das SVM-Verfahren herangezogen werden, welches neben der Bayes'schen Klassifikation ebenfalls gute Ergebnisse liefert (LEE et al. 2011).

3.4.2.2 Unüberwachte Verfahren

Neben den bereits vorgestellten überwachten Klassifikationsverfahren gibt es auch solche, die keinen Trainingsdatensatz benötigen. Dies umfasst in der Regel die Clusterung oder Segmentierung von Datensätzen. Die unüberwachten Verfahren dienen in der Regel dazu, bestimmte Gruppen (Cluster) zu bilden, da diese noch nicht vorab bekannt sind.

K-Means

K-Means ist einer der am meisten verwendeten Klassifikatoren im Bereich der Clusterbildung und unüberwachten Klassifikation. Bei dem Algorithmus muss die Anzahl K der zu bildenden Cluster vorgegeben werden. Anschließend werden die Werte, z. B. eine Auswahl von Textdokumenten, dem Zentroid des am nächsten liegenden Clusters zugewiesen. Anschließend werden die Zentroide auf Basis der Zuweisung neu berechnet. Dieser Vorgang wird so lange wiederholt, bis die Zentroide sich nicht mehr verschieben (HOTHO et al. 2005).

Neuronale Netzwerke

Unter die unüberwachten Verfahren fallen auch bestimme CNNs oder das sogenannte Deep Learning. In Bezug auf die Textanalyse in Sozialen Medien ist dieser Ansatz jedoch eher ungebräuchlich, obgleich es gerade im Bezug zu CNN auch einige Arbeiten gibt, die diese zur Analyse von Texten nutzen (STOJANOVSKI et al. 2015). Das Deep Learning wird häufig auch vom konventionellen Machine Learning abgegrenzt.

Unüberwachte CNNs stellen besondere NNs dar, welche vor allem Daten in mehreren Arrays klassifizieren sollen. Der Ausgangsdatensatz wird nach dem kompletten Element (Convolutional Layer) und den einzelnen Arrays (Pooling Layer) zerlegt. Es wird anschließend durch mehrere Ebenen iteriert und anhand des Ausgangsdatensatzes nach Mustern gesucht, die sich in den anderen Ebenen wiederfinden lassen. Diese Muster sind jeweils mit der höher liegenden Ebene verbunden. Dieses Verfahren kommt vor allem bei der Bildanalyse zur Anwendung (LECUN et al. 2015).

Latent Dirichlet Allocation

Die Latent Dirichlet Allocation (LDA) wurde erstmals von BLEI et al. (2003) vorgestellt und ist eines der potentesten Verfahren, Informationen aus Texten zu extrahieren (WOLD et al. 2016). Die LDA behandelt hierbei jedes Dokument als einen Vektor bestehend aus der Anzahl der Wörter (BoW). Jedes Dokument beinhaltet dabei diverse Themen, wobei das

einzelne Thema wiederum eine Wahrscheinlichkeitsverteilung über die Anzahl von Wörtern ist. Der Algorithmus der LDA weist jedem Dokument, in diesem Fall also jedem Tweet, ein Thema entsprechend der multinominalen Verteilung der enthaltenen Wörter zu (BLEI et al. 2003). Allerdings kann die LDA auch in Verbindung mit Bi- oder Trigrammen verwendet werden (PRABHAKARAN 2018).

Bei der Textanalyse steht man häufig vor dem Problem der Kategorisierung und Einordnung. So wird Twitter nach KWAK et al. (2010) als „*News Media*“ bezeichnet, wobei gerade bei News die Identifikation von Trends und Kategorien von großer Bedeutung ist. Hierfür ist ein unüberwachtes Verfahren notwendig, da im Vorhinein kaum entsprechende Themen identifiziert werden können, sondern diese erst im Laufe der Zeit entstehen. Daher ist die LDA dafür sehr gut geeignet (WOLD et al. 2016, BLEI et al. 2003).

3.4.3 Ableitung von Ortsbezügen

Da im Kern das Ziel der Arbeit sein soll, Ortsbezüge aus Social Media Nachrichten abzuleiten, soll an dieser Stelle auf die speziellen Vorgehensweisen eingegangen werden, wie hier zum Ergebnis gelangt werden kann. In die Extraktion werden in der Regel mehrere Verfahren eingebunden, weshalb diesem Kapitel das des maschinellen Lernens vorangestellt ist. Die Ableitung der Ortsbezüge ist dabei kein triviales Problem, auch wenn im Vergleich zur reinen Textanalyse weitere Möglichkeiten zur Verfügung stehen. In die Methodik fließen nicht nur die Texte selbst, sondern häufig auch Metainformationen ein, aus denen sich Aussagen zur Zeitzone, zum Wohnort des Nachrichtenschreibers oder aber dessen Freundschaftsbeziehungen ableiten lassen

3.4.3.1 Gazetteer Matching

Ein Gazetteer stellt ein Wörterbuch dar, welches sowohl den Ort, seine Art als auch seine geografischen Koordinaten enthält. Dadurch ist es möglich, mit Hilfe eines Gazetteers Begriffe aus Texten Orten zuzuordnen (BILL 2016). Das größte Problem dabei ist die Identifikation des Toponyms selbst, insbesondere in den kurzen und unstrukturierten Texten in Sozialen Netzwerken (ZHANG & GELERNTER 2014, MILLER & GOODCHILD 2015).

Die in der Textanalyse zur Anwendung kommenden Verfahren stellen meist eine Kombination verschiedener Verfahren des maschinellen Lernens und spezieller Textanalyse-Algorithmen dar. Vor allem auch Metainformationen, die über die API’s der einzelnen Sozialen Netzwerke gewonnen werden können, sind eine nützliche Ergänzung zum reinen Gazetteer-Matching. So bietet die Java Script Object Notation-Datei (JSON), die jedem Tweet zugrunde liegt, zahlreiche weitere Informationen (ZHANG & GELERNTER 2014). Auch die vorgestellten Methoden des Textvergleichs (Kapitel 3.4.3.2) bieten eine sinnvolle Ergänzung, da so beispielsweise Tippfehler weniger ins Gewicht fallen.

Damit stellt das Gazetteer Matching den letzten Schritt bei der Ortsbestimmung dar. Wesentlich ist, dass die Verortung jedoch nur dann mit einer hohen Genauigkeit erfolgen kann, wenn verschiedene Methoden des maschinellen Lernens und des Textvergleichs als auch Metainformationen mit einbezogen werden. Schließlich ist das Design des Gazetteers von Bedeutung, da so entscheidungsbasierte Regeln integriert werden können, welche die Genauigkeit sowie die Auflösung des Matchings erhöhen (VETTERMANN et al. 2017a).

3.4.3.2 Text-Similarity Funktionen

Im Bereich der Textanalyse gibt es eine große Zahl an Algorithmen, die sowohl zum Vergleich von Texten (Similarity Funktionen) als auch zur thematischen Einordnung von Tex-

ten dienen. Diese sind insbesondere von Bedeutung, wenn es um den Vergleich mit Wörterbüchern geht. Durch sie lassen sich Wortvorschläge ermitteln, Tippfehler korrigieren oder Wörter aus einem Wörterbuch selektieren, deren Ähnlichkeit sehr hoch ist.

Um einen Überblick über die Funktionen zu geben sei auf Tabelle 3-1 verwiesen. Grundsätzlich lassen sich die Vergleichsalgorithmen in drei Bereiche einteilen: textbasiert, korpusbasiert und wissensbasiert. Textbasierte Methoden (auch lexikalischer Textvergleich) umfassen Algorithmen wie die Jaro-Winkler-Distanz, die Levenshtein-Distanz, N-Grams u. v. m. (GOMAA & FAHMY 2013). Die korpusbasierten sowie die wissensbasierten Methoden hingegen lassen zusätzlich die Beziehung der Wörter untereinander einfließen. Daher werden sie als linguistische bzw. semantische Vergleichsmethoden bezeichnet (GOMAA & FAHMY 2013). Da in der vorliegenden Arbeit kein Korpus nur für die Textoperationen erstellt werden soll, kommen nur die Funktionen des textbasierten Vergleichs sowie die wissensbasierten Methoden in Frage. Durch die Kombination verschiedener Verfahren soll die Genauigkeit des Textvergleichs verbessert werden. Dass eine kombinierte Herangehensweise erfolgsversprechend im Bereich der Identifikation und Zuordnung von Location Indicative Words (LIWs) ist, haben KIM et al. (2016) gezeigt. Vor allem die Integration einer räumlichen Komponente kann hier für eine weitere Verbesserung sorgen.

Bezüglich der Genauigkeit der textbasierten Methoden soll an dieser Stelle noch die Arbeit von RECCHIA & LOUWERSE (2013) angeführt werden. Für den Textvergleich raumbezogener Wörter in der deutschen Sprache empfiehlt sich mit einer Genauigkeit (Precision/Recall) von 0.75/0.74 die Skip-Grams-Methodik. Auch Bigrams (0.75/0.73) und Trigrams (0.75/0.74) liefern sehr gute Ergebnisse. Die Levenshtein-Distanz hingegen schneidet mit 0.61/0.60 deutlich schlechter ab. Daher empfiehlt es sich für diese Arbeit die Nutzung von Trigrams oder Skip-Grams im Rahmen des Textvergleiches anzuwenden.

Tabelle 3-1: Ausgewählte Text-Similarity-Funktionen (eigene Erstellung, nach GOMAA & FAHMY 2013).

Kategorie	Methode	Beschreibung
Textbasiert - Zeichenbasiert	Longest Common Substring (LCS)	Misst die längste Kette gleicher Buchstaben in beiden Texten.
	Levenshtein	Unterschiedlichkeit entspricht der Anzahl an Operationen um Text 1 in Text 2 umzuwandeln.
	Jaro	Basiert auf der Ordnung und der Nummer der Buchstaben in beiden Texten.
	Jaro-Winkler	Wie Jaro, wobei die Ähnlichkeit anhand einer vordefinierten Skala beschrieben wird.
	Needleman-Wunsch	Dynamischer Algorithmus aus der Bioinformatik, der mit Hilfe von Backtracking zwei Elemente miteinander vergleicht um global die optimale Anpassung zu identifizieren.
	Smith-Waterman	Ähnlich Needleman-Wunsch, allerdings mit Anpassung auf lokaler Ebene.
	N-Gram	Der Text wird in Sequenzen der Länge n zerlegt und mit dem zweiten verglichen. Die Distanz errechnet sich aus der Anzahl der gleichen N-Gramme durch die maximale Anzahl.
Textbasiert - Satzbasiert	Block-Distanz	Berechnet sich aus der Entfernung auf einem Grid zwischen zwei Punkten.
	Kosinus-Ähnlichkeit	Wird aus dem Kosinus des Winkels zweier Vektoren, d. h. den beiden Texten berechnet
	Dice Koeffizient	Errechnet sich aus dem Anteil der N-Gramme, die in beiden Texten vorhanden sind, geteilt durch die Gesamtzahl der N-Gramme. Beim Überlappungskoeffizienten wird eine vollständige Übereinstimmung angenommen, wenn der zweite Text eine Teilzeichenfolge des ersten ist.
	Euklidische Distanz	Ist die Wurzel aus der Summe der quadrierten Differenzen zweier Vektoren.
	Jaccard Ähnlichkeit	Anzahl gleicher Textelemente gegenüber unterschiedlicher Textelemente.
	Matching-Koeffizient	Summe der gleichen Textelemente in beiden Texten.
Korpusbasiert	Hyperspace Analogue to Language (HAL)	Es wird davon ausgegangen, dass die Nachbarn eines Wortes seinen Kontext beschreiben. Mit einer 10-Wort-Matrix wird nun über den Text iteriert. Sie scheinen dann ähnlich/ gleich, wenn sie von denselben Nachbarn umgeben sind.
	Latent Semantic Analysis (LSA), Generalized Latent Semantic Analysis (GLSA)	Es wird angenommen, dass Worte gleicher Bedeutung in den gleichen Textabschnitten vorkommen. Aus der Anzahl der Wörter je Absatz werden die Wörter bzw. die Textelemente (GLSA) mit Hilfe des Kosinus ihrer beiden Vektoren verglichen.
	Explicit Semantic Analysis (ESA)	Ein Wikipediabasierter Ansatz, bei dem die Texte als Vektoren mit TF-IDF gewichtet werden. Der Vergleichsvektor entstammt aus einem Wikipedia-Artikel. Der Zusammenhang wird wieder durch den Kosinus zwischen den Vektoren beschrieben.
	Normalized Google Distance (NGD)	Basiert auf der Anzahl, wie viele Ergebnisse die Google-Suche für ein Set an Schlüsselwörtern zurückgibt. Sind die Suchergebnisse zwischen zwei Wörtern komplett unterschiedlich, d. h. der Begriff kommt niemals zusammen auf zwei Seiten vor, so ist die Distanz unendlich.
Wissensbasiert	Informationen aus semantischen Wortnetzen	Der Grad der Übereinstimmung wird mit Hilfe semantischer Wortnetze ermittelt. Die Ähnlichkeit selbst wird dabei auf Basis von Beziehungen oder der semantischen Ähnlichkeit bestimmt.

3.4.3.3 Named Entity Recognition

Die Eigennamenerkennung, auch Named Entity Recognition (NER) genannt, bezeichnet ein Verfahren zur Textanalyse, mit dessen Hilfe der Typ eines Wortes erkannt werden kann (Ort, Unternehmen, Person, …). Das Verfahren wird in der Regel unter Verwendung von Methoden des maschinellen Lernens angewendet. Es können hierbei verschiedene Verfahren aus dem Bereich der überwachten Methoden zur Anwendung kommen. Darunter fallen beispielsweise CRF (MCCALLUM & LI 2003), HMM (BIKEL et al. 1997), Entscheidungsbäume (SEKINE et al. 1998) oder SVM (EKBAL & BANDYOPADHYAY 2008). Aber auch unüberwachte Verfahren können beispielsweise in Verbindung mit semantischen Wortnetzen wie *WordNet*[24] zur Anwendung kommen. Dabei werden die Typen der Eigennamen aus WordNet bezogen und anschließend die Wörter im Text entsprechend zugeordnet (ALFONSECA & MANANDHAR 2002). Auch Online-Zeitungsartikel können genutzt werden, da bestimmte Entitäten immer zeitgleich auftauchen. Dadurch können seltene Entitäten zugeordnet werden (SHINYAMA & SEKINE 2004).

Gerade bei Tweets stößt die NER allerdings häufig an ihre Grenzen. Die Ursachen dafür liegen zum einen an der Begrenzung auf 140 respektive 280 Zeichen. Dadurch fehlt ein Gros an Kontextinformation, wie sie beispielsweise bei Zeitungsartikeln vorhanden ist. Zum anderen gibt es eine große Anzahl an Typen, die erkannt werden müssen, über die aber nur selten geschrieben wird. So muss eine NER in der Lage sein, neben Orten auch Bands, Filme, Produkte, Firmen etc. zu erkennen. Daher liefert die *Stanford NER*[25] auch nur Ergebnisse mit unzureichender Genauigkeit. So fällt die Genauigkeit von 97 % auf 80 % bei Tweets ab (RITTER et al. 2011).

Eine reine Lokalisation mittels NER bietet den Vorteil, dass eine aufwendige Gazetteer-Erstellung nicht mehr notwendig ist. Diese Methodik führt allerdings zu recht ungenauen Ergebnissen, weshalb sich eine Methodenkombination empfiehlt (GELERNTER & MUSHEGIAN 2011). Durch eine NER lässt sich beispielsweise zuordnen, bei welchen Wörtern es sich um Ortsbezeichnungen handelt. Anschließend werden nur noch die als Orte getaggten Wörter mit dem Gazetteer verglichen. Durch diese Methodik kann der Aufwand, ein Matching mit dem Gazetteer herzustellen, erheblich verringert werden.

3.4.4 Sentimentanalyse

Ein wesentlicher Teil bei der Untersuchung und Analyse von Texten ist die Identifikation von Stimmungen, auch als Sentimentanalyse bezeichnet. In Sozialen Netzen werden diese sehr häufig durch Smilies und Emoticons zum Ausdruck gebracht bzw. unterstützt. Smilies sind hierbei in der Regel eine Abfolge verschiedener Satzzeichen wodurch z. B. Lachen oder Traurigkeit zum Ausdruck gebracht wird („:-) :P :-/“). Emoticons hingegen stellen Unicode-Elemente dar, welche die Stimmungen in einem schöneren Design visualisieren. Daneben gibt es auch eine große Zahl neutraler Emoticons, wie Züge, Fahnen o. ä. Gerade in kurzen Textnachrichten spielen sie eine besondere Rolle, da so mit wenigen Zeichen dem geschriebenen Text eine zusätzliche emotionale Komponente und Bedeutung zugewiesen werden kann (DESHWAL & SHARMA 2016, SIDARENKA & STEDE 2016, AGGARWAL 2011, AGARWAL et al. 2011).

Neben den Smilies und Emoticons spielt zudem der Text selbst eine bedeutende Rolle. Einzelne Wörter (hass, gut, böse) oder aber Wortkombinationen (na toll, mega schnell, einfach klasse) sind dabei mit einer positiven bzw. negativen Bedeutung belegt. Um diese Wortkombinationen zu identifizieren, kann auf die N-Gramme zurückgegriffen werden.

[24] https://wordnet.princeton.edu/wordnet/
[25] https://nlp.stanford.edu/ner/

Dadurch können, nachdem der Text in Einzelwörter umgewandelt wurde, verschiedenste Wortkombinationen erfasst und zum Trainieren des Klassifikators verwendet werden.

Ein Verfahren, was vor allem bei großen Dokumenten häufig zur Anwendung kommt, ist Term Frequency – Inverse Document Frequency (TF-IDF, BODENDORF 2003). Hierbei wird die Häufigkeit des Auftretens eines bestimmten Terms oder Suchworts ins Verhältnis aller Terme bzw. Wörter des Dokuments gesetzt. Durch die inverse Dokumenthäufigkeit wird nun die Bedeutung des einzelnen Terms für alle betrachteten Dokumente ermittelt. Die Gleichung zur Berechnung der TF-IDF lautet wie folgt:

$$tf(t,D) = \frac{t,D}{max_{t' \epsilon D}(t',D)} \quad \text{tf: Häufigkeit; t: Term; D: Dokument; t': alle Terme} \quad (3)$$

$$idf(t) = log \frac{\mathbb{N}}{\sum_{D:t \epsilon D} 1} \quad \text{idf: Inverse Häufigkeit; } N\text{: Anzahl aller Dokumente} \quad (4)$$

$$tf.idf(t,D) = tf(t,D) * idf(t) \quad (5)$$

TF-IDF wird dabei im Rahmen der Sentimentanalyse vor allem genutzt, um die Bedeutung der einzelnen Wörter herauszufinden, da Wörter die häufig in einer kleinen Gruppe von Dokumenten auftreten, einen höheren Wert haben, als sehr häufige, jedoch informationslose Wörter wie beispielsweise der, die, das oder dann (BESBINAR et al. o. J.). Zudem gibt es noch diverse Gewichtungen der einzelne Terme, wodurch die Formel sich an bestimmte Anforderungen besser anpassen lässt (WU et al. 2008).

Zur Identifikation von Stimmungen kommen häufig die SVMs zur Anwendung. Es hat sich gezeigt, dass die Genauigkeiten sowohl bei Unigrammen als auch Bigrammen über denen anderer Klassifikationsmethoden wie Maximum Entropie oder Naive Bayes liegen (KHARDE & SONAWANE 2016).

Ein anderer Ansatz, der vor allem bei der Verwendung Neuronaler Netze zur Anwendung kommt, ist die Umwandlung der Wörter in Vektoren (word2vec). Durch die Umwandlung der Wörter in einen Zahlenwert sind diese für das Training Neuronaler Netze nutzbar, zum anderen lässt sich so ein n-dimensionaler Merkmalsraum für jedes Wort aufspannen (Abbildung 3-9).

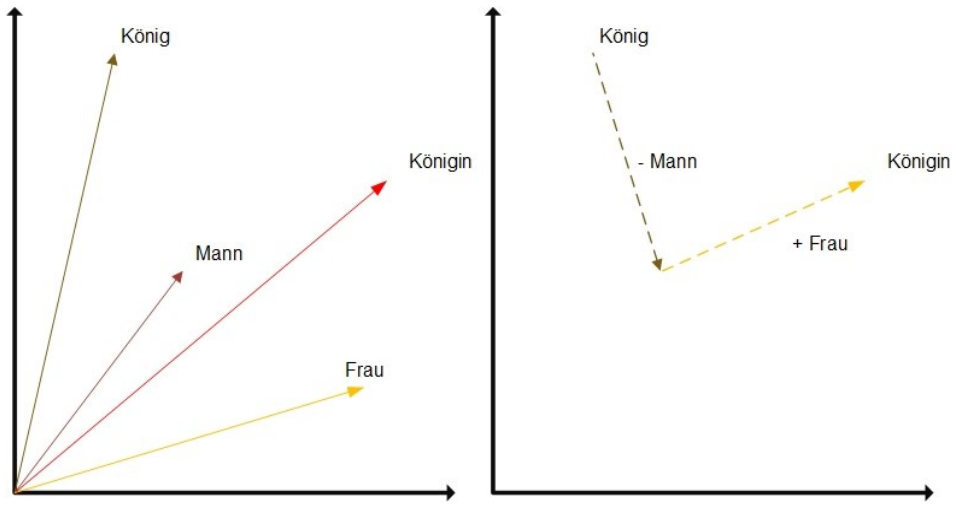


Abbildung 3-9: Konzept des Vektormodells bei Wörtern anhand der Anfrage König - Mann + Frau = ? (eigene Erstellung, nach COLYER 2016).

Der Ansatz ist nicht neu, wurde aber durch MIKOLOV et al. (2013b) überarbeitet und deutlich effizienter gestaltet durch die Verwendung eines kontinuierlichen Bag-of-Words (Continuous Bag of Words - CBOW) Modells oder aber durch die Anwendung eines kontinuierlichen Skip-Gram Modells. Dadurch wird das Training hochdimensionaler Vektoren erst mit einem vertretbaren Aufwand möglich. Die Ableitung der Vektoren findet dabei ebenfalls über Neuronale Netze statt. Bei der CBOW-Methode werden die einzelnen Wörter, die den Kontext des jeweils fokussierten Wortes bilden, in sogenannte One-Hot-Layer überführt, wobei nur das einzelne Wort den Wert eins erhält, alle anderen auf null gesetzt werden (V-dimensionaler Vektor). Diese Vektoren werden nun einem einfachen neuronalen Netz zum Training übergeben. Das Trainingsziel besteht darin, die bedingte Wahrscheinlichkeit zu maximieren, dass das tatsächliche Ausgabewort (das Fokuswort) unter Berücksichtigung der eingegebenen Kontextwörter in Bezug auf die Gewichte ausgegeben wird. Beim Skip-Gram-Modell wird das Vorgehen umgekehrt. Hier ist das Fokuswort der Eingangswert und das Ziel sind die Kontextwörter (Abbildung 3-10). Zudem gibt es noch eine ganze Reihe weiterer Verbesserungen für das Modell, wobei die Skip-Gram-Methodik hier die besten Ergebnisse aufweist (MIKOLOV et al. 2013a).

In der Regel werden Dimensionen zwischen 52 und 300 gewählt, um die Wörter optimal abbilden zu können. Dadurch können ähnliche Wörter sowie Beziehungen zwischen Wörtern modelliert werden. Deutlich wird dies an einem beliebten Beispiel (COLYER 2016). Aus diesem werden die Vorteile des Vektorenmodells deutlich. So lassen sich logische Abfragen stellen und Wörter als quasi-mathematische Ergebnisse und Elemente berechnen. So ergibt die im Beispiel dargestellte Gleichung, dass eine Königin letztlich ein weiblicher König ist, was entsprechend durch die jeweiligen Vektoren dargestellt wird. Weitere Beispiele dazu führen auch MIKOLOV et al. (2013b) an.

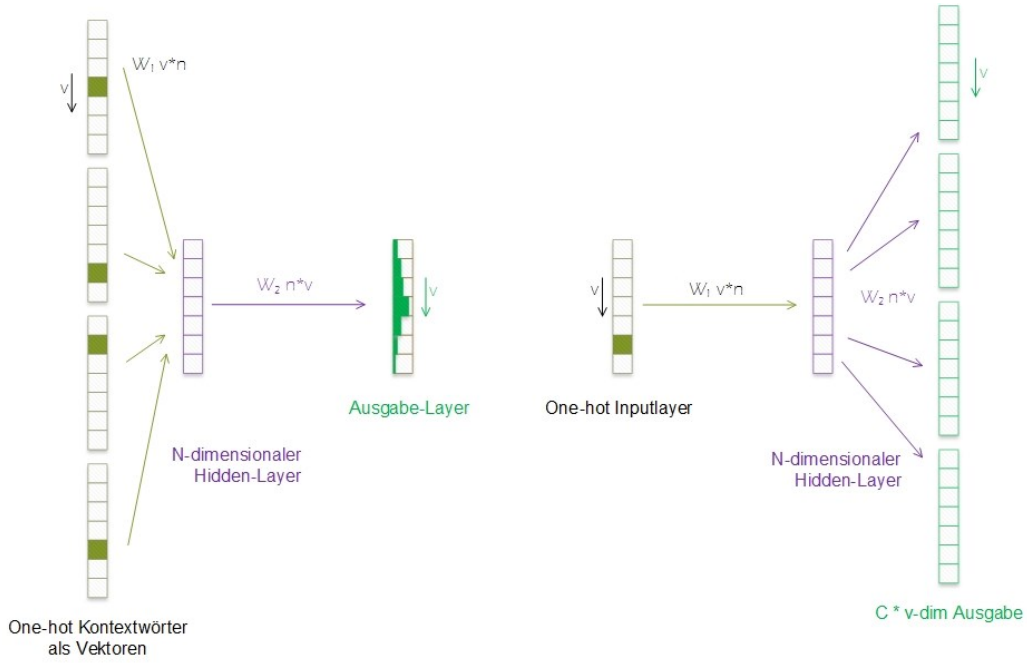


Abbildung 3-10: Berechnung des Vektormodells mittels CBAG (links) und N-Grammen (rechts) (eigene Erstellung, nach MIKOLOV et al. 2013a).

4 Programmier- und Softwareaspekte

"Software is like Sex, it's best if it's free."

Linus Thorwald (1969), finnisch-amerikanischer Informatiker

4.1 Datenbanken und Programmierung

4.1.1 PostgreSQL

PostgreSQL stellt das nach eigenen Angaben am weitesten fortgeschrittene Open Source Objekt-Relationale Datenbankmanagementsystem (ORDBMS) dar. Die Entwicklung von Postgres, welches die Basis von PostgreSQL darstellt, wurde 1986 an der Universität von Kalifornien in Berkeley gestartet. 1996 wurde die bis dato aktuelle Version Postgres95 in PostgreSQL umbenannt. Seither besitzt das Softwarepaket seinen Namen (POSTGRESQL GLOBAL DEVELOPMENT GROUP 2017). Die aktuelle Version 9.6 beinhaltet zahlreiche, moderne Features. Dazu zählen:

- Komplexe Queries
- Fremdschlüssel
- Trigger
- Updatebare Views
- Transaktionale Integrität
- Multiversion Concurrency Control (MCC)

Zudem ist der Anwender in der Lage, PostgreSQL nach seinen Wünschen zu erweitern. Es lassen sich neue Datentypen, Funktionen, Operatoren, Sprachen und Indexmethoden integrieren. Der große Vorteil ist, dass PostgreSQL unter einer Open Source Lizenz steht, wodurch es kostenfrei an die Bedürfnisse eines jeden Anwenders angepasst werden kann (POSTGRESQL GLOBAL DEVELOPMENT GROUP 2017).

Durch die weite Verbreitung von PostgreSQL bietet es auch eine Vielzahl von Erweiterungen, die letztlich der ausschlaggebende Grund für die Verwendung waren. Dazu zählt insbesondere die Bibliothek PostGIS. PostGIS ermöglicht es, Geodaten in PostgreSQL abzuspeichern. Durch die Erweiterung können ortsbasierte Abfragen in PostgreSQL durchgeführt werden, womit das Plugin für alle geografischen Fragestellungen von essentieller Bedeutung ist (POSTGIS PROJECT STEERING COMMITTEE 2015). Neben PostGIS bietet das *pg_similarity*[26] Plugin die Möglichkeit, verschiedene Textvergleichsalgorithmen auf Datenbankebene einzubinden (OLIVIERA 2012). Diese Kriterien waren für die Auswahl von PostgreSQL als Datenbankmanagementsystem für die vorliegende Arbeit ausschlaggebend.

4.1.2 Python

Python ist eine Programmiersprache, welche 1994 durch Guido van Rossum in der Version 1.0 veröffentlicht wurde. Seither sind die Versionen 2 und 3 erschienen, die jeweils eine Vielzahl an neuen Funktionen integriert haben. Die aktuellste Version ist Python 3.7.2 (Stand: 06.02.2019). Das Ziel Pythons ist es, Programmcode möglichst einfach und über-

[26] https://github.com/eulerto/pg_similarity

sichtlich zu gestalten. So besitzt die Sprache eine relativ geringe Anzahl an Schlüsselwörtern und die Strukturierung des Codes wird anstatt mit Klammern mit Einrückungen durchgeführt. Zudem ist Python Open Source (PYTHON SOFTWARE FOUNDATION 2017).

Python wurde als Programmiersprache ausgewählt, da hier, ähnlich wie bei PostgreSQL, eine Vielzahl an Erweiterungen existiert, die für die vorliegende Arbeit von großer Bedeutung sind. Zuerst sei hier die Bibliothek *twython*[27] genannt. Mit dieser ist es möglich, auf die Twitter API zuzugreifen und den Nachrichtenstream entsprechend zu verarbeiten (MCGRATH 2018). Für Instagram existiert die Schnittstelle *Python-Instagram*[28].

Die Programmbibliothek *psycopg2*[29] stellt die Schnittstelle zwischen Python und PostgreSQL her. Diese Verbindung ist sowohl zum Speichern als auch zum Laden von Informationen essentiell. Über sie wird beispielsweise der Gazetteer in Python eingebunden (GREGORIO & VARRAZZO 2018).

Für die Sprachanalyse wird auf das *Natural Language Toolkit (NLTK)*[30] zurückgegriffen. Diese Bibliothek ermöglicht es, eine Vielzahl von Operationen des Textprocessings in Python durchzuführen. Dazu zählt die Bildung von N-Grammen oder die Identifikation von Wörtern (BIRD et al. 2009).

Die Algorithmen des maschinellen Lernens werden über die Bibliothek Scikit-Learn realisiert (PEDREGOSA et al. 2011). Über die Bibliothek lässt sich eine Vielzahl von überwachten und unbewachten Klassifikationsverfahren in Python realisieren. Zudem bietet sie Möglichkeiten der Genauigkeitsanalyse. Des Weiteren werden mittels der Bibliothek Tensorflow und Keras die Algorithmen des Deep Learnings bzw. der Neuronalen Netze in Python integriert (ABADI et al. 2015, CHOLLET 2015). Das unüberwachte Verfahren der LDA hingegen ist über die Bibliothek Gensim eingebunden (REHUREK & SOJKA 2010).

Daneben existiert eine Vielzahl weiterer Erweiterungen, die in dieser Arbeit zur Anwendung kommen, jedoch nicht explizit erwähnt werden sollen. Häufig handelt es sich dabei um Abhängigkeiten von den verschiedenen Programmbibliotheken. Auf die weiteren Bibliotheken wird entsprechend bei der Beschreibung ihrer Anwendung eingegangen.

4.2 Programmschnittstellen

4.2.1 Twitter Streaming API

Die Programmschnittstellen (APIs) stellen den Zugangspunkt zu den einzelnen Social Media Portalen dar. Über sie ist es möglich, die Daten zu beziehen, aber auch bei Bedarf selbst Nachrichten in die Netzwerke einzuspeisen. Durch die APIs ist zudem das Auslesen der Metainformationen möglich. Jede Twitter-Nachricht lässt sich als JSON-Datei auslesen, welche diese Metainformationen beinhaltet. Eine ausführliche Auflistung aller enthaltenen Elemente ist Tabelle 4-1 zu entnehmen, die ortsrelevanten Informationen sind entsprechend hervorgehoben.

Die Streaming API bietet insgesamt drei Endpunkte, auf die zugegriffen werden kann. Dies ist zum ersten der öffentliche Stream, der alle Daten umfasst, die ohne Zugriffsbeschränkung gepostet werden. Der zweite Endpunkt sind die spezifischen Nutzer-Streams, welche

[27] https://github.com/ryanmcgrath/twython
[28] https://github.com/facebookarchive/python-instagram
[29] http://initd.org/psycopg/
[30] http://www.nltk.org/

alle Nachrichten eines Nutzers umfassen. Der dritte, sich noch in der Beta-Phase befindende Endpunkt, sind die Seiten-Streams. Diese stellen letztlich einen Multi-Nutzer-Stream dar (TWITTER INC. 2016b).

Tabelle 4-1: Beschreibung der einzelnen Elemente eines Tweets (eigene Erstellung, nach TWITTER INC. 2016b).

Bereich	Name	Beschreibung	Datentyp
Tweets	Contributor	Informationen über den Tweet-Autor	Array
	Created_at	UTC (Coordinated Universal Time) Zeit der Nachrichtenerstellung	String
	Current_user_retweet	Retweet der eigenen Nachricht	Objekt
	entities	Entities die im Text enthalten sind	Entities
	Favorite_count	Anzahl der Likes für die Nachricht	Integer
	favorited	Angabe ob Tweet geliked wurde oder nicht	Boolean
	Filter_level	Gibt das Filterlevel an, ab wann die Nachricht im Stream angezeigt wird	String
	geo	Veraltet	Object
	Id, id_str	ID des Tweets	Integer, String
	In_reply_to_	Angabe, ob der Tweet eine Reaktion ist und auf wen bzw. welche Nachricht	diverse
	lang	Sprache nach BCP (Best Current Practice) 47[31] Code	String
	place	Beschreibung eines vorgegebenen Ortes mit begrenzendem Rechteck, Name, etc.	Places
	Possibly_sensitive	Wahr, wenn Nachricht eine URL enthält	Boolean
	Quoted_status	Enthält die ID des zitierten Tweets, sollte es ein Zitat sein	Integer
	scopes	Verschiedene Wertepaare die den Kontext der Nachricht einordnen. Wird für besonders angebotene Nachrichten verwendet.	Objekt
	Retweet_count	Anzahl wie häufig Nachricht geretweeted wurde.	Integer
	retweeted	Angabe, ob Nachricht geretweeted wurde	Boolean
	Retweeted_status	Geretweetete Nachricht	Tweet
	source	Tool, mit dem Nachricht gepostet wurde als HTML	String
	text	UTF-8 (Unicode Transformation Format) Text der Nachricht	String
	truncated	Gibt an, ob der Tweet-Text eingekürzt wurde. Dies ist erkennbar an … am Nachrichtenende	Boolean
	Withheld_copyright	Festlegung eines Copyrights	Boolean
	Withheld_in_countries	Länder, in denen Nachricht nicht dargestellt wird	String Array
	Withheld_scope	Gibt an, ob der Status oder z. B. die User-Informationen dargestellt werden	String
Coordinates	Coordinates	Lat- & Lon Werte der Koordinaten	Floats
	Typ	Datentyp für die Koordinaten, z. B. Point	String
Users	Contributors_enabled	Nutzer mit aktiviertem contributor mode	Boolean
	Created_at	UTC Zeit wann Account erstellt worden ist	String
	Default_profile	Gibt an, ob Hintergrund oder Thema des Nutzerprofiles verändert worden ist	Boolean
	Default_profile_image	Gibt an, ob das Profilbild geändert wurde oder nicht	Boolean
	description	UTF-8 Beschreibung des Nutzeraccounts	String

[31] https://tools.ietf.org/html/bcp47

	entities	Entities die im Description-Feld vorkommen	Entities
Users	Favorites_count	Die Anzahl an Nachrichten die der Nutzer bisher geliked hat	Integer
	Follow_request_sent	Anzeige, ob man selbst eine Follow-Anfrage an den Nutzer gesendet hat	Boolean
	following	Veraltet	Boolean
	Followers_count	Anzahl an Followern die der Account derzeit hat	Integer
	Friends_count	Anzahl an Nutzern die dieser Account folgt	Integer
	Geo_enabled	Gibt an, ob die Funktion des Geotags für die Nachrichten des Nutzers aktiviert ist	Boolean
	Id, Id_str	Nutzer ID	Integer, String
	Is_translator	Gibt an, ob der Nutzer am Twitter-Übersetzungsprogramm[32] teilnimmt	Boolean
	lang	BCP 47 Code für die vom Nutzer angegebene Sprache	String
	Listed_count	Anzahl an öffentlichen Listen in denen der Nutzer gelistet ist	Integer
	location	Benutzerdefinierter Ort wo der Account beheimatet ist	String
	name	Benutzername	String
	notifiactions	Veraltet	
	Profile_background;profile_image	Angaben zum Nutzerbild, Nutzerhintergrundbild und Darstellung des Nutzeraccounts im Allgemeinen	Diverse
	protected	Gibt an, ob die Nachrichten geschützt sind oder nicht	Boolean
	Screen_name	Anzeigename des Nutzers	String
	Show_all_inline_media	Nicht mehr verwendet	Boolean
	status	Status eines Nutzers	Tweet
	Statuses_count	Anzahl der Nachrichten des Nutzers	Integer
	Time_zone	Zeitzone des Nutzers	String
	url	Url, welche vom Nutzer zur genaueren Profilbeschreibung angegeben werden kann	String
	Utc_offset	Unterschied von GMT in der Zeit in Sekunden	Integer
	verified	Zeigt an, ob der Account verifiziert ist	Boolean

Twitter bietet drei Versionen seiner Streaming API an. Die erste ist die freie Echtzeit-Streaming API. Mit ihr ist es möglich, maximal 1 % aller Nachrichten zu erhalten. Sind allerdings entsprechende Filter gesetzt, ist es auch möglich, 100 % der gewünschten Nachrichten zu erhalten. Die zweite API ist Decahose. Decahose umfasst 10 % aller Nachrichten. Vollzugriff auf alle Nachrichten ist nur mit der kostenpflichtigen Firehose API möglich (TWITTER INC. 2016b).

Der Zugriff ist hierbei wie in Abbildung 4-1 dargestellt geregelt. Die Authentifizierung wird über *OAuth*[33] durchgeführt. Es lassen sich zudem bestimmte Parameter für die Filterung des Streams angeben. Darunter fallen Schlagworte (track), Orte innerhalb einer geografischen Begrenzung (locations), der durch Twitter vergebene Sprachcode (language) sowie bestimmte Nutzer, denen gefolgt werden kann (follow). Mehrere Angaben können dabei kombiniert werden, wobei hier das logische *oder* gilt. Zudem kann jeder Parameter mehrere Werte erhalten, indem diese in einer durch Kommata getrennten Liste angefügt werden (TWITTER INC. 2016b). Der programmtechnische Zugriff erfolgt schließlich über diverse Bibliotheken die für Python, Java, C# u. v. m. verfügbar sind.

[32] https://translate.twitter.com/
[33] https://oauth.net/

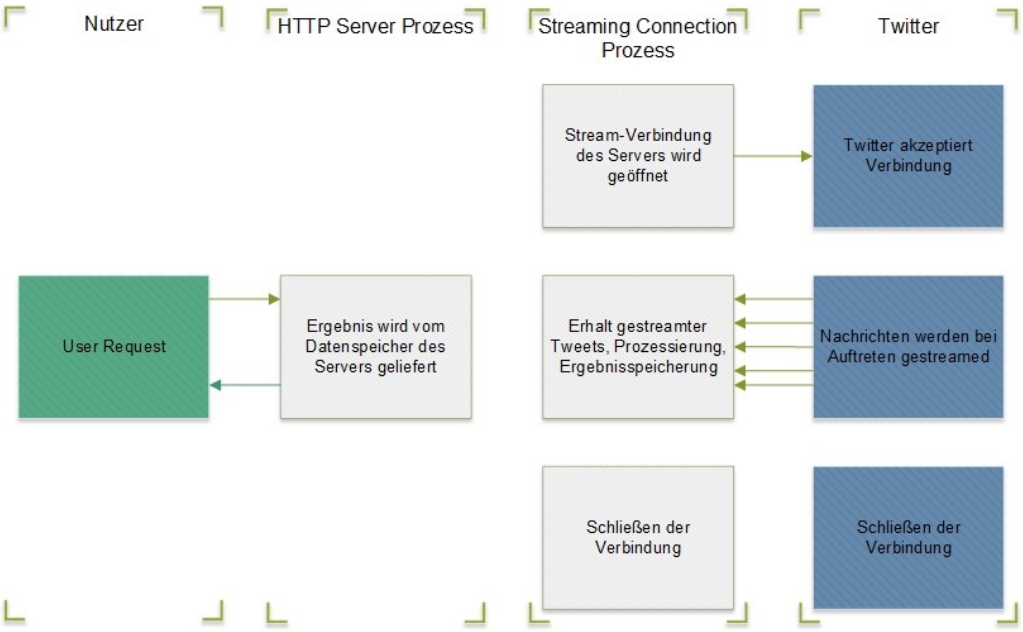


Abbildung 4-1: Funktionsweise der Twitter Streaming API (eigene Erstellung, nach TWITTER INC. 2016b).

4.2.2 Twitter REST API

Neben der Streaming API bietet Twitter auch eine Representational State Transfer-Schnittstelle (REST) an. Hier können ebenfalls Nachrichten abgerufen oder gepostet werden. Allerdings findet dabei nur ein einmaliger Zugriff statt, bei dem die Nachrichten zurückgeliefert werden (Abbildung 4-2). Eine Echtzeit-Verarbeitung wie bei der Streaming API findet nicht statt. Zudem sind nur 180 Anfragen innerhalb von 15 Minuten gestattet. Zur Authentifizierung reicht in diesem Fall allerdings bereits OAuth1 aus. Bei einer Anfrage erhält man die ebenfalls wieder filterbaren Ergebnisse aus den letzten sieben Tagen. Allerdings ist die Anzahl der Nachrichten pro Anfrage auf 100 begrenzt (TWITTER INC. 2016b). Für die in dieser Arbeit behandelte Fragestellung eignet sich die Twitter REST API aufgrund der fehlenden Echtzeit-Komponente und der begrenzten Anzahl an zurückgegebenen Anfragen nicht.

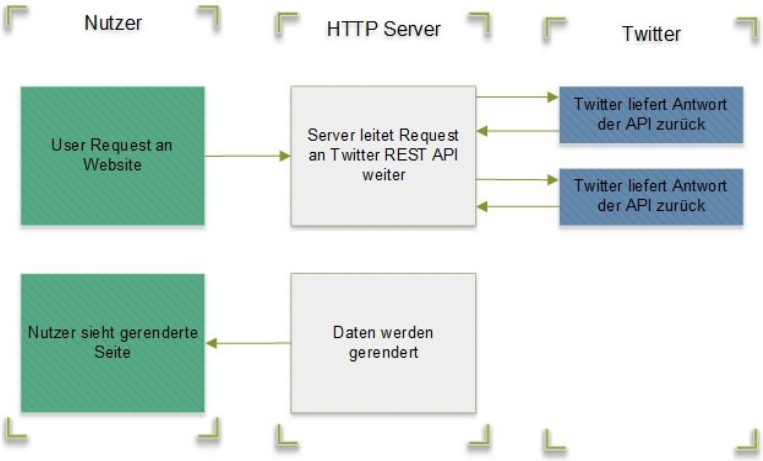


Abbildung 4-2: Funktionsweise der Twitter REST API (eigene Erstellung, nach TWITTER INC. 2016b).

4.3 Geodateninfrastrukturen

4.3.1 OpenLayers 3

OpenLayers (OL) wurde 2006 durch das OGC als frei verfügbare Alternative zu Google Maps entwickelt und veröffentlicht. Bekannt wurde das Projekt vor allem im Zuge von OSM. Die neueste Version ist OL 3, welches eine komplette Überarbeitung der älteren Versionen darstellt (SANTIAGO 2015).

OL ist eine JavaScript-Bibliothek, welche alle benötigten Komponenten zur Arbeit mit Geodaten im Browser zur Verfügung stellt. So lassen sich mit OL Daten verschiedenster Formate und Projektionen darstellen und bearbeiten. Dazu zählen GML, Keyhole Markup Language (KML) und Geo-Java Script Object Notation (GeoJSON). Des Weiteren ist OL OGC konform, d. h. es kann mit diversen Services und Formaten gearbeitet werden, die das OGC definiert hat (WMS, WFS, …). Die Darstellung der Layer ist über die sogenannte Styled Layer Description (SLD) gelöst (SANTIAGO 2015). Die Darstellung erfolgt dabei immer clientseitig, d. h. OL kommuniziert über die OGC-Standards, z. B. WMS zur Kartendarstellung, mit einem Server, der wiederum die Karte, z. B. OSM, bereit stellt (Abbildung 4-3). Die Aufrufe finden dabei mit Asynchronous JavaScript XML (AJAX) statt.

Die Überarbeitung und Umstellung von OL 2 auf OL 3 geschah insbesondere bezüglich der Anpassung auf mobile Anwendungen. Auch die Größe der API wurde drastisch reduziert und neue Renderer (WebGL, Canvas, DOM) integriert. Dabei basiert der Code auf der *Closure*[34]-Bibliothek, welche eine hohe Performance für die Anwendung liefert (SANTIAGO 2015). Des Weiteren ist mit OL 3 die Bibliothek *Cesium*[35] integriert worden. Diese ermöglicht die Darstellung von 3D-Formaten im Browser (CESIUM CONSORTIUM 2017).

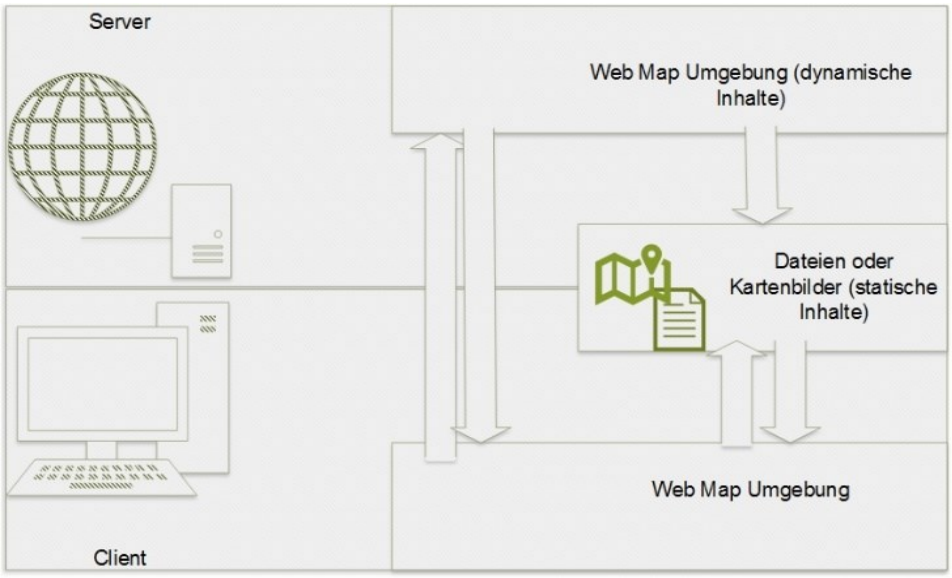


Abbildung 4-3: Client-Server-Prinzip bei OL (GRATIER et al. 2015).

[34] https://developers.google.com/closure/
[35] https://cesiumjs.org/

4.3.2 GeoNetwork

GeoNetwork stellt ein Open-Source-Metainformationssystem dar, dessen Entwicklung 2001 im Rahmen eines FAO-Projektes begonnen wurde. Es bietet durch die Integration von GeoServer die Möglichkeit, über die OGC-Standards WMS und WFS Dienste bereitzustellen und die Bearbeitung über sogenannte WPS (Web Processing Service) zu ermöglichen (SEIP et al. 2017). Die Visualisierung erfolgt via OpenLayers 3 mit seiner 3D-Erweiterung Cesium. Des Weiteren liefert GeoNetwork eine Harvesting-Funktion, mit deren Hilfe sich OGC-konforme Metadaten integrieren lassen, um eine redundante Datenhaltung zu verhindern. Der Harvester unterstützt die Dienste WMS, WFS, WCS, WPS, CSW sowie SOS (Sensor Observation Service, GEONETWORK OPENSOURCE 2014). Der grundlegende Aufbau ist in Abbildung 4-4 dargestellt.

Durch die Möglichkeit Thesauri zu integrieren, lässt sich die Suche nach Schlagwörtern deutlich erleichtern und ein hierarchischer multilingualer Aufbau erstellen. Die Integration erfolgt über das Ressource Description Framework (RDF). Diese erleichtert sowohl die automatisierte als auch die manuelle Suche deutlich (MORENO-SANCHEZ 2009). Daneben verbessert es die Handhabe mit Schlagwörtern bei der Dateneingabe, da durch vorgefertigte Thesauri die Auswahl an Schlagwörtern begrenzt wird. Über die Thesauri auf Basis des Simple Knowledge Organization Systems (SKOS) lässt sich schließlich die gesamte Datenstruktur des Portals organisieren (W3C 2009). Dies bedeutet, dass jeder Datensatz über die ihm zugewiesenen Schlüsselwörter einem Thema zugeordnet und mittels sogenannter Facets im GeoNetwork wieder aufgefunden werden kann (GEONETWORK OPENSOURCE 2014).

Um die Konformität mit anderen Datenkatalogen herzustellen, werden die ISO-Standards für Metadaten ISO 19115, 19139 (ISO-Standard für Geodaten) und Dublin Core (W3C-Standard für Dokumente, Textdateien, etc.) verwendet. Zudem kann jeder Datensatz auf seine INSPIRE-Konformität geprüft werden (GEONETWORK OPENSOURCE 2014). Um Eingabefehler gerade beim Datenupload zu verhindern, lässt sich eine Template-Bibliothek für die verschiedenen Datenbereiche einbinden. Durch diese wird der Eingabeaufwand und somit die Hemmschwelle zur Verwendung des Portals deutlich verringert. Um den beteiligten Akteuren die Möglichkeit zu geben, Datensätze direkt online zu bearbeiten, wird auf WPS zurückgegriffen (BILL 2016). Diese werden als WebGIS-Funktion des Datenportals integriert (HÜBNER 2016).

Im Rahmen des Projektes KOGGE ist dazu eine entsprechende Infrastruktur aufgebaut worden (HÜBNER & VETTERMANN 2016, HÜBNER et al. 2016), welche auch im Rahmen dieser Arbeit genutzt und entsprechend weiterentwickelt wurde.

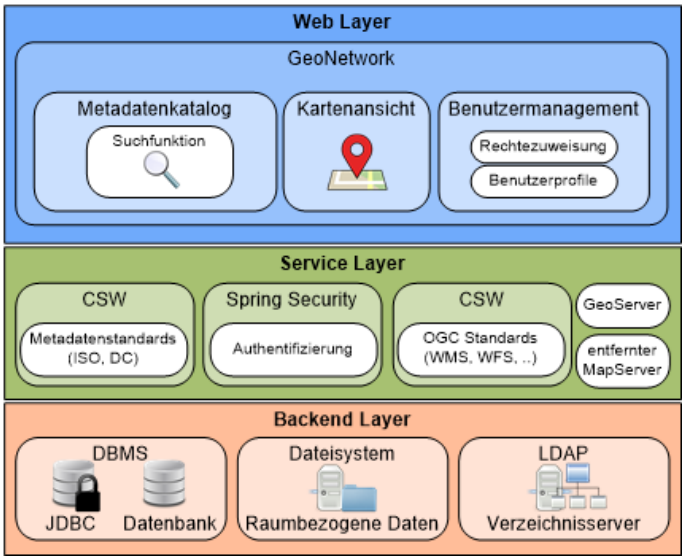


Abbildung 4-4: Architektur von GeoNetwork (KOLDRACK et al. 2017).

4.3.3 GeoServer

GeoServer ist ein freier Open Source Mapserver, der auf den Spezifikationen des OGC basiert und diese über das Internet nutzt bzw. anbietet. Dieses Grundprinzip ist in Abbildung 4-5 dargestellt. Er wurde 2001 durch The Open Planning Project (TOPP) entwickelt. Die Vision der Entwickler war dabei die Etablierung eines Geospatial Webs, welches analog zum WWW gesehen werden kann. Durch die Zusammenarbeit mit dem GeoTools-Projekt, welches ein auf Java basiertes Open Source GIS-Toolkit darstellt, konnte schließlich die Unterstützung von Shapefiles, die Einbindung von ArcSDE sowie Oracle-Datenbanken realisiert werden. Außerdem sind zeitgleich die Standards WMS sowie GML durch das OGC als auch PostGIS als freies, räumliches Datenbanksystem entwickelt worden. Inzwischen stellt GeoServer eine Referenz hinsichtlich der Implementierung von WFS, WCS und WMS dar und ist einer der am meisten genutzten Mapserver (GEOSERVER 2014).

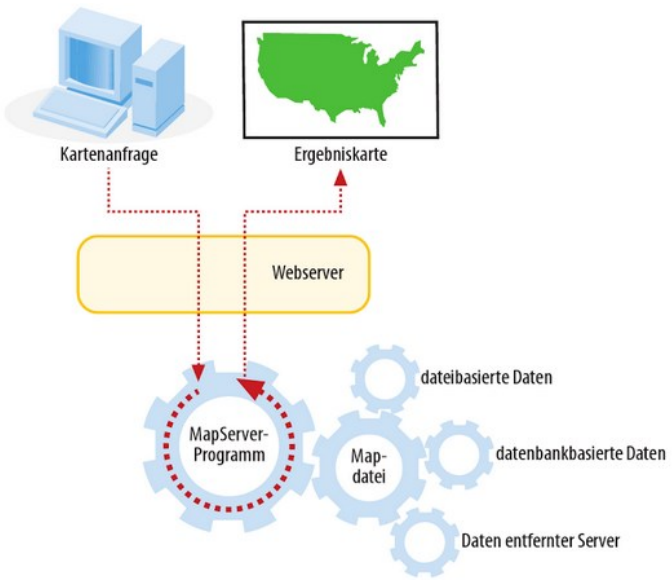


Abbildung 4-5: Grundfunktionen einer MapServer-Anwendung (MITCHELL 2008).

4.4 Auswertung und Visualisierung

4.4.1 GIS-Software

Das von ESRI entwickelte ArcGIS stellt ein Standardwerkzeug zur Erstellung von Karten, aber auch zum Ableiten geografischer Informationen dar. ESRI selbst wurde 1969 in Kalifornien gegründet. Mit der Veröffentlichung von ARC/INFO 1982 wurden die ersten Schritte in Richtung eines kommerziellen GIS-Produktes getan. Im Jahr 2004 wurde ArcGIS 9 vorgestellt, welches nicht mehr nur eine einfache Desktop-Applikation war, sondern sowohl eine Server-Plattform als auch eine Entwicklungsumgebung für eigene Erweiterungen umfasst. Damit bietet sich ArcGIS als Infrastruktur zum Erstellen von Karten und geografischen Informationen an (ESRI o.J.). In dieser Arbeit kommt das Kernprodukt ArcGIS for Desktop als auch ArcGIS Pro zur Anwendung.

Neben dem kommerziellen ArcGIS kommen im Rahmen der Arbeit zum Teil die Funktionalitäten zur Darstellung über QGIS zur Anwendung (QGIS DEVELOPMENT TEAM 2019). QGIS hat den großen Vorteil, dass es komplett Open Source ist und, dass es eine Vielzahl an Erweiterungen gibt, die den Funktionsumfang erheblich vergrößern.

4.4.2 Gephi

Die Visualisierung von Netzwerken in der Form von Edges und Vertices ist im Rahmen dieser Arbeit ebenfalls ein gewichtiger Punkt. Diese Visualisierung wird mit Hilfe von Gephi durchgeführt (BASTIAN et al. 2009). Gephi ist Open Source und aktuell in der Version 0.9.2 verfügbar (Stand: 06.02.2019). Mit dem Werkzeug ist man in der Lage, Beziehungen zwischen Elementen sowie Netzwerken, seien es biologische als auch soziale, zu visualisieren und deren Parameter zu berechnen. Zudem ist es gut dokumentiert und durch seinen offenen Charakter existieren einige Plugins, um den Funktionsumfang zu erweitern.

5 Das Projekt KOGGE und die Hansestadt Rostock als Untersuchungsgebiet

„Die Landschaft erobert man mit den Schuhsohlen, nicht mit den Autoreifen."

Georges Duhamel (1884 - 1966), französischer Humanist

5.1 Das Projekt KOGGE

KOGGE[36] ist ein vom BMBF (Bundesministerium für Bildung und Forschung) gefördertes Projekt. In diesem Projekt stehen aus Sicht der Geoinformatik die Erstellung einer modernen webbasierten Geodateninfrastruktur für die Projektpartner und die webbasierte Öffentlichkeitsbeteiligung im Fokus. Beteiligte an dem Projekt sind die Universität Rostock (Professur für Wasserwirtschaft, Professur für Hydrologie und Angewandte Meteorologie, Professur für Geodäsie und Geoinformatik), die EURAWASSER Nord GmbH, der Wasser- und Bodenverband „Untere Warnow-Küste" sowie biota - Institut für ökologische Forschung und Planung GmbH. Diese kooperieren mit der Hansestadt Rostock, dem Senatsbereich Bau und Umwelt, dem Warnow-Wasser- und Abwasserverband, dem Staatlichen Amt für Landwirtschaft und Umwelt Mittleres Mecklenburg und dem Landesamt für Umwelt, Naturschutz und Geologie Mecklenburg-Vorpommern. Somit sind alle Stakeholder der Wasserwirtschaft in der Region beteiligt. Da im Stadtgebiet Rostock immer wieder Hochwasserereignisse auftreten, wurde für die Stadt bereits ein integriertes Entwässerungskonzept (INTEK) entwickelt (MEHL et al. 2015). An dieses schließt KOGGE an, um die Kleinst- und Kleingewässer hinsichtlich der Einhaltung der WRRL und des Hochwasserschutzes weiter zu untersuchen. Im Ergebnis soll das Entwässerungskonzept für die Hansestadt weiter verbessert werden (KOGGE 2015).

Bedingt durch die zahlreichen Stakeholder ist es notwendig, den Datenaustausch und die Dokumentation zu vereinfachen. Hierfür bietet sich eine webbasierte Geodateninfrastruktur (GDI) an, durch welche Metadaten, Geodaten, Datendienste, Standards und Zugriffsregelungen organisiert werden können (GDI-DE 2015). Des Weiteren soll die Webanbindung des Projektes der Bürgerbeteiligung dienen. Moderne Geodatenportale kommen hier bereits vielfach zur Anwendung (GDI-DE 2015).

Die Öffentlichkeitsarbeit stellt einen der Schwerpunkte des Projektes dar. So zielt die GDI des Projektes darauf ab, Projektergebnisse und Daten sowohl der Öffentlichkeit als auch Akteuren aus Forschung und Wirtschaft zur Verfügung zu stellen. Des Weiteren fand bereits eine *Online-Umfrage*[37] zur Zahlungsbereitschaft der Bürger in Rostock statt, um einen finanziellen Rahmen eventueller wasserbaulicher Maßnahmen abschätzen zu können (MEHL et al. 2017). KOGGE stellt damit den Rahmen sowie die Motivation der vorliegenden Arbeit dar. Im konkreten Fall soll aus Social Media Beiträgen und Nachrichten ein konkreter Mehrwert für die Stadtplanung in der Hansestadt Rostock geschaffen werden. Grundlegend geschieht das Sammeln der Nachrichten auch in Bezug auf wasserwirtschaftliche Themen. Allerdings soll gezeigt werden, dass sich mit diesen Daten und Konzepten ein deutlich breiteres Anwendungsfeld erschließen lässt.

[36] https://www.kogge.auf.uni-rostock.de/
[37] www.umfrage-kogge.uni-rostock.de

5.2 Die Hansestadt Rostock

Die Hansestadt Rostock ist mit knapp 210 000 Einwohnern die größte Stadt Mecklenburg-Vorpommerns. Erwähnt wurde sie zum ersten Mal im Jahr 1160 und hat seither eine wechselhafte Geschichte durchlebt. Vor allem mit der Gründung der Hanse 1358 und dem Beitritt zu dieser erlebte die Stadt ihre Blütezeit und erreichte überregionale Bedeutung. Auch in der jüngeren Geschichte besaß Rostock eine große Bedeutung. So war es zwischen 1945 und 1990 das „Tor zur Welt“ der ehemaligen DDR. Dies spiegelt sich in dem rasanten Bevölkerungsanstieg auf über 250 000 Einwohner wieder (HANSESTADT ROSTOCK 2018). Rostock ist an der Warnow gelegen, welche im Stadtteil Warnemünde in die Ostsee mündet. Damit stellt Rostock eine der wenigen Großstädte in Deutschland dar, die direkt am Meer liegen, was einen nicht zu unterschätzenden Einfluss auf die touristische Attraktivität der Stadt hat. Durch seine Lage gilt Rostock als Regiopole, da die Stadt ein Bindeglied zwischen verschiedenen Metropolen im Ostseeraum ist (Abbildung 5-1). Insbesondere durch ihren Hafen ist die Stadt international eingebunden (HANSESTADT ROSTOCK 2013).

Im Folgenden soll detaillierter auf die geografischen und soziokulturellen Gegebenheiten (Kapitel 5.3) sowie auf die Bevölkerungsstruktur (Kapitel 5.4) Rostocks eingegangen werden, da anzunehmen ist, dass dies entscheidende Parameter für die zu erwartende Nachrichtendichte sind. Zudem soll die wirtschaftliche Bedeutung beleuchtet werden, da hier insbesondere der Werbeaspekt der Nachrichten von Interesse ist.

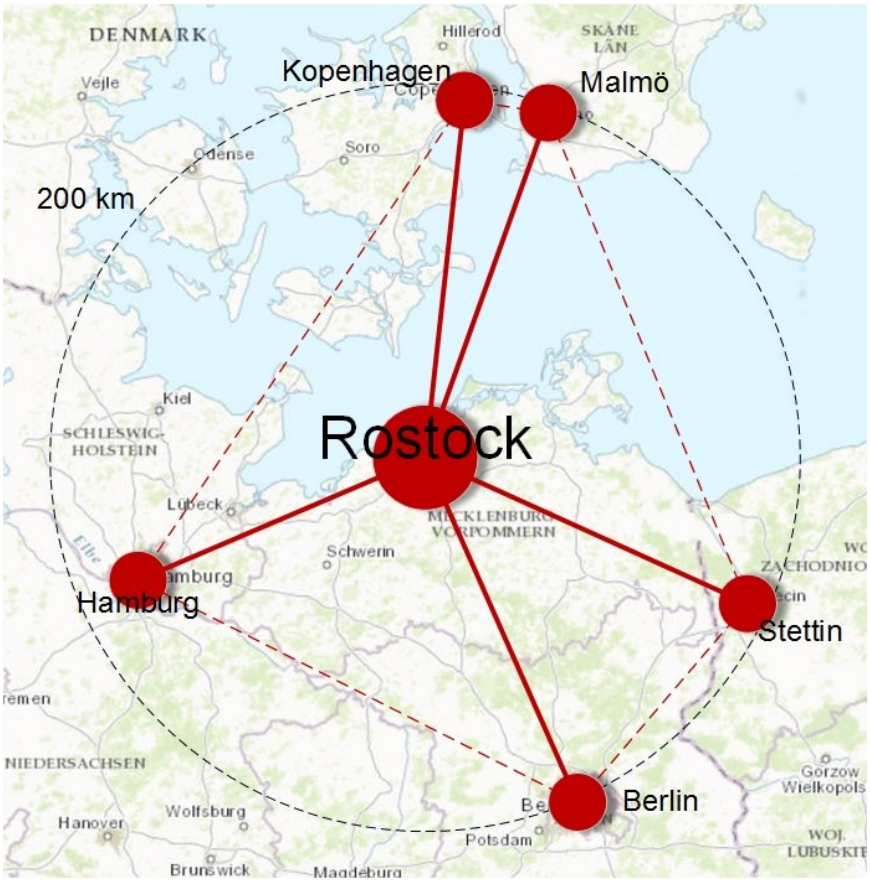


Abbildung 5-1: Regiopole Rostock (eigene Erstellung, nach HANSESTADT ROSTOCK 2013).

5.3 Geografische, wirtschaftliche und soziokulturelle Hotspots

Wirtschaftlich positioniert sich Rostock als Schnittstelle nach Nord- und Osteuropa, bedingt durch seine gute infrastrukturelle Anbindung (Überseehafen, Autobahnen A20 und A19, Flughafen Rostock-Laage). Vor allem die maritime Wirtschaft bildet dabei das Rückgrat. Dazu zählt insbesondere die Reederei AIDA sowie der gesamte Kreuzfahrttourismus (HANSESTADT ROSTOCK 2013).

Rostock gilt als touristisch attraktiv, insbesondere die direkte Lage an der Ostsee macht Rostock zu einem begehrten Reiseziel. So zählt die Stadt 2017 knapp 2 Mio. Übernachtungen. Die durchschnittliche Verweildauer beträgt pro Person 2.6 Tage (Rostock gesamt) bzw. drei Tage (Warnemünde) (HANSESTADT ROSTOCK 2018). Der Hauptteil der Übernachtungen von 65 % entfällt dabei auf die Monate von Mai bis Oktober. Dazu kommen etwa 10 Mio. Tagesausflügler und Geschäftsreisende (Stand: 2010). Insgesamt werden durch den Tourismus in Rostock knapp 500 Mio. Euro umgesetzt (HANSESTADT-ROSTOCK 2012). Herausragende Ziele sind v. a. im Sommer die Strandpromenade in Warnemünde, die Hotels Neptun, A-ja, Hübner und das Strandresort Hohe Düne, sowie der Teepott und der Warnemünder Leuchtturm, die als Wahrzeichen der Stadt gelten (Abbildung 5-2). Ebenfalls in Warnemünde ist das Kreuzfahrtterminal angesiedelt. Mit 190 an- und ablegenden Kreuzfahrtschiffen und 383 000 Personen im Jahr 2017 kann dieser durchaus als touristischer Hotspot in der Hansestadt gelten (HANSESTADT ROSTOCK 2018). Gleiches gilt für den Fährterminal im Überseehafen, von wo aus diverse Ziele an der Ostseeküste mehrfach täglich angefahren werden. 2017 sind durch die Fähren 2.1 Mio. Personen befördert worden (HANSESTADT ROSTOCK 2018). Dieser ist ebenfalls ein wichtiger Bestandteil der innereuropäischen Infrastruktur (HANSESTADT ROSTOCK 2013).

Abseits von Warnemünde sind die Altstadt und die Stadtmitte Rostocks zu nennen, welche ebenfalls zahlreiche Besucher mit ihren historischen Bauten, den Stadttoren (Steintor, Kröpeliner Tor) und den großen Kirchen (Marienkirche, Petrikirche) anziehen (Abbildung 5-2). Das abendliche kulturelle Leben konzentriert sich insbesondere im Szeneviertel Kröpeliner-Tor-Vorstadt (KTV). Hier bieten die zahlreichen Restaurants und Bars den abendlichen Anlaufpunkt vieler Menschen. Im Stadthafen Rostocks lässt sich zudem vor allem im Sommer viel maritimes Flair erleben. So bieten hier das Schiff „Stephan Jantzen“ und die zahlreichen Restaurants beliebte Anlaufstellen sowohl für Touristen als auch für die Einwohner der Stadt. Neben dem Stadtzentrum sind als potentielle Hotspots kulturellen Lebens noch der Zoo der Stadt sowie das ehemalige Gelände der internationalen Gartenausstellung (IGA), das Traditionsschiff „Dresden“ mit seinem Schifffahrtsmuseum als auch das Messegelände der Stadt zu nennen.

Rostock bietet eine große Vielfalt an regelmäßigen Veranstaltungen. Dazu zählen die Spiele diverser Sportvereine, allen voran die des FC Hansa Rostock im 29 000 Zuschauer fassenden Ostseestadion. Aber auch Theaterveranstaltungen sowie Konzerte auf diversen Bühnen und Clubs der Stadt sind hier zu nennen (HANSESTADT ROSTOCK 2013).

Zudem bietet Rostock diverse Großveranstaltungen. Als erstes ist hier die Hanse Sail mit ihren ca. 1 Mio. Besuchern zu nennen (NDR 2016). Sie gilt als eine der größten Festveranstaltungen Deutschlands und unterstreicht das maritime Flair der Stadt Rostock. Im maritimen Bereich sind außerdem die Warnemünder Woche als Segelevent sowie das jährliche Leuchtturmglühen zu nennen. Daneben gibt es zahlreiche weitere Festveranstaltungen. Dazu zählen die Musikfestivals Rostock Rockt und N-Joy The Beach, aber auch Großveranstaltungen im Stadion des FC Hansa Rostock und der Stadthalle. Zuletzt sei noch der Rostocker Weihnachtsmarkt erwähnt (HANSESTADT ROSTOCK 2013).

Es lässt sich also erkennen, dass es eine ganze Reihe diverser Hotspots des kulturellen Lebens in Rostock gibt, die nicht zwingend der Bevölkerungsverteilung der Stadt entsprechen. Daher ist es letztlich interessant zu wissen, ob sich die Nachrichtenverteilung v. a. an den Hotspots oder der Bevölkerungsverteilung orientiert und welche Themen in den einzelnen Gebieten die am häufigsten diskutierten sind.

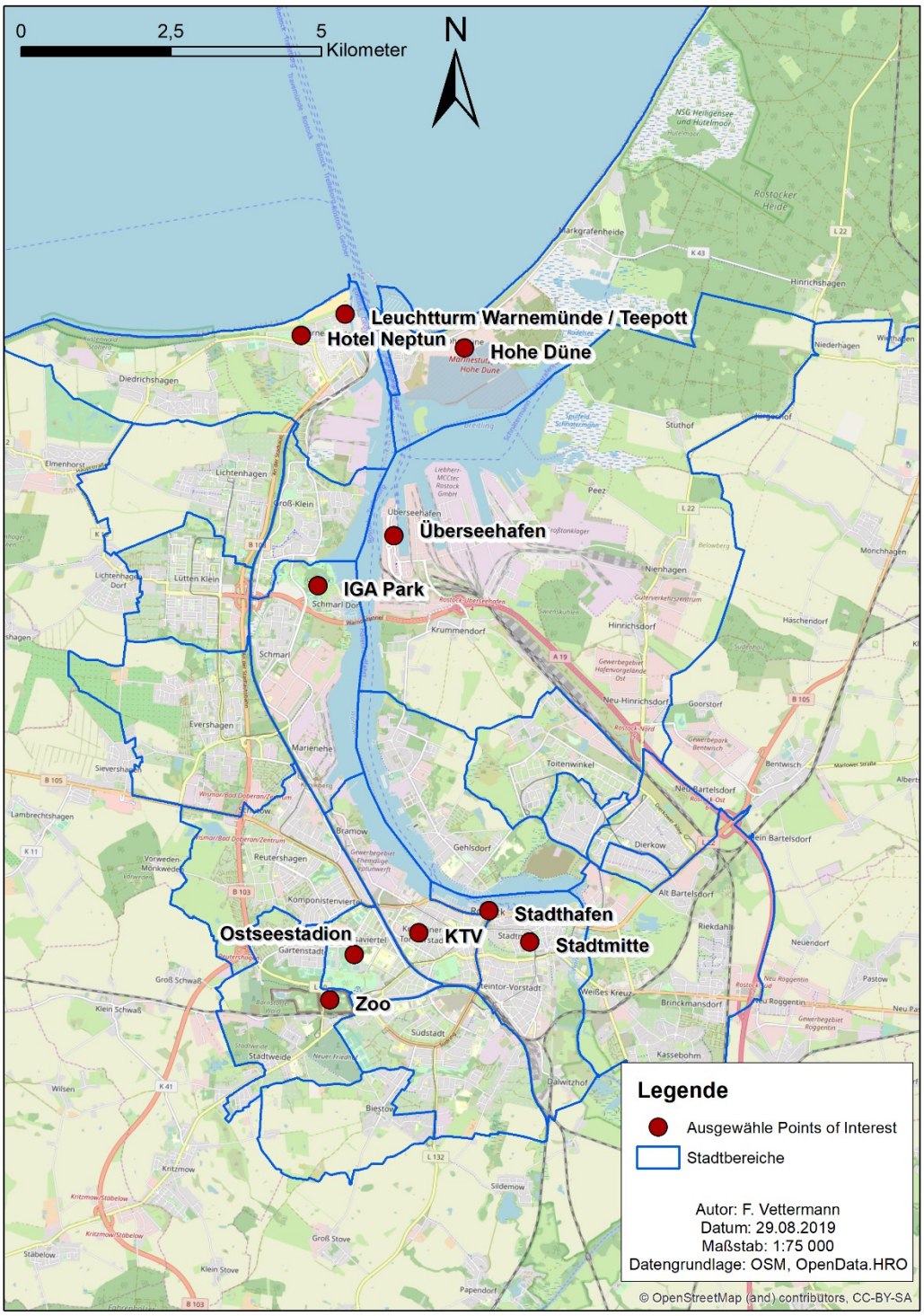


Abbildung 5-2: Überblick über die Stadtbereiche der Hansestadt Rostock mit ausgewählten Points of Interest (POI).

5.4 Bevölkerungsentwicklung und -struktur

Von wesentlichem Interesse ist die Bevölkerungsstruktur Rostocks, da davon auszugehen ist, dass insbesondere die jüngere Generation verstärkt Soziale Medien nutzt (SCHOLZ 2014). Es zeigt sich, dass sich seit dem Tiefpunkt 2002 mit 194 978 Einwohnern die Einwohnerzahl kontinuierlich auf 208 409 Einwohner 2017 erhöht hat (Abbildung 5-3). Bis zum Jahr 2035 wird ein weiterer Bevölkerungsanstieg in Rostock auf nahezu 230 000 Einwohner prognostiziert (HANSESTADT ROSTOCK 2018). Der Anteil der Personen, welche jünger als 45 Jahre sind, folgt verzögert der Entwicklung der Bevölkerungsanzahl und ist von 56 % im Jahr 2000 auf unter 50 % im Jahr 2014 gefallen. Seither ist wieder von einem steigenden Anteil an der Gesamtbevölkerung Rostocks auszugehen. Neben einer gestiegenen Geburtenrate von 1.24 Kindern / Frau (2007) auf 1.40 Kindern / Frau (2014) ist insbesondere das Wanderungssaldo für den Anstieg der Gesamtbevölkerung verantwortlich, da die Zahl der Lebendgeborenen weiterhin niedriger liegt als die Zahl der Todesfälle (HANSESTADT ROSTOCK 2016). Der Ausländeranteil in Rostock liegt bei 4.8 % (Stand 2015, HANSESTADT ROSTOCK 2018).

Von Interesse ist vor allem die Bevölkerungsverteilung in der Stadt Rostock nach Stadtbereichen (Abbildung 5-5). So lassen sich zum einen die bevölkerungsreichsten Stadtteile, aber auch die jüngsten und somit die vermeintlich prosperierenden identifizieren. Gleichzeitig gilt hier natürlich wieder die Annahme, dass, je höher der Anteil der jüngeren Bevölkerung ist, auch die Anzahl der Nachrichten in Sozialen Netzwerken steigt. Diese These wird sich allerdings erst beweisen müssen. Durch die ohnehin schon kleinteilige Analyse soll auf Parameter wie Bildung verzichtet werden.

Es ist ersichtlich, dass vor allem die Szeneviertel KTV sowie die Stadtmitte die am dichtesten besiedelten Stadtbereiche sind (Abbildung 5-5). Allerdings stechen auch die großen Wohngebiete Reutershagen, Groß Klein und Lütten Klein mit 15 - 20 Tsd. Einwohnern heraus. Bis zum Jahr 2025 ist prognostiziert, dass die Einwohnerzahl Lichtenhagens ebenfalls ansteigen wird. Daneben werden v. a. Biestow, Gehlsdorf und die Südstadt einen Einwohnerzuwachs verzeichnen. Grund dafür ist u. a. der geplante neue Stadtteil Groß-Biestow (SCHWERINER VOLKSZEITUNG 2016).

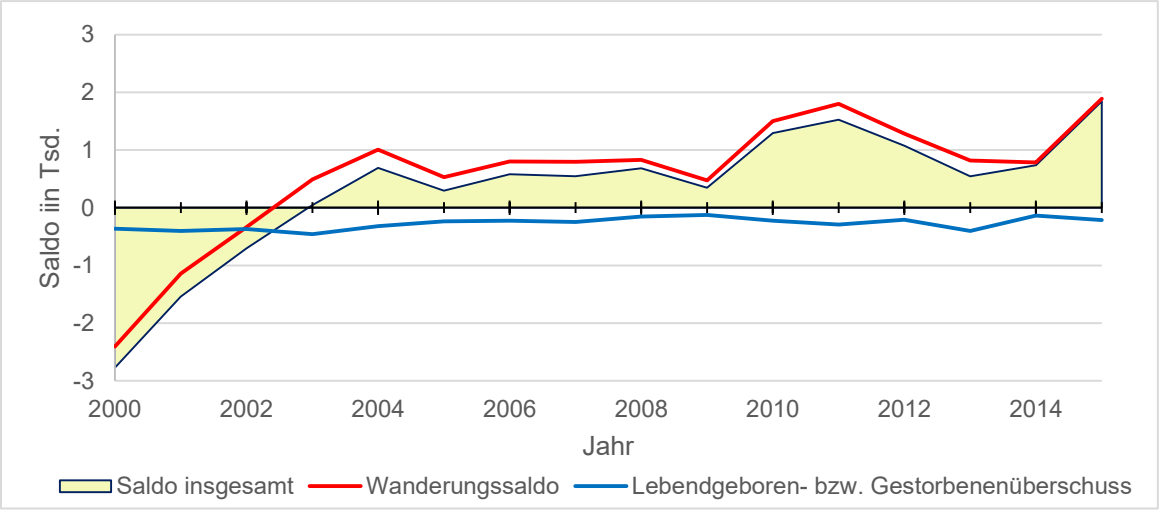


Abbildung 5-3: Bevölkerungssaldo der Hansestadt Rostock (eigene Erstellung, nach HANSESTADT ROSTOCK 2018).

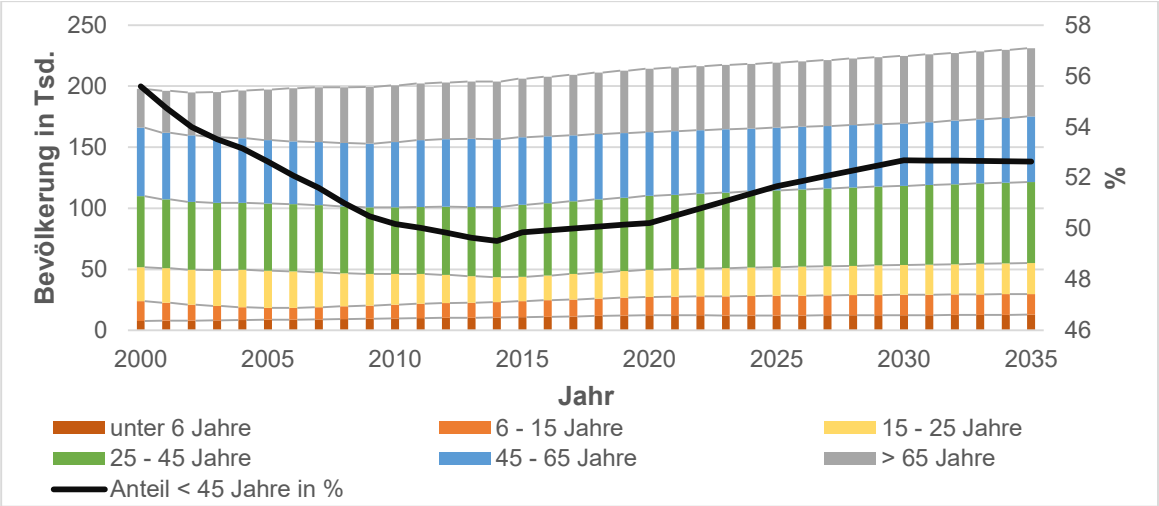


Abbildung 5-4: Aktuelle und prognostizierte Bevölkerungsstruktur der Hansestadt Rostock (eigene Erstellung, nach HANSESTADT ROSTOCK 2018, 2016).

Dies schlägt sich in der Altersstruktur der Bevölkerung nieder. So wird der Stadtbereich Biestow bis 2025 mit einem Anteil der unter 45-Jährigen von über 90 % der jüngste Stadtteil werden. Die Szeneviertel KTV und Stadtmitte bleiben dabei mit einem Anteil von über 70 % der unter 45-Jährigen weiterhin sehr junge Stadtviertel, obgleich auch hier der Altersschnitt sinken wird. Dierkow West, Dierkow Ost sowie Warnemünde sind und bleiben weiterhin die Stadteile mit der ältesten Bevölkerung. Nichtdestotrotz sorgt das Wachstum der Stadt insgesamt für eine Verjüngung. So soll das Durchschnittsalter von 44.9 Jahren (Stand: 2015) auf 44.3 Jahre (2035) zurückgehen (HANSESTADT ROSTOCK 2016). Die Daten spiegeln damit letztlich die Szeneviertel der Hansestadt wieder (KTV, Stadtmitte), aber sie zeigen auch, dass zukünftig die Analyse von Nachrichten in Sozialen Netzwerken einen größeren Bevölkerungsteil abbilden kann, da die Grundgesamtheit wächst und gleichzeitig das Alter der Bevölkerung sinkt.

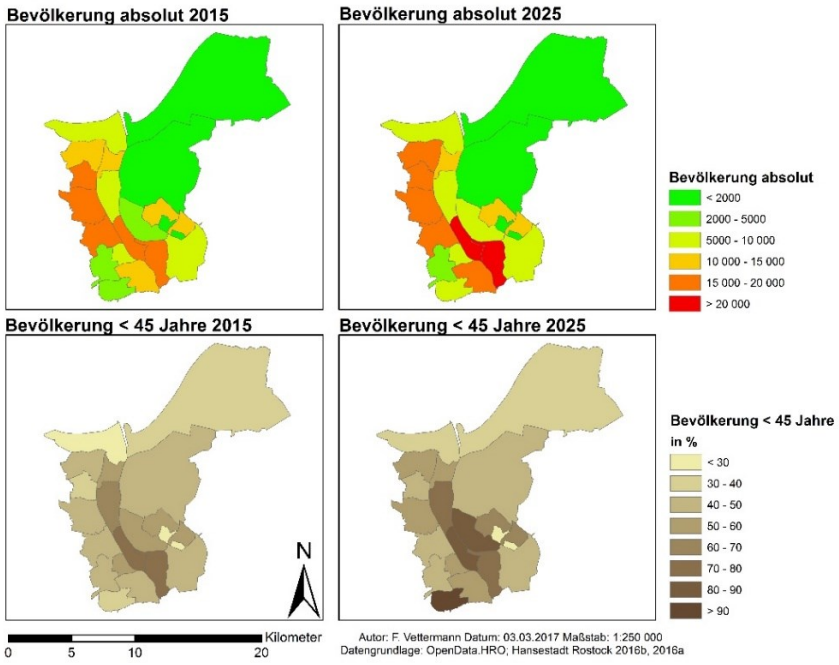


Abbildung 5-5: Bevölkerungsentwicklung der Hansestadt Rostock bis 2025.

5.5 Bürgerbeteiligung und -information in Rostock

In Rostock sind bezüglich der Bürgerbeteiligung insbesondere die 19 Ortsbeiräte zu nennen. Diese widmen sich den jeweiligen Anliegen in den Stadtteilen und können hierzu Empfehlungen und Anträge an die Bürgerschaft einreichen. Daneben stellen sie eine der wesentlichen Kontaktstellen für Bürgeranliegen dar (DIE GRÜNEN 2016). Daneben kommen weitere Methoden zur Bürgerbeteiligung wie kommunale Bürgerforen, Webplattformen sowie Informationsveranstaltungen zur Anwendung. Ein gutes Beispiel ist hierfür die Rostocker Plattform zur Beteiligung an der Bebauungsplanung des Werftdreiecks[38]. Um zu einem definierten Verfahren bei der Bürgerbeteiligung zu kommen, wird aktuell ein Leitfaden für die Hansestadt ausgearbeitet (HANSESTADT ROSTOCK 2017).

Daneben haben sich weitere Verfahren der Bürgerbeteiligung sowie für VGI in der Hansestadt etabliert. Dazu zählt zum einen die Plattform Klarschiff.HRO, zum anderen die Dateninfrastruktur der Hansestadt - *OpenData.HRO*[39]. Dem übergeordnet ist die neueste Plattform der Hansestadt, der *Geolotse.HRO*[40], der Zugang zu allen raumbezogenen Daten Rostocks bieten soll.

5.5.1 Klarschiff.HRO

Klarschiff.HRO stellt eine im März 2012 gestartete Online-Bürgerbeteiligungsplattform in der Hansestadt Rostock dar. Es handelt sich hierbei um eine Serviceplattform, mit deren Hilfe die Bürger in der Lage sind, „[…] *die Verwaltung auf Probleme in der Stadtentwicklung und der öffentlichen Sicherheit und Ordnung hinzuweisen.*“ (NEITZ et al. 2010). Die Plattform bietet neben der Visualisierung auf dem herkömmlichen Desktop auch den Zugriff mittels Smartphone. Ziel ist es, das Engagement der Bevölkerung für ihre Stadt zu steigern, jüngere Menschen aktiv einzubeziehen und die Reaktionszeit auf bestimmte Schäden deutlich zu reduzieren.

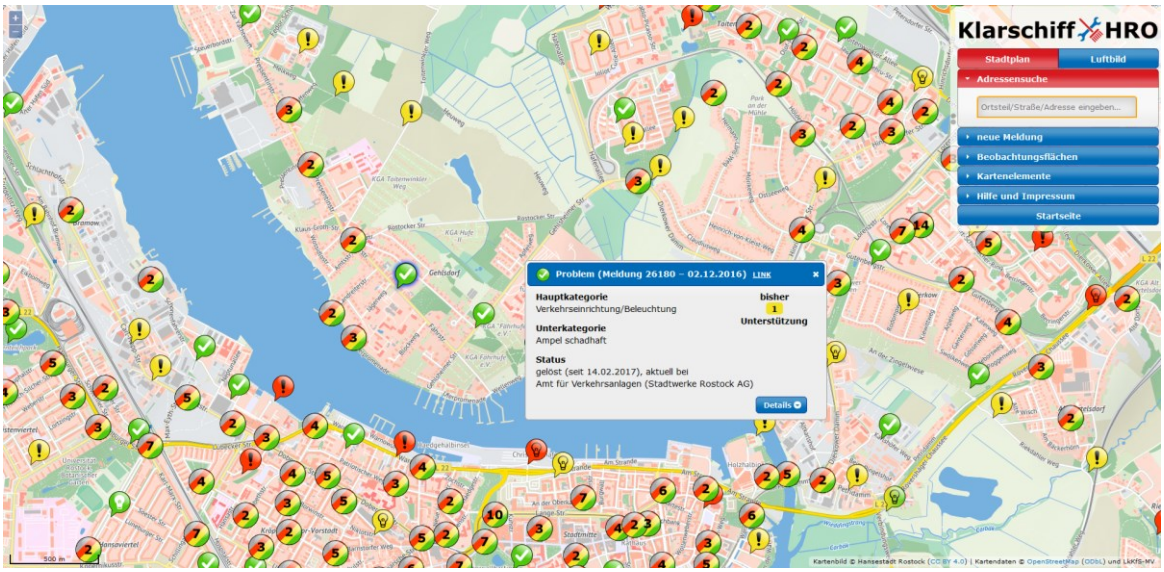


Abbildung 5-6: Web-Applikation von Klarschiff.HRO.

[38] http://werftdreieck-rostock.de/
[39] [39] https://www.opendata-hro.de/
[40] https://geo.sv.rostock.de/geolotse/de

Ein großer Vorteil der Plattform ist es, dass sie auf die vom OGC definierten Standards setzt und komplett frei zugänglich ist. Das Portal basiert auf OpenLayers sowie MapBender. Des Weiteren lassen sich die Meldungen über eine eigens entworfene API auch in andere GDIs einbinden. Während des Entwicklungsprozesses wurden bereits alle wichtigen Stellen mit eingebunden (Abbildung 5-7). Dadurch sind die Zuständigkeiten geklärt und die Meldungen gelangen direkt zum jeweils zuständigen Bearbeiter (NEITZ et al. 2010).

Dass das Portal sehr gut angenommen wird, wird an der Zahl der Meldungen und Aufrufe ersichtlich. So sind aktuell (Stand 06.02.2019) über 37 000 Meldungen eingegangen, davon allein 1608 Meldungen in den vergangenen zwölf Monaten. Von diesen wiederum sind 612 gelöst und 568 aktuell in Bearbeitung. Allein diese Zahlen zeigen deutlich, dass Online-Bürgerbeteiligung sehr gut funktionieren kann. Zudem spart sie bares Geld für die Stadt, denn durch die schnelle Meldung und Bearbeitung können Folgeschäden vermieden und auf eine Ortsbesichtigung kann häufig verzichtet werden (KOMMUNE21 2014).

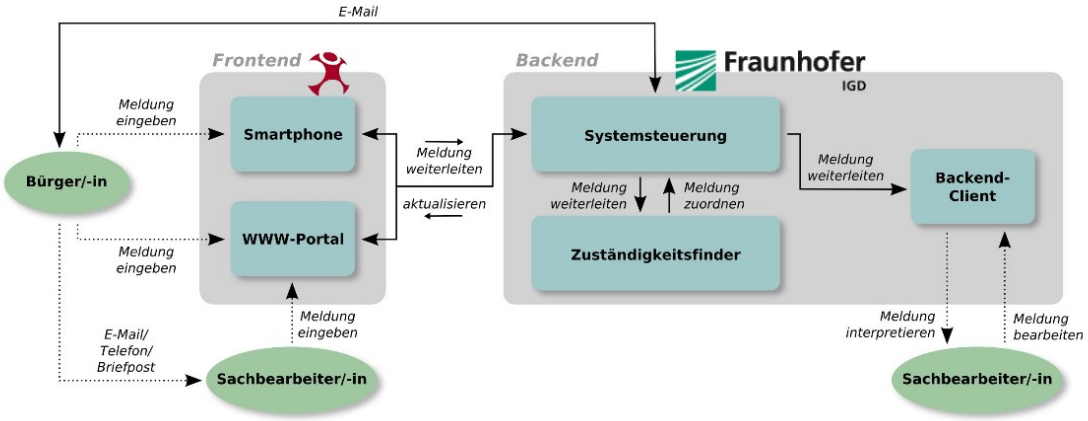


Abbildung 5-7: Prozessabläufe in Klarschiff.HRO.

5.5.2 OpenData.HRO

OpenData.HRO stellt eine offene Datenplattform der Hansestadt Rostock dar. Hierbei werden kommunale Daten unter Creative Commons (CC) Lizenz abgelegt und für die Nutzer zugänglich gemacht. OpenData.HRO basiert dabei auf der Open-Source-Software *CKAN*[41]. Es bietet eine Weboberfläche, über die Datensätze gesucht, Metadaten angezeigt oder Daten hinzugefügt werden können. Auch der Zugriff über eine API ist möglich (NEITZ et al. 2013). Damit zählt das Portal weniger zur Bürgerbeteiligung, sondern dient der Bürgerinformation.

Der in der Datenplattform abgelegte Datenkatalog enthält statistische Daten sowie zahlreiche Geodaten auf deren Basis neue, auch kommerzielle Produkte, hergestellt werden können. Die Datensätze werden in verschiedenen Formaten, darunter GeoJSON, KML oder Shape zur Verfügung gestellt. Zudem werden WMS oder WFS angeboten (NEITZ et al. 2013).

[41] http://ckan.org/about/

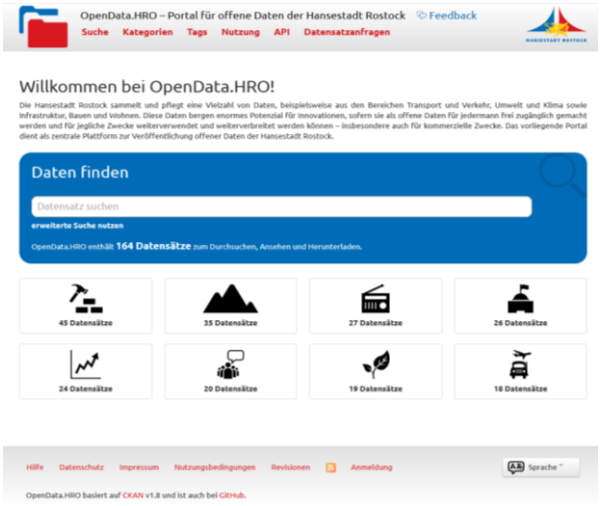


Abbildung 5-8: Web-Interface von OpenData.HRO.

5.5.3 Zahlungsbereitschaftsanalyse zu den Klein- und Kleinstgewässern

Neben den vorgestellten Portalen der Hansestadt zur Bürgerbeteiligung und -information können auch Umfragen einen wertvollen Beitrag leisten. Gerade die geringen Hürden und die i. d. R. einfache Handhabe von Online-Umfragen ermöglicht es, die Meinung der Bevölkerung zu spezifischen Themen mit einem relativ geringen Aufwand zu erfassen (THIELSCH & WELTZIN 2012).

Zwar weisen Online-Umfragen spezifische Nachteile, wie eine Vor-Selektion der Teilnehmer aufgrund der Beschränkung auf Personen mit Internetzugang, auf, jedoch überwiegen die Vorteile die Nachteile deutlich. Aus diesem Grund lief im Rahmen des Projektes KOGGE eine Online-Umfrage mit dem Titel „Was sind Ihnen unsere kleinen städtischen Gewässer und Feuchtgebiete wert?“ für die Dauer von ca. einem Jahr. Ziel war es, die potentielle Zahlungsbereitschaft der Rostocker Bevölkerung zu erfassen und damit den Wert zu ermitteln, den die Rostocker ihren Gewässern beimessen (MEHL et al. 2017).

5.5.3.1 Zielstellungen und Aufbau

Die Umfrage ist auf Basis von Limesurvey erstellt worden, einer Open Source Softwareumgebung, mit der sich komplexe Umfragen, auch mit räumlichen Karten- und Matrixabfragen, erstellen lassen (SCHMITZ 2015). Durch den Open Source-Charakter der Software ist es möglich, die gesamte Umfrage entsprechend der eigenen Anforderungen anzupassen. Die Antworten werden in einer PostgreSQL-Datenbank gespeichert. Die Umfrage selbst lief auf einem SSL-gesicherten Server.

Strukturell ist die Umfrage in sechs Bereiche gegliedert, wobei drei die jeweiligen Fragen beinhalten (Abbildung 5-9). Der erste Teil beinhaltet die Erfassung der personenbezogenen allgemeinen Informationen, der zweite die Erfassung der persönlichen Einschätzungen und Präferenzen bei Feuchtgebieten und Gewässern in Rostock und schließlich folgt der letzte Teil mit der Abfrage der Zahlungsbereitschaft. Letztere stellt den Kern der Umfrage dar, d. h., der Nutzer wird langsam mit Zusatzinformationen ausgestattet und ist gezwungen, sich mit der Thematik zu beschäftigen, bevor er seine Zahlungsbereitschaft angeben soll. Die Dauer der Umfrage beträgt dabei etwa zehn Minuten, wodurch sie eine große Zahl an Nutzern erreichen soll, ohne, dass diese bedingt durch die Länge entnervt die Umfrage abbrechen.

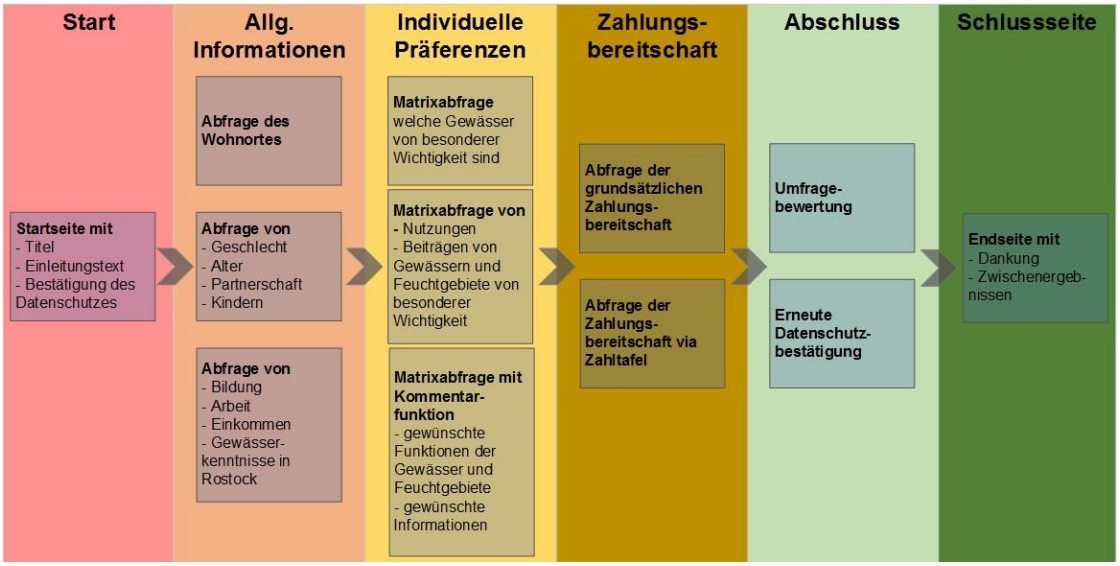


Abbildung 5-9: Umfragestruktur der KOGGE Online-Befragung.

5.5.3.2 Ergebnisse

Nach ca. einem Jahr Laufzeit haben sich 1 101 Befragte vollständig bzw. auswertbar an der Umfrage beteiligt. Bedingt durch die Verteilungswege der Online-Umfrage konnte keine repräsentative Grundgesamtheit erreicht werden. Auch eine entsprechende nachträgliche Gewichtung empfahl sich nicht, da manche Teile der Bevölkerung der Hansestadt völlig unzureichend erfasst wurden. So lag das Gros der Befragten bei Akademikern und Studenten sowie Angestellten im öffentlichen Dienst. Dennoch lassen sich einige interessante Aspekte aus der Umfrage ableiten, insbesondere da sich die Ergebnisse weitgehend mit denen vorangegangener Untersuchungen hinsichtlich der Ökosystemdienstleistungen decken (MEHL et al. 2017). Insofern sind die Ergebnisse differenziert zu betrachten, bringen aber durchaus einen Mehrwert.

So treten einige interessante Aspekte zu tage. Den Rostockern ist die hohe ökologische Funktionsfähigkeit der Gewässer sowie Naturnähe und Biodiversität besonders wichtig. Demgegenüber wird Wasser als Gestaltungselement als unwichtigster Aspekt betrachtet, auch die Regulierungsleistung der Moore im Hinblick auf die Treibhausgasbindung/-freisetzung wird offenkundig als zu abstrakt und damit relativ unwichtig angesehen. Die präferierten Nutzungen der Befragten entfallen auf die Kategorien „Naherholung“, „Naturerleben“, „Baden/Schwimmen“ sowie „Abenteuerspielplatz“ (MEHL et al. 2017). Interessant ist der Vergleich der Präferenzen mit der Naturbewusstseinsstudie 2015 (BMUB & BFN 2016). Bei dieser zeigte sich ebenfalls, dass sich eine Mehrzahl der Befragten (94 %) für eine gute Zugänglichkeit von Natur im urbanen Umfeld ausspricht. Zwei von drei Befragten sprechen sich zudem für innerstädtische Sukzessionsflächen aus.

Immerhin 62 % der Befragten haben im Rahmen der Umfrage ihre grundsätzliche Zahlungsbereitschaft erklärt. Die mittlere Zahlungsbereitschaft je Monat und Haushalt über alle Befragten beträgt danach 5.27 € (63,24 € pro Jahr). Extrapoliert man diese Angaben würde man bei 210 146 Einwohnern bzw. 118 406 Haushalten bei etwa 7.5 Mio. € pro Jahr landen, die die Rostocker ihren Gewässern an Wert beimessen (HANSESTADT ROSTOCK 2015b). Diese Werte bewegen sich im Rahmen anderer aktueller repräsentativer Umfragen (vgl. MEHL et al. 2017, Tabelle 5-1).

Tabelle 5-1: Geäußerte Zahlungsbereitschaften (ZB) für naturnahe Fließgewässer und Feuchtgebiete anhand ausgewählter Quellen (MEHL et al. 2017).

Thema [Literaturquelle]	ZB pro Haushalt und Jahr	Literatur	Bemerkungen
Schutz der biologischen Vielfalt durch Auenrenaturierung an der Elbe (bzw. an Rhein und Weser)	5.00 €…15.00 €	MEYERHOFF 2002	
Mehrwert naturnaher Fließgewässer	172 CHF (ca. 185 €)	ARNOLD et al. 2009	Durchschnitt von 4 Varianten in der Schweiz; durchschnittliche Haushaltsgröße angesetzt mit 2 Personen
Renaturierung von Gewässern und Auen	109.44 €	MEYERHOFF et al. 2012	
Gewässerzustandsverbesserung (gute Qualität nach WRRL) in der Region Berlin/Brandenburg	154.00 €	MEYERHOFF et al. 2014	792 nach repräsentativen Kriterien ausgewählte Teilnehmer, die ZB ist in Berlin ca. doppelt so hoch wie im Brandenburger Umland
Leitbild „hohe Biodiversität“ in einer Flussauenlandschaft	53.39 €	HORBAT et al. 2016	Repräsentative telefonische Umfrage in Nordostdeutschland
Leitbild „naturnahe Auenlandschaft“	61.37 €	HORBAT et al. 2016	Repräsentative telefonische Umfrage in Nordostdeutschland
Umsetzung der Ziele der WRRL für kleine urbane Gewässer und Feuchtgebiete	63.24 €	MEHL et al. 2017	KOGGE-Online-Umfrage, 80 % Teilnehmer aus der Hansestadt Rostock

6 Daten und Methodik

„Ist dies schon Wahnsinn, so hat es doch Methode.“

William Shakespeare (1564 - 1616), englischer Dichter

6.1 Aufbau einer GDI

Die GDI ist im Rahmen des Projektes KOGGE aufgebaut wurden, um hier einen Datenaustausch, aber auch die Analyse und Visualisierung von Daten online zu ermöglichen. Um die problemlose Datenverwaltung und den Datenaustausch zu gewährleisten, sind die folgenden Anforderungen an das Portal zu stellen (GDI-DE 2015):

1. Formatunabhängige Datenverfügbarkeit für jeden überall
2. Benutzer- und Rechteverwaltung
3. Einsicht in Metainformationen
4. Nachvollziehbarkeit von Änderungen
5. Auffindbarkeit der Datensätze
6. Neuintegration von Datensätzen
7. Visualisierung und Bearbeitung

Diesen Anforderungen genügt im Wesentlichen die Open Source Software GeoNetwork. Zudem ist sie komplett den eigenen Bedürfnissen und Wünschen anpassbar. Vor allem durch die Möglichkeit, Geodaten zu visualisieren, empfiehlt sie sich im Rahmen dieser Arbeit angewendet zu werden. Zudem muss so kein neues Portal aufgebaut werden, sondern es können bestehende Strukturen genutzt werden. Nachfolgend sollen kurz die Struktur des Portals und seine Besonderheiten beschrieben werden.

6.1.1 GeoNetwork

Die eigene GeoNetwork-Integration wurde auf einem KOGGE-eigenen Server aufgesetzt und entsprechend den projektinternen Anforderungen gestaltet. Dazu wurden sowohl ein Hilfebereich, ein Kontaktbereich sowie ein Impressumsbereich hinzugefügt. Außerdem wurde das gesamte Design des Portals überarbeitet (HÜBNER & VETTERMANN 2016, HÜBNER et al. 2016). Zudem sind eine ganze Reihe optische und funktionelle Anpassungen in die GDI eingeflossen.

Eine Besonderheit des Portals ist, dass das Frontend der Datenstruktur mit Hilfe von Thesauri gestaltet worden ist (vgl. Kapitel 4.3.2; Abbildung 6-1). Es ist für jeden Daten-Teilbereich des Portals ein eigenes Wortnetz angelegt worden, welches alle wesentlichen Schlagwörter enthält und in eine hierarchische Struktur bringt (Abbildung 6-2). Ziel war es, die Hierarchie so flach wie möglich zu gestalten, um die gesuchten Elemente schnellstmöglich auffindbar zu machen. Anschließend sind jedem Daten- bzw. Metadatensatz ein oder mehrere entsprechende Schlagwörter zugeordnet worden. Nachteilig ist, dass bei neuen Schlagwörtern die Thesauri manuell ergänzt werden müssen.

Ein besonderer Augenmerk bei der Erstellung lag auf der Vereinfachung der Dateneingabe, um die Hemmschwelle bei allen Projektbeteiligten zu senken, Daten in das Portal aufzunehmen (HÜBNER et al. 2016). Daher sind für alle wichtigen Thematiken Templates erstellt worden, bei denen bereits das Gros aller Angaben ISO-konform ausgefüllt sind. Lediglich Angaben über das Datum oder auch den hochladenden Nutzer müssen manuell ergänzt werden. Daneben ist die Darstellung der Metadaten selbst in Tabs organisiert und

eine ISO-Validierung implementiert worden. So lassen sich nur noch korrekt beigetragene Datensätze auch tatsächlich im Portal speichern.

Des Weiteren ist die Arbeit von HÜBNER (2016) in das Portal eingeflossen. So ist es erst einmal möglich, dass WPS sich in GeoNetwork nutzen lassen (HÜBNER et al. 2016). Dies hat vor allem für die Darstellung der Nachrichten aus den Sozialen Netzwerken Auswirkungen. So ist es nicht notwendig, diese erst mühsam aufzubereiten, sondern dieser Arbeitsschritt kann direkt im Portal durchgeführt werden, wodurch die Daten sich in der gewünschten Form anzeigen lassen.

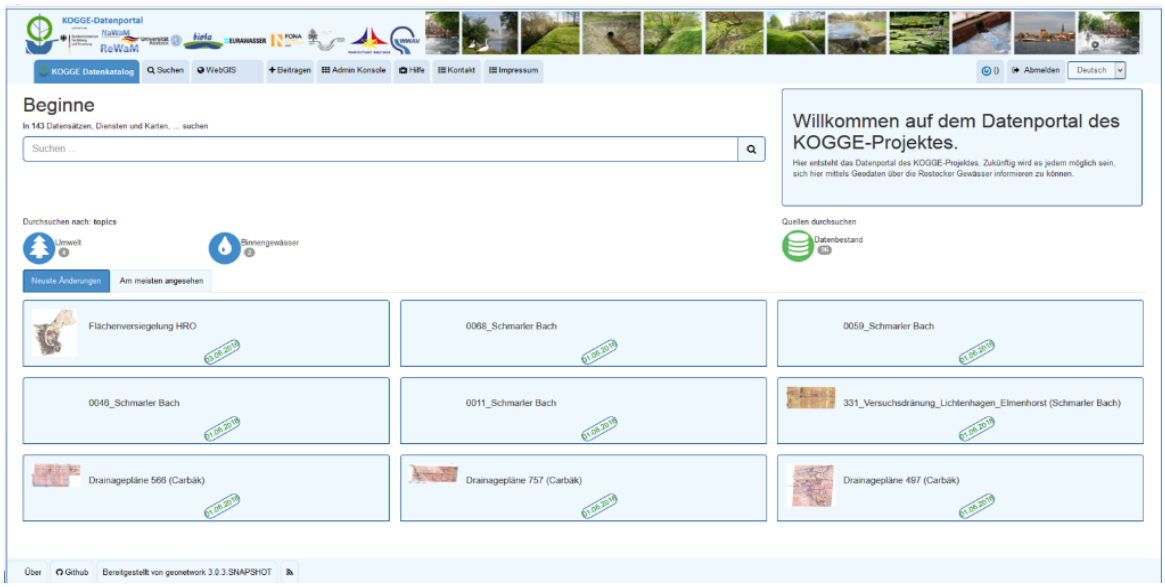


Abbildung 6-1: Screenshot vom Frontend des Datenportals.

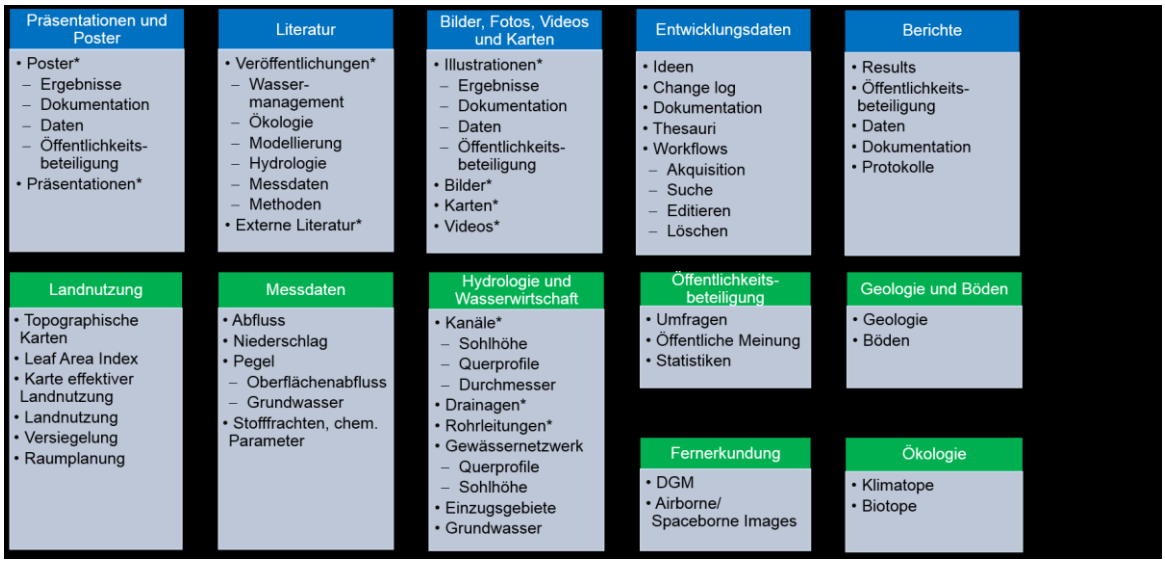


Abbildung 6-2: Grundlegende Datenstruktur des Datenportals auf Basis von Thesauri.

6.1.2 GeoServer

Da GeoServer direkt in GeoNetwork eingebettet ist, wird dieser schließlich zur Visualisierung der einzelnen Datensätze verwendet. An GeoServer selbst wurden im Rahmen der Erstellung des Portals keine größeren Änderungen vorgenommen. Lediglich die WPS-Funktionalitäten sowie einige Projektionen sind ergänzt worden (HÜBNER 2016). Zudem erfolgte für die darzustellenden Datensätze die manuelle Erstellung und Einbindung von SLDs. Ansonsten wurde auf die bereits durch GeoNetwork erstellte Grundstruktur zurückgegriffen.

6.2 Social Media Harvesting

Das Harvesten der Daten aus den Sozialen Netzen stellt den ersten und wichtigsten Schritt für die in dieser Arbeit durchgeführte Analyse dar. Da für Python von allen gewünschten Netzwerken entsprechende Bibliotheken für den API-Zugriff zur Verfügung stehen, kommt diese Programmiersprache bei der Umsetzung zum Zug (vgl. Kapitel 4.1.2).

Neben dem reinen Harvesting ist vor allem die Live Integration von großer Wichtigkeit. Daher sind die meisten Funktionen ebenfalls direkt in Python eingebunden, um beispielsweise die Texte zu korrigieren oder das Gazetteer-Matching durchzuführen. Diese Funktionen benötigen eine hohe Performanz, damit es nicht zu einem Nachrichtenstau kommt und nur noch ein Teil der auflaufenden Nachrichten verarbeitet werden können.

Alle Verfahren sind dabei zuerst auf Basis von Twitter entwickelt worden. Bei Bedarf lassen sich die Algorithmen zur Textprozessierung aber auch auf andere Netzwerke wie z. B. Instagram übertragen.

6.2.1 Datenbankerstellung

Der erste Schritt bei der Umsetzung des Harvester stellt das Aufsetzen einer geeigneten Datenbank dar. Diese muss sowohl den Gazetteer für die Lokalisationsoperationen als auch spezielle Tabellen, beispielsweise für die Expertenfindung oder die Zuweisung von Themen bereitstellen. Als Umgebung wurde PostgreSQL mit der Erweiterung PostGIS verwendet (vgl. Kap. 4.1.1).

In der Datenbank selbst finden keine Funktionen und Skripte Eingang. Die Datentabellen werden ausschließlich aus Python heraus gefüllt, geupdated und abgefragt. Untereinander sind alle Tabellen über den Ortsnamen (Tweets – Gazetteer, Gazetteer – Gazetteer Polygone, Gazetteer – Exception List), die User-ID (Tweets – Follower-List) oder mittels der Thematik (Tweets – Topics) verknüpft.

In Abbildung 6-3 ist dabei die grundsätzliche Struktur der Datentabellen und deren Verbindung untereinander dargestellt. Aus dieser wird ersichtlich, dass die drei Kerntabellen die Tweets als solches, die Follower Liste sowie der Gazetteer darstellen. Alle Operationen werden letztlich auf der Basis dieser drei Tabellen durchgeführt. Grundsätzlich findet anhand der Tweets die Textanalyse statt, anhand des Gazetteers die Ortszuweisung und mittels der Follower-List die Expertenfindung. In den folgenden Kapiteln werden die dazu zugehörigen Funktionen und deren Bedeutung genauer betrachtet.

In einem letzten Schritt werden die Tweets der letzten zwei Wochen jede Minute in eine separate Datentabelle kopiert für die Darstellung über GeoServer.

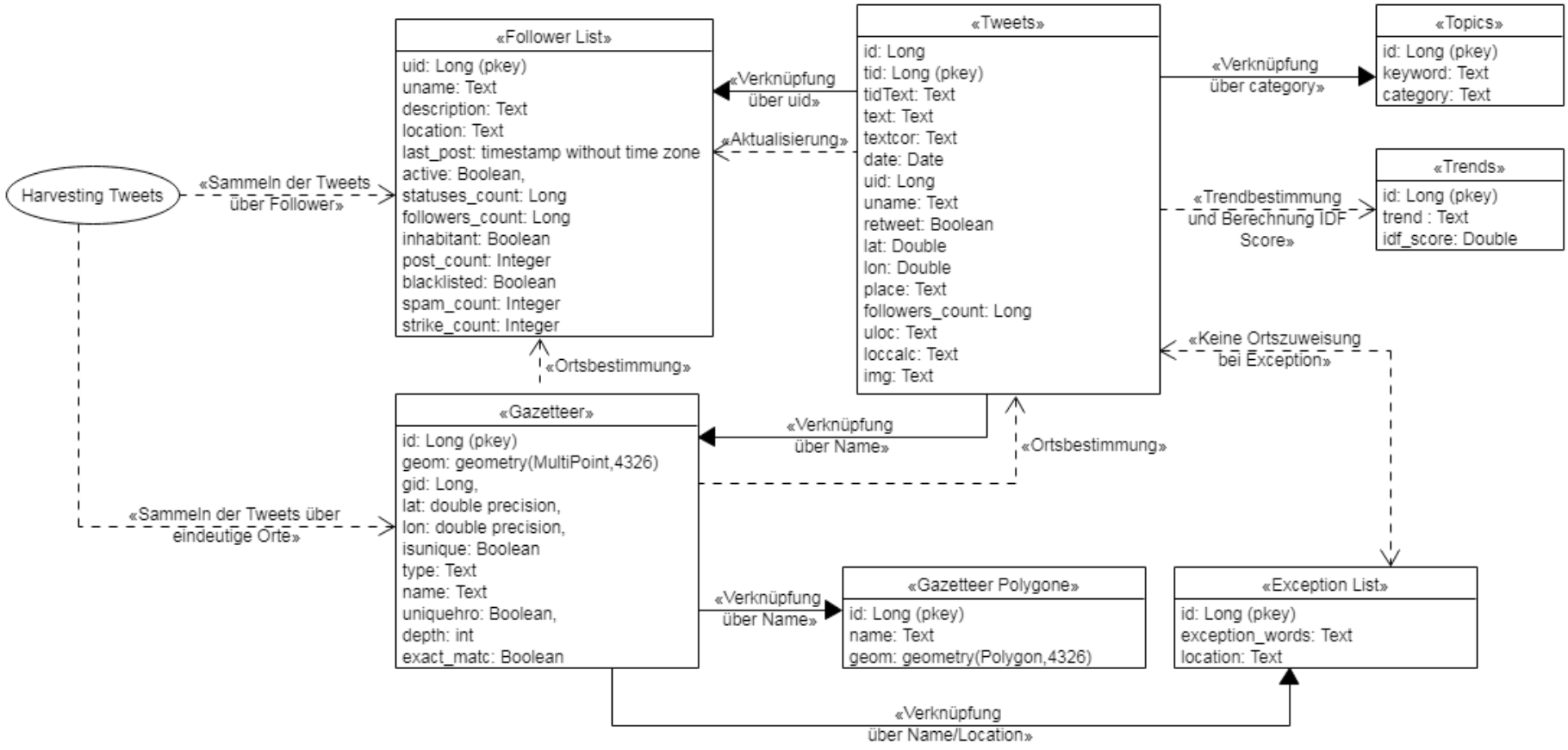


Abbildung 6-3: Datenbankstruktur.

6.2.2 Gazetteererstellung

Der Gazetteer stellt eines der wesentlichsten Elemente zur Verortung der Nachrichten dar. Seine Struktur bestimmt maßgeblich die Genauigkeit der Lokalisation sowie den zugehörigen Algorithmus. Zudem lassen sich nur die Ortsbezeichnungen in den Nachrichten verorten, die korrekt in den Gazetteer eingepflegt worden sind.

Der Gazetteer besteht aus vier Schichten, die sich der Skala nach von klein (Rostock als Ort) nach groß (Point of Interest - POI, Gebäude) ordnen (Abbildung 6-4). Die zweite Ebene, die Stadtteile bzw. Gebiete, ist mit einer zweiten Datentabelle verknüpft, die die zugehörige Polygoninformation enthält. Ansonsten wurde von allen Linien- oder Polygonobjekten der Schwerpunkt gebildet, d. h., dass jede Verortung auf der Basis eines Punktes stattfindet. Umgesetzt wurde diese Architektur mittels der PostgreSQL-Erweiterung PostGIS. Erstellt wurde der Gazetteer manuell aus OSM, dem offenen Datenportal der Hansestadt Rostock OpenData.HRO, den Lagekarten der Hanse Sail sowie individuellen Ergänzungen, die erst beim Test des Systems aufgefallen und integriert worden sind (z. B. Dobi für Doberaner Platz). Alle Elemente des Gazetteers und deren Verteilung sind in Abbildung 6-5 dargestellt.

Insgesamt besitzt der Gazetteer 5 254 Einträge. Davon sind 5 183 den beiden tiefsten Schichten, d. h. Straßen sowie Gebäuden/POI's zuzuordnen (Stand: 30.08.2018). Allerdings wird der Gazetteer kontinuierlich erweitert. Einige Orte besitzen hierbei mehrfache Bezeichnungen, da es für diese nicht nur eine eindeutige Beschreibung gibt. Dazu zählt zum Beispiel das Stadion (Ostseestadion, FC Hansa) oder auch der Doberaner Platz (auch als Dobi bezeichnet). Außerdem wurde jeder Eintrag auf seine Einzigartigkeit deutschlandweit sowie in Rostock geprüft. So lassen sich aus dem Gazetteer die Schlagworte für das Abgreifen der jeweiligen Nachrichtenstreams ableiten. Der Gazetteer wird durch eine PostgreSQL-Datenbank mit PostGIS-Erweiterung bereitgestellt.

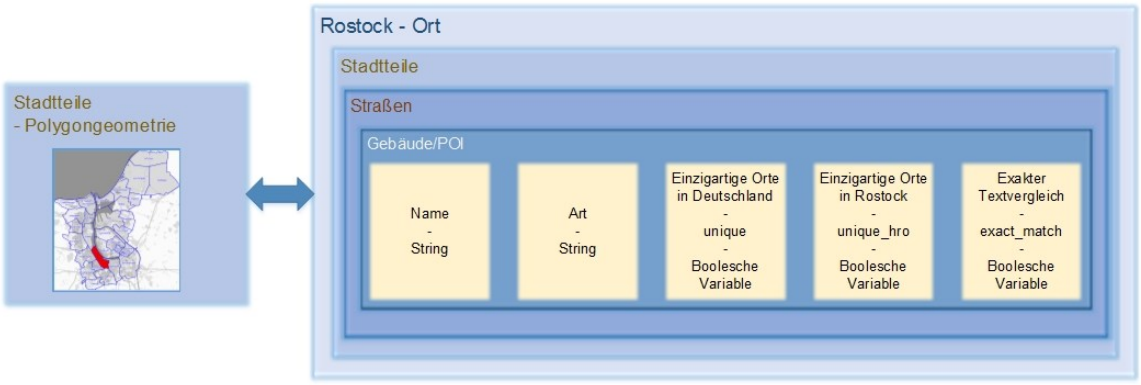


Abbildung 6-4: Struktur des vierschichtigen Gazetteers.

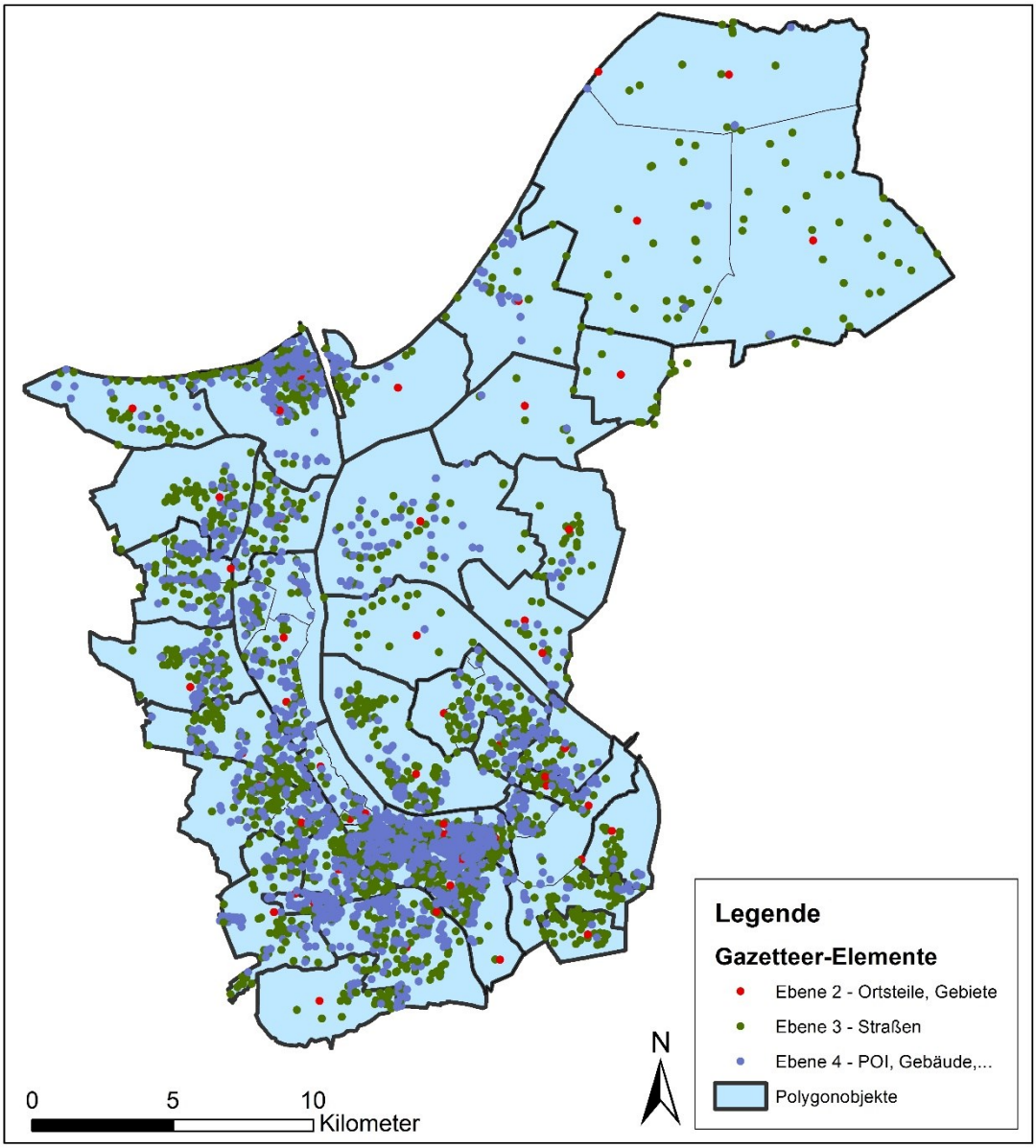


Abbildung 6-5: Verteilung der Ortsnamen in Rostock.

6.2.3 Filterstrategie und Harvesting

Prinzipiell lassen sich verschiedene Ansätze für die Filterung der Social-Media-Daten finden. Hierbei geht es vor allem darum, auf welche Datenbasis zurückgegriffen wird und an welcher Stelle die relevanten von den nicht-relevanten Daten getrennt werden (Abbildung 6-6). Eine mögliche erste Methode setzt nur auf die direkte Filterung des Streams. Hierbei sind die Filtermöglichkeiten jedoch stark begrenzt. So bietet Twitter lediglich die Möglichkeit 400 Suchbegriffe abzufragen (TWITTER INC. 2016b). Dies ist für die Komplexität der Abfragen zu gering.

Die zweite Methode wäre, die Nachrichten innerhalb der Applikation zu filtern und nur die gefilterten Elemente in die Datenbank zu schreiben. Dieses Vorgehen ermöglicht eine bessere Filterstrategie als das erste Verfahren, ist aber programmiertechnisch aufwendiger. Zudem besteht hier das Problem, dass alle Daten im Speicher vorgehalten werden müssen.

Der dritte Ansatz ist, eine möglichst große Menge an Daten zu sammeln und erst im zweiten Schritt diese wieder einzulesen und entsprechend zu filtern. Dies hat den Vorteil, dass eine große Datenbasis existent ist. Zudem müssen die Daten nicht im Speicher vorgehalten werden. Allerdings ist anzumerken, dass auch hier ein größerer Programmieraufwand zu Buche schlägt.

Als idealer Weg erweist sich eine Mischung aus Verfahren eins und drei. Da der Harvester mehrstufig aufgebaut ist, greifen hier verschiedene Filteralgorithmen. Zuerst wird eine Vorselektion getroffen, um nur die Nachrichten zu sammeln, welche sich tatsächlich auf Rostock oder Orte in Rostock beziehen. Diese Nodes müssen entsprechende amtliche Begriffe beinhalten oder einen Geotag besitzen (vgl. Kap. 6.2.2). Diese werden dann aus dem Gazetteer selektiert. Anschließend wird allen Nachrichten dieses Nutzers gefolgt. Die Idee ist, dass dieser Nutzer offenbar aus Rostock stammt oder sich in Rostock befindet. Damit können seine Nodes nun auch nach Toponymen durchsucht werden, die nicht nur für Rostock alleine gelten. Erwähnt der sich höchstwahrscheinlich in der Hansestadt befindliche Nutzer beispielsweise den Strand, ist er mit hoher Wahrscheinlichkeit an der Ostseeküste zu verorten. Somit kann er Gebieten in Rostock zugeordnet werden, ohne dass er diese bei ihrem exakten Namen nennen muss. Allerdings muss dieses Vorgehen eingeschränkt werden, da es sich bei dem Nutzer um einen Touristen oder Besucher der Hansestadt handeln kann. Daher wird ein Timer gesetzt. Dieser wurde entsprechend der durchschnittlichen Verweildauer auf drei Tage festgesetzt (vgl. Kap. 5.3). Außerdem soll dieser Timer ein Überlaufen der Follower-Liste verhindern, da die Twitter-Streaming API nur 5 000 Accounts zulässt, denen man parallel folgen kann (TWITTER INC. 2016b). Diese Expertenfindung wird dabei in Kapitel 6.2.8 genauer betrachtet.

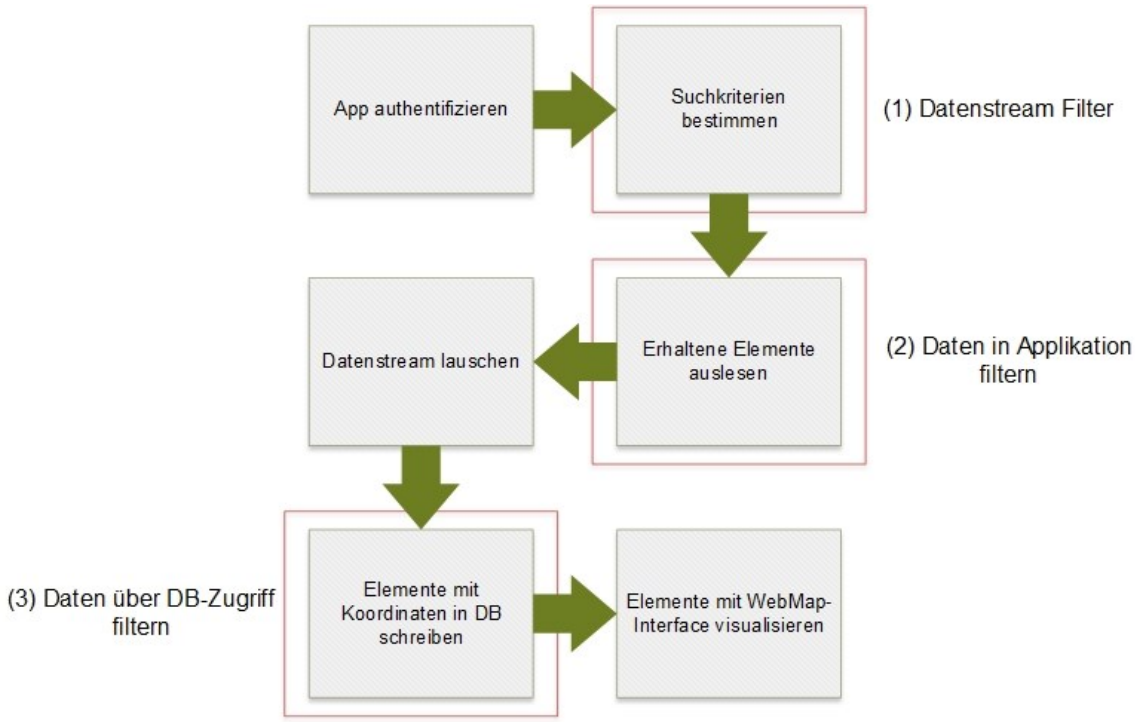


Abbildung 6-6: Grundlegender Aufbau des Streaming-Algorithmus.

Wesentlich für die Prozessierung der Nachrichten ist, dass diese relativ performant ist. So sind insgesamt relativ wenige Nachrichten zu erwarten (bei Twitter ca. 15 pro Stunde, vgl. VETTERMANN et al. 2018), dennoch sollte das Entstehen einer langen Warteschlange auflaufender Nachrichten vermieden und diese in naher Echtzeit abgearbeitet werden, da sonst u. U. Datenverluste auftreten könnten (Abbildung 6-7).

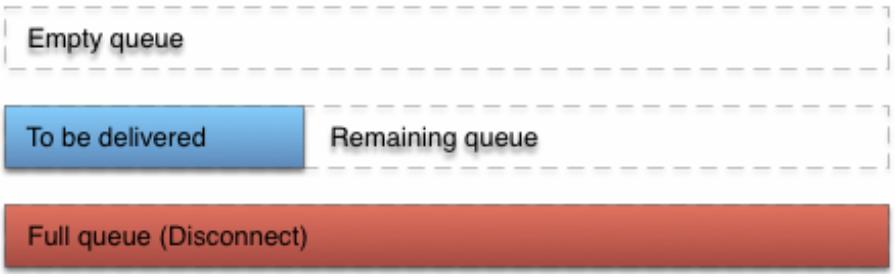


Abbildung 6-7: Warteschlange bei Twitter (TWITTER INC. 2016b).

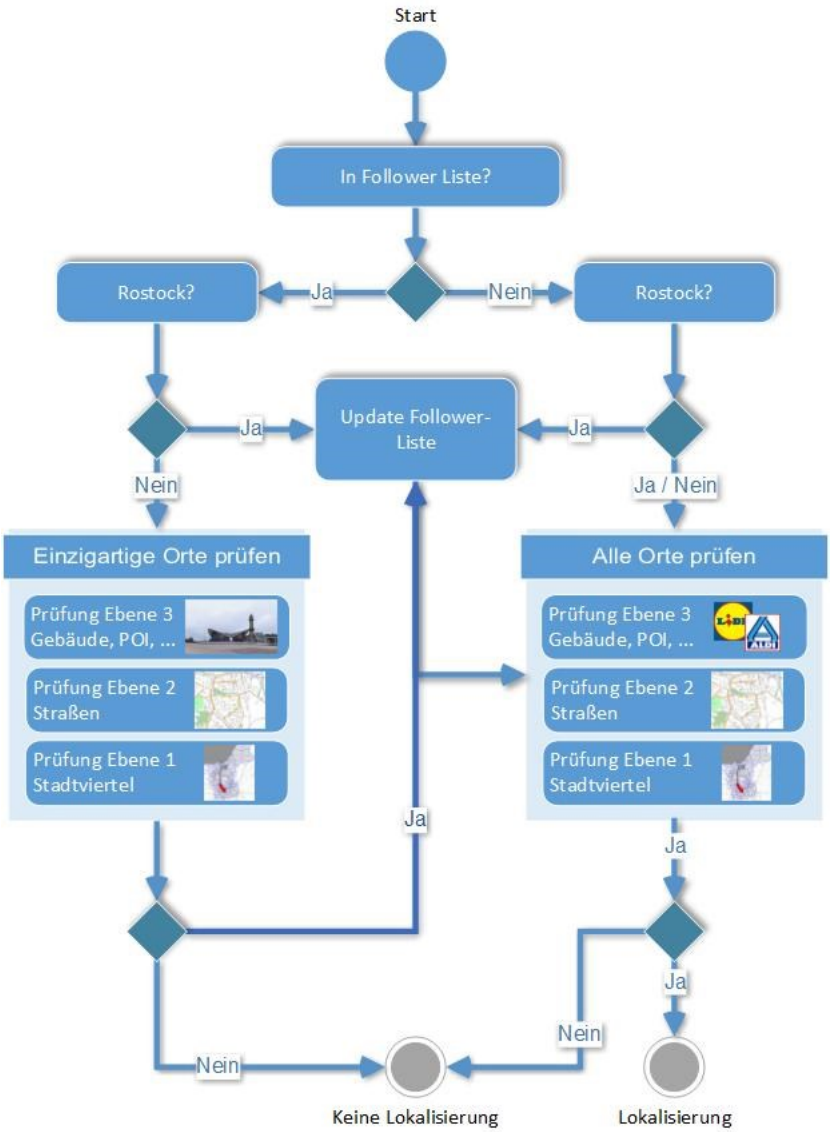


Abbildung 6-8: Grundlegende Struktur des Harvesters.

Da der Harvester darauf abzielt, die Daten live zu verarbeiten, unterscheidet sich die Struktur in wesentlichen Punkten von der in VETTERMANN et al. (2017a) vorgestellten (Abbildung 6-8). Die Anforderung, die Daten in naher Echtzeit zu verarbeiten, bereitzustellen und nach verschiedenen Kategorien und Themen zu visualisieren, erfordert zahlreiche Anpassungen, wie beispielsweise eine Methodik zur Expertenfindung, Anpassungen im Matching-Algorithmus mit dem Gazetteer sowie die Zuweisung von Themen (Topics), die Identifikation von Trends sowie das Zuweisen von Stimmungen mit Hilfe der Methoden der Sentimentanalyse (VETTERMANN et al. 2018).

Im Folgenden sollen die einzelnen Schritte, die zur Erfassung, Klassifizierung und Verortung der Tweets beitragen, näher beschrieben werden, um ein vollumfängliches Bild des Vorgehens aufzuzeigen.

6.2.4 Textaufbereitung und Normalisierung

Die Aufbereitung des Textes ist ein wesentlicher Schritt bei der Verarbeitung der Nachrichten. Sie umfasst eine Normalisierung, um spezifische Inhalte zu entfernen wie der Identifikation von Stimmungen in Form von Smilies und Gelächter (Tabelle 6-1), die Stemmatisierung, Tokenisierung und schließlich die Bildung von N-Grammen. Eine pauschale Orthografieprüfung der Tweets findet nicht statt, lediglich bei Wörtern mit drei oder mehr Wiederholungen eines Buchstabens wird nach einer passenden Ersetzung gesucht. Grund ist die informelle Sprache, weshalb eine pauschale Textkorrektur keine sinnvollen Ergebnisse bringt (ZHANG & GELERNTER 2014, MILLER & GOODCHILD 2015).

Tabelle 6-1: Social Media spezifische Inhalte.

Art	Beispiel
Smilies	:D :-) :)
Emoticons	😈 ☺ ☹
Buchstabenwiederholungen	Hallooooo, Mooooorgeeeeen
Umlaute	Ä, Ü, Ö, ß
Links	http://www.google.de
E-Mails	emailadresse@gmail.com
angesprochene Personen	@Sarazarlp, @TobiasHauke
Hashtags	#boehmermann, #olympia
Retweet Tags	@ RT

6.2.4.1 Erkennung von Links, E-Mails, angesprochenen Personen und Retweets

Die Erkennung von Links, E-Mails, angesprochenen Personen, Retweets sowie Hashtags ist durch die Suche nach dem ersten Buchstaben bzw. der beginnenden Zeichenfolge respektive der Endung (#, @, http, @ RT, .de) realisiert. Für die Sentimentanalyse sind dabei die enthaltenen Informationen unwichtig, für die Analyse des Ortsbezugs hingegen von Bedeutung. Daher wird hierfür, wie von SIDARENKA et al. (2013) vorgeschlagen, lediglich der # entfernt, das Wort aber beibehalten. Die Links und Mailadressen werden in jedem Fall gelöscht. Handelt es sich bei einer Nachricht um einen Retweet, wird dies ebenfalls in der Datenbank vermerkt, der Retweet-Tag wird ebenso gelöscht.

6.2.4.2 Erkennung von Smilies und Emoticons

Für die Erkennung von Smilies wurde der Algorithmus von RITTER et al. (2011) in bearbeiteter Form verwendet[42]. Dieser basiert im Wesentlichen auf einer Ersetzung auf Basis von

[42] https://github.com/aritter/twitter_nlp/blob/master/python/emoticons.py

Regular Expressions (RegEx). Anschließend werden diese je nach ihrer Stimmung mit „POSITIVE“, „NEGATIVE“ oder „NEUTRAL“ ersetzt (BESBINAR et al. o. J.). Für die Erkennung von Emoticons wird nach den einzelnen Unicode-Bereichen gesucht und diese ebenfalls mit den oben genannten Tags ersetzt.

6.2.4.3 Erkennung von Lachen

Um Lachen in den Nachrichten zu erkennen, wird auf einen eigens entwickelten Algorithmus zurückgegriffen. Dieser macht sich zunutze, dass im Deutschen bei geschriebenem Gelächter immer der Buchstabe h häufiger als alle anderen Buchstaben in einem Wort vorkommt. Dabei ist nach dem herkömmlichen Gelächter („hahaha“) und abgewandelten Formen („hihiahahaha“) zu unterscheiden. Bei ersterem greift die Regel, dass der Buchstabe h mindestens genauso häufig vorkommen muss wie der zweite Buchstabe. Im zweiten Fall muss der Buchstabe h mindestens zweimal und genauso häufig wie der zweithäufigste Buchstabe vorkommen. Zudem muss der String in jedem Fall mit dem Buchstaben h beginnen, da sonst Wörter wie Schach ebenfalls als Gelächter interpretiert werden würden. Allerdings hat dies den Nachteil, dass Varianten wie „ahahaha“ nicht als Gelächter klassifiziert werden können. Schließlich wird der String in jedem Fall durch „POSITIVE“ ersetzt.

6.2.4.4 Prüfung von Umlauten

Umlaute spielen insbesondere im Deutschen eine große Rolle. Häufig gibt es aber das Problem, dass diese in Twitter als ue, oe oder ae geschrieben werden (SIDARENKA et al. 2013). Für diesen Fall sollen diese durch ü, ö und ä ersetzt werden. Werden entsprechende Zeichenfolgen gefunden und ersetzt, wird anschließend geprüft, ob diese in einem vollumfänglichen deutschen Korpus der Universität Leipzig[43] vorkommen (GOLDHAHN et al. 2012). Dies soll dazu dienen, fehlerhafte Ersetzungen (Israel, Uerdingen, …) zu verhindern.

6.2.4.5 Normalisierung von Wiederholungen

Ein großes Problem sind ebenfalls Buchstabenwiederholungen wie „Halloooo“ oder „ahaaaaa“. Hierbei wird ebenfalls auf einen eigens entwickelten Algorithmus zurückgegriffen. Dabei wird zuerst das Auftreten aller Buchstaben gezählt und solche, die dreimal vorkommen, in ein Array gespeichert. Zuerst werden durch den NLTK-Tokenizer bereits alle Wiederholungen auf maximal drei eingekürzt. Anschließend werden alle Kombinationen mit allen sich wiederholenden Buchstaben des Wortes in ein weiteres Array geschrieben und gegen den Korpus der Universität Leipzig geprüft. Kommt davon eine Variante vor, wird das Wort durch diese ersetzt, kommt keine der Varianten im Wörterbuch vor, wird das originale Wort mit den Wiederholungen beibehalten.

Die Korrektur wird auf Basis von NORVIG (2016) durchgeführt. Dieses auf Bayes basierende Verfahren lieferte im Vergleich zu PyEnchant die besten Ergebnisse (KELLY 2014). Hierzu wurde der Wortschatz der Universität Leipzig integriert, auf dessen Basis ein Vergleich der Wörter durchgeführt wird (GOLDHAHN et al. 2012). Dieser Wortschatz umfasst insgesamt vier Dateien (deu_mixed-typical_2011_1M-words, deu_news_2015_1M-words, deu_newscrawl_2017_1M-words, deu_wikipedia_2016_1M-words), welche zusammengeführt worden sind und insgesamt 2 358 257 Wörter umfasst. Damit sollte ein Großteil des deutschen Sprachschatzes abgebildet werden. Dennoch kann keine absolut vollständige Korrektur gewährleistet werden. Die zahlreichen Abkürzungen und die in Sozialen Medien genutzte informelle Sprache machen dies fast schier unmöglich (ZHANG & GELERNTER 2014, MILLER & GOODCHILD 2015).

[43] http://wortschatz.uni-leipzig.de/de/download

6.2.4.6 Entfernung von Interpunktion und Stoppwörtern

Bei der Normalisierung werden zuletzt alle Satzzeichen sowie die Stoppwörter entfernt, da diese keinen Einfluss auf den Inhalt oder das Sentiment der jeweiligen Nachricht haben. Bei der Bereinigung um die Stoppwörter wurden potentiell stimmungsrelevante entfernt. Die Liste der Stoppwörter ist im Anhang angefügt.

6.2.4.7 Sprachzuordnung

Des Weiteren wurde der Track nach der Twitter-internen Sprachzuweisung nach deutschen Nachrichten gefiltert (TWITTER INC. 2016b). Dies ist sinnvoll, da die Sprachanalyse nur auf Deutsch durchgeführt werden soll, d. h. anderssprachige Nachrichten nicht berücksichtigt werden. Dadurch lässt sich ein konsistentes Bild der deutschsprachigen Twitter-Nutzer zeichnen. Nichtdestotrotz liefert die Twitter-interne Sprachzuordnung auch fehlerhafte Ergebnisse und es ist nicht eindeutig, welcher Algorithmus dieser zugrunde liegt (SCHEFFLER 2014). Eine nach SCHEFFLER (2014) vorgeschlagene In-Time-Anwendung der Python-Bibliothek langid wurde, bedingt durch die relativ lange Berechnungsdauer, ebenfalls verworfen (LUI & BALDWIN 2012). Ein direktes Einbinden dieser führt ebenfalls zu einem Überlaufen der Warteliste. Da langid sehr gute Ergebnisse hinsichtlich der Spracherkennung liefern soll, wurde eine anschließende Prüfung sowie ein Vergleich mit der Google Chromium Bibliothek durchgeführt (SITES 2014, AL-RFOU 2015).

Hierbei hat sich gezeigt, dass beide Bibliotheken hinsichtlich der korrekten Spracherkennung relativ schlechte Ergebnisse liefern. Dafür wurde jeweils eine Stichprobe von 10 000 Nachrichten zufällig aus dem Korpus ausgewählt und geprüft, ob diese jeweils als nicht Deutsch durch langid oder die Chromium Bibliothek klassifiziert wurden. So waren bei langid lediglich 11 % der Nachrichten richtigerweise als nicht Deutsch klassifiziert worden. 59 % dieser Nachrichten wurden hierbei von der Chromium Bibliothek richtig identifiziert. Wurde Chromium für die Erstklassifikation genutzt, zeigte sich, dass hier lediglich 7 % richtigerweise als nicht Deutsch eingeordnet wurden. 75 % der Nachrichten sind dabei durch langid korrekt identifiziert wurden. Dennoch weisen beide Bibliotheken eine so geringe Genauigkeit gegenüber der bereits von Twitter vorgefilterten Nachrichten auf, dass auf eine zweite Sprachprüfung verzichtet worden ist und auf die von Twitter verwendete Zuordnung zurückgegriffen wird.

6.2.5 Themenidentifikation

Die Zuordnung von Themen findet auf zwei Wegen statt. Zum einen werden die Themen mittels einer vordefinierten Datentabelle anhand von Schlagwörtern zugeordnet. Dieses einfache Matching bezieht sich auf folgende, vorab definierte übergeordnete Themen:

1. Wasser
2. Urlaub
3. Arbeit
4. Sicherheit
5. Veranstaltungen
6. Sport

Die zugehörigen Schlagwörter wurden ebenfalls manuell erfasst. Eine Integration von Methoden des maschinellen Lernens findet an dieser Stelle nicht statt. Dennoch sind die Themen von großer Bedeutung, da anhand dieser, sowohl für das Projekt KOGGE als auch für die Stadtplanung relevanten Kategorien, die Nachrichten letztlich analysiert sowie dargestellt werden sollen.

6.2.6 Trendidentifikation

Ein wesentlicher Bestandteil bei der Tweetanalyse ist die Identifikation von Trends. Durch die Identifikation von Trends sollen die Gesprächsthemen in der Hansestadt identifiziert werden. In Verbindung mit der Sentimentanalyse (vgl. Kap. 6.2.7) lässt sich so ein umfassendes Bild von Themen und Stimmungen in der Hansestadt erstellen.

Zur Trendidentifikation wird auf TF-IDF zurückgegriffen. Da das Verfahren darauf abzielt, die wesentliche Information aus Texten zu filtern (vgl. Kap. 3.4.4), erscheint es für die Identifikation von Trends ideal. Hierzu ist der TF-IDF-Prozess aus Scikit-Learn genutzt worden. Wesentlich ist, dass als N-Gram-Bereich eins bis zwölf Gramme gewählt worden. Dadurch werden auch komplette Tweets in die Trends miteinbezogen. Grund dafür ist, dass bei geringeren Grammen mehrere Trends für denselben Tweet erstellt werden und somit ebenfalls der gesamte Tweet als Trend gewertet wird. Um dies zu verhindern, werden anschließend alle Trends, die doppelte Elemente enthalten, um diese bereinigt. Abschließend werden nur Trigramme berücksichtigt.

Dieses Verfahren erscheint zwar nicht ideal, allerdings lassen sich die wesentlichen Themen so direkt identifizieren ohne nur auf Hashtags o. ä. Rücksicht zu nehmen. In die Trendbestimmung fließen dabei die Nachrichten der letzten 24 Stunden ein. Des Weiteren muss jedes der gefundenen Gramme mindestens sieben Mal vorkommen. Abschließend werden die zehn nach dem TF-IDF-Score bedeutendsten Trends in die zugehörige Datentabelle geschrieben.

Im Rahmen der Arbeit ist allerdings noch ein zweites Verfahren zur Identifikation von Trending Topics integriert worden, welches auf der LDA basiert (BLEI et al. 2003). Dieses überschneidet sich allerdings sehr stark mit der Themenidentifikation, wobei hierfür die Themen nicht vorgegeben werden müssen. Lediglich die Anzahl der Themen ist bei der Berechnung der LDA über Gensim vorzugeben (REHUREK & SOJKA 2010). In die LDA fließen dabei die Tweets als BOW-Modell ein.

6.2.7 Sentimentanalyse

Die Sentimentanalyse dient dazu, den einzelnen Nachrichten eine Stimmung (Positiv – Neutral – Negativ) zuzuweisen. Durch die Zuweisung von Stimmungen soll ermöglicht werden, dass die Einstellungen der Bürger als auch der auf den Sozialen Plattformen aktiven Medien in Rostock zu bestimmten Themen räumlich visualisiert, aber auch quantifiziert werden können.

Am bedeutendsten dafür ist eine korrekte Aufbereitung des Textes. Vor allem Smilies und Emoticons, aber auch geschriebenes Lachen sind für eine korrekte Klassifikation von großer Bedeutung (DESHWAL & SHARMA 2016, SIDARENKA & STEDE 2016, AGARWAL et al. 2011). Für die Aufbereitung wird auf die in Kap. 3.4.3.2 beschriebenen Verfahren zurückgegriffen.

6.2.7.1 Korpus-Erstellung

Die Umsetzung findet über die Python-Bibliotheken NLTK und Scikit-Learn unter Verwendung von Algorithmen des maschinellen Lernens statt (vgl. Kap. 3.4.2, Kap. 4.1.2). Wesentlich ist dabei das ausgewählte Verfahren sowie der verwendete Korpus für die Klassifikation. Im deutschsprachigen Raum stehen vier Korpora (PotTS – Potsdam Twitter Sentiment Corpus, MGS – Multilingual Sentiment Corpus, DAI – Distributed Artificial Intelligence, SB10k – Spinning Bytes 10 000 Wörter) für die Klassifikation zur Verfügung (CIELIEBAK et al. 2017). Einer (PotTS - SIDARENKA 2016) kann jedoch nicht direkt für eine

Weiterverarbeitung genutzt werden. Der MGS-Korpus weist eine unzureichende Genauigkeit zwischen den einzelnen Annotatoren auf, weshalb auch dieser verworfen wird (MOZETIČ et al. 2016). Der DAI-Korpus entfällt ebenfalls, da dieser nur 1 800 Nachrichten umfasst und somit relativ klein ist (NARR et al. 2012). Daher verbleibt nur der SB10k Korpus von CIELIEBAK et al. (2017) zur Nutzung.

Nach Angaben von CIELIEBAK et al. (2017) weist dieser bei der Zuordnung der Stimmungen von Positiv, Neutral und Negativ mittels SVM eine Genauigkeit (F1) von 56.98 % auf. Das Verhältnis Test- zu Trainingskorpus betrug 9:1. Größte Schwachstelle ist dabei die Zuordnung der negativen Tweets (Tabelle 6-2). Erst unter Verwendung eines CNN werden die Genauigkeiten deutlich verbessert. Dies zeigt auch der Test an zwei anderen Korpora (MGS, DAI).

Tabelle 6-2: Genauigkeiten des SB10k Korpus (CIELIEBAK et al. 2017).

Klassifikator	Trainingskorpus	Testkorpus	$F1_{pos}$	$F1_{neg}$	$F1_{neutral}$	F1
SVM	SB10k	SB10k	66.16	47.80	81.32	56.98
CNN	SB10k	SB10k	71.46	58.72	81.19	65.09
SVM	SB10k	MGS	49.50	38.62	66.41	44.06
CNN	SB10k	MGS	50.41	44.19	71.81	47.30
SVM	SB10k	DAI	62.30	61.40	81.22	61.85
CNN	SB10k	DAI	62.79	58.43	79.92	60.61

Da die Weitergabe von eindeutigen Tweets nach den Nutzungsbedingungen von Twitter verboten ist (TWITTER INC. 2016b), muss der Korpus mit Hilfe eines Scriptes[44] aus den Tweet-ID's bezogen werden. Dies ist insofern problematisch, da ein Teil der Nachrichten nicht mehr verfügbar ist. Der Korpus umfasst somit anstatt der ursprünglichen 9 738 Tweets nur noch 7 342 Tweets (Stand: 07.12.2017). Aus diesem Grund ist ein negatives Abweichen der Genauigkeiten zu erwarten, weshalb im Rahmen dieser Arbeit mehrere Klassifikationen getestet worden sind, auch unter Verwendung von regelbasierten Ansätzen. Zudem sind 1 680 positive, 1 094 negative sowie 4 567 neutrale Tweets vorhanden, wodurch ein deutliches Ungleichgewicht zugunsten der neutralen Tweets gegeben ist.

Jede Nachricht wurde anschließend mit denselben Verfahren aus Kapitel 6.2.4 aufbereitet, wobei deren Bigramme gebildet werden. Diese zeigen für die Sentimentanalyse die besten Ergebnisse, da sich spezifische Wortkombinationen, welche positiv bzw. negativ wirken, abbilden lassen. Beispiele hierfür sind z. B. Kombinationen wie sehr gut, ziemlich gut oder nicht gut. Daraus wird deutlich, dass gut alleine nicht zwingend für ein positives Sentiment stehen muss. Erst in Verbindung mit sehr oder nicht wird das Sentiment tatsächlich definiert (PFAFFENBERGER 2016).

Zusätzlich wurde ein grob selektierter und gelabelter Korpus nach DERIU et al. (2017) bzw. GO et al. (2009) erstellt. Dafür wurden ca. eine Million Tweets (168 000 negativ, 876 000 positiv) mit Hilfe der Track-Variablen :((für negative Tweets) bzw. :) (für positive Tweets) gesammelt. Dieser als grob klassifizierte Korpus kann nun ebenfalls als Trainingsdatensatz angewendet werden, da der SB10k Korpus für ein umfassendes Training als zu klein zu werten ist. Zudem ist es so möglich zu testen, wie sich die Klassifikatoren bei größeren Trainingsdatensätzen verhalten. Mit Hilfe des Transfer-Learnings kann die Sentiment-Klassifikation auf diesem Wege verbessert werden (DERIU et al. 2017).

[44] https://github.com/AnthonyMRios/Sentiment-Classification-Example/tree/master/sentiment-Data/train/twitter_download-master

6.2.7.2 Auswahl des Klassifikators

Für die Sentimentanalyse steht eine ganze Reihe an Klassifikatoren zur Verfügung (vgl. Kap 3.4.2). Der verwendete Korpus von CIELIEBAK et al. (2017) wurde zur Auswahl in einen Test- und Trainingsdatensatz geteilt (70 % Training, 30 % Test), um eine grundlegende Vorauswahl des Klassifikators durchzuführen. Zur besseren Vergleichbarkeit fand ein Oversampling mit Hilfe der Synthetic Minority Over-sampling Technique (SMOTE) statt. Grund dafür ist die Ungleichverteilung der Klassen im SB10k Korpus. Zur Anwendung kamen hierbei die beiden Python-Bibliotheken Scikit-Learn sowie NLTK mit den Klassifikatoren Logistic Regression, SVM (LinearSVC sowie SVC mit linearem Kernel, vgl. SCIKIT-LEARN 2017), Decision Tree, Random Forest sowie Naive Bayes (Abbildung 6-9)

Der Vergleich der Validierungsgenauigkeiten zeigt, dass die beiden SVM-Klassifikatoren die höchste Genauigkeit liefern. Lediglich die Logistische Regression kann mit einer Genauigkeit von 0.69 mithalten, welche aber dennoch um fünf Basispunkte geringer ist als die der SVM. Aus dem Vergleich von Trainings- und Validierungsgenauigkeit wird dabei deutlich, wie sensitiv die Klassifikatoren hinsichtlich der Klassifikationsgenauigkeit auf größere Trainingsdatensätze reagieren. Dabei zeigt sich, dass dies sowohl auf die Logistische Regression, Naive Bayes als auch für die SVM zutrifft. Allen Klassifikatoren ist gemein, dass die negativen Tweets mit unzureichender Genauigkeit klassifiziert werden. Daher empfehlen sich letztlich ein weiterer Test sowie ein Vergleich mit CNN und SVM.

6.2.7.3 Support Vector Machines

Wesentlich für die Performanz des SVM-Klassifikators ist die Struktur der Eingangsdaten, d. h. welche N-Gramme hier die besten Ergebnisse liefern und wie der Text in Vektoren überführt wird. Hierzu sind zwei verschiedene Textvectorizer (TF-IDF, Count Vectorizer) von Scikit-Learn mit Uni-, Bi-, Tri- und Tetragrammen getestet worden (Abbildung 6-10). Aus dem Test wird deutlich, dass die TF-IDF-Verfahren deutlich bessere Ergebnisse als die Count Vectorizer liefern. Außerdem wird deutlich, dass die höheren Gramme keine besseren Ergebnisse liefern als die Uni- und Bigramme. Allerdings sind die Unterschiede mit unter einem Basispunkt marginal und bei einer größeren Datenbasis könnten die höheren Gramme u. U. bessere Ergebnisse liefern.

Unterstützt wird die Nutzung von Tetragrammen durch die Grid-Search Methode von Scikit-Learn. Bei dieser werden verschiedene Parameterkombinationen gegeneinander getestet, um die bestmöglichen zu finden. Für den SVM-Klassifikator (SVM mit linearem Kernel) ergeben sich die folgenden Parameter:

N-Gram Range: 1.4

Gamma: 1

C: 100

Die Genauigkeit liegt dabei bei 73 %. Problematisch sind jedoch weiterhin die Genauigkeiten in der Klassifikation des positiven und negativen Sentiments. Dies geht aus der Konfusions-Matrix der Klassifikation in Tabelle 6-3 als auch aus den klassenspezifischen Genauigkeiten hervor.

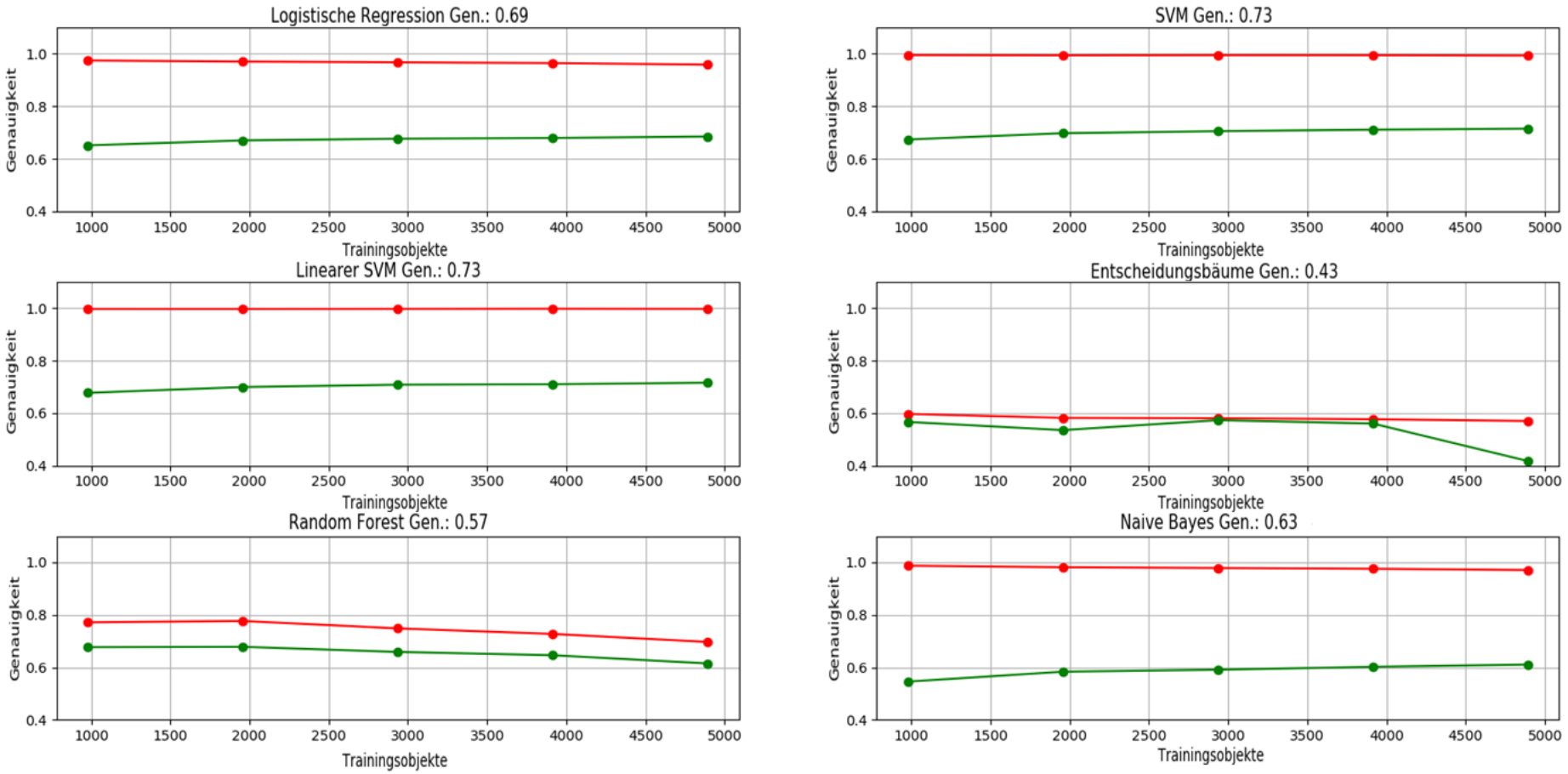


Abbildung 6-9: Vergleich der Trainingsgenauigkeit (Rot) und der Validierungsgenauigkeit (Grün) für mehrere Klassifikatoren zur Sentimentbestimmung.

1. Negative Tweets
    a. Präzision: 52 %
    b. Recall: 20 %
    c. F1: 29 %
2. Neutrale Tweets
    a. Präzision: 74 %
    b. Recall: 89 %
    c. F1: 81 %
3. Positive Tweets
    a. Präzision: 54 %
    b. Recall: 44 %
    c. F1: 49 %

Tabelle 6-3: Konfusionsmatrix des SVM-Klassifikators.

	Negativ	**Neutral**	**Positiv**
Negativ	52	154	57
Neutral	30	1 035	100
Positiv	18	209	181

Aus diesem Grund wurde die Klassifikation deutlich vereinfacht sowie die Datenbasis vergrößert, da gerade der SVM-Klassifikator gezeigt hat, dass mehr Trainingsdaten die Genauigkeit verbessern können. Insbesondere der kleine Korpus scheint hier eines der wesentlichsten Probleme darzustellen. Hier kommt der grob selektierte und gelabelte Korpus nach Smilies zur Anwendung. Mit diesem ist jedoch nur eine Klassifikation nach positivem und negativem Sentiment möglich. Im Rahmen der Arbeit sollte dies aber ausreichend sein, vor allem wenn die genauere Klassifikation keinen wirklichen Mehrwert aufgrund der hohen Ungenauigkeiten bringt.

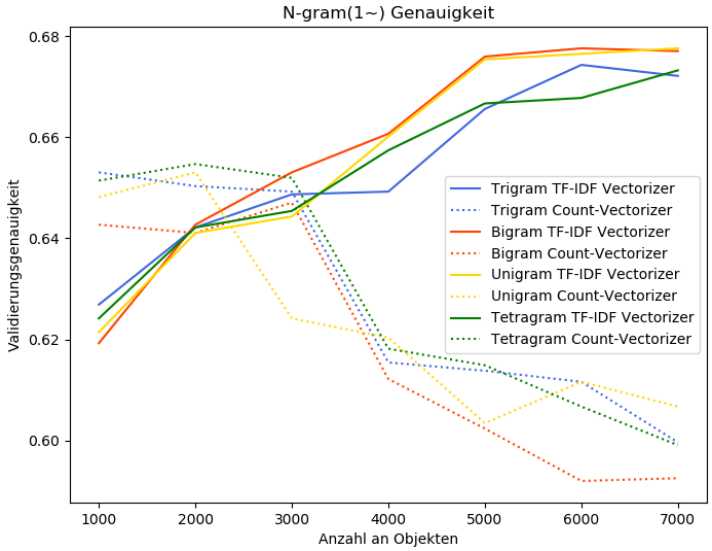


Abbildung 6-10: Vergleich der unterschiedlichen Text-Vectorizer auf Basis des SB10k Korpus und Klassifikation in Positiv, Neutral und Negativ.

In Abbildung 6-11 wird deutlich, dass bei einer Genauigkeit von über 80 % kein nennenswerter Anstieg der Genauigkeiten trotz einer höheren Anzahl an Trainingsdaten festzustellen ist. Aus diesem Grund fließen 150 000 positive und 150 000 negative Tweets in die Klassifikation ein. Dies führt zu einer Genauigkeit von 83 % und folgender klassenspezifischer Genauigkeiten:

1. Negative Tweets
    a. Präzision: 81 %
    b. Recall: 84 %
    c. F1: 82 %
2. Positive Tweets
    a. Präzision: 83 %
    b. Recall: 80 %
    c. F1: 81 %

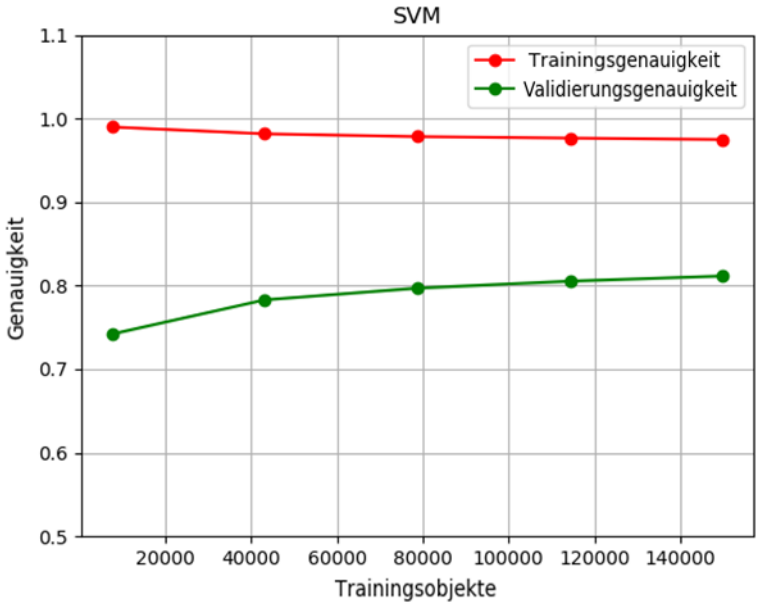


Abbildung 6-11: Genauigkeit des SVM-Klassifikators im Verhältnis zur Anzahl an Trainingsdaten.

Aus diesen wird deutlich, dass der Klassifikator mit einer Wahrscheinlichkeit von insgesamt über 80 % die Tweets richtig einer positiven bzw. negativen Stimmung zuordnen kann. In Tabelle 6-4 ist die zugehörige Konfusionsmatrix dargestellt. Durch die Möglichkeit, die Zuordnungswahrscheinlichkeit zu berechnen, lassen sich ebenfalls neutrale Tweets identifizieren.

Zudem wurde der Klassifikator auf den SB10k Korpus angewendet und sollte die positiven und negativen Tweets dieses Korpus identifizieren. Eine Genauigkeit von 67 % (Negativ: 56 %; Positiv: 78 %) zeigt, dass eine akzeptable Performanz mit dem angepassten SVM-Klassifikator erreicht werden kann.

Tabelle 6-4: Konfusionsmatrix des SVM-Klassifikators.

	Negativ	**Positiv**
Negativ	31 360	6 059
Positiv	7 573	30 008

6.2.7.4 Convolutional Neural Networks

Bedingt durch die guten Ergebnisse, die CIELIEBAK et al. (2017) mit CNNs im Rahmen der Sentimentanalyse erreicht haben, ist der zugehörige SB10k Korpus zum Training eines einfachen CNNs verwendet worden. Zudem fand eine Umwandlung der Wörter in Vektoren statt. Hierzu ist ein mit Word2Vec trainierter Embedding Layer[45] auf Basis von 200 Mio. Tweets mit 200 Dimensionen je Wort gewählt worden (CIELIEBAK et al. 2017). Anschließend wurden die Wörter des SB10k Korpus den Vektoren aus dem Embedding-Layer zugewiesen. Zum Vergleich wurde mittels Word2Vec zudem ein eigener Embedding-Layer auf Basis des Korpus trainiert. Zur Darstellung der Vorteile der vortrainierten Embeddings sind die beiden Layer jeweils in zwei Dimensionen umgewandelt und dargestellt worden. Aus dem Vergleich der beiden Embeddings wird deutlich, dass mittels des vortrainierten Korpus sich klar Cluster ähnlicher Wörter bilden (Abbildung 6-12). Gerade hinsichtlich der Sentimentanalyse ist dies das gewünschte Ergebnis.

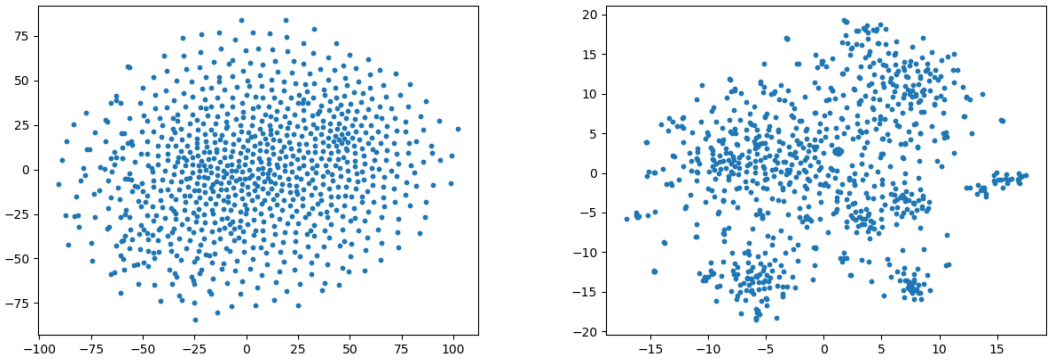


Abbildung 6-12: word2vec-Modell vor (links) und nach der Verwendung der Word-Embeddings.

Dem CNN liegt der in Abbildung 6-13 gezeigte Aufbau zugrunde. Insgesamt besitzt es 4 202 491 Parameter, von denen allerdings nur 341 891 trainiert werden. Die restlichen entstammen aus den Word Embeddings, welche selbst nicht trainiert werden.

Das Netz erreicht nach 20 Trainingsdurchläufen eine Genauigkeit für die Validierungsdaten von 59 %. Grund für die schlechte Genauigkeit sind die Klassifikationsgenauigkeiten für positive und negative Tweets von 37 % bzw. 21 %. Lediglich die neutralen Tweets werden mit einer ausreichenden Genauigkeit von 77 % klassifiziert.

Daneben wurde ein LSTM-Network getestet. Die Wörter wurden ebenfalls wieder über die vortrainierten Embeddings eingebunden. Die Struktur des Netzes ist in Abbildung 6-13 dargestellt. Die Gesamtgenauigkeit lag hier bei 71 %. Allerdings waren die Genauigkeiten für positive Tweets mit 42 % und negative Tweets mit 32 % deutlich zu gering. Aus diesem Grund wurde der Ansatz verworfen, da der Zwei-Klassen-SVM-Klassifikator bereits bessere Genauigkeiten lieferte und zudem für die vorliegende Aufgabe ausreichend ist.

[45] http://www.spinningbytes.com/resources/wordembeddings/

CNN	LSTM
Layer – Dimensionen – Parameter	Layer – Dimensionen – Parameter
Input – 134 – 0	Input – 23 – 0
Embedding – 134, 200 – 3 860 600	Embedding – 23, 200 – 171 400
Conv1D – 130, 128 – 128 128	Zero-Padding 1D – 33, 200 – 0
MaxPooling – 43, 128 – 0	Conv1D – 28, 200 – 240 200
Conv1D –39, 128 – 82 048	MaxPooling – 13, 200 – 0
MaxPooling – 13, 128 – 0	Conv1D – 8, 200 – 240 200
Conv1D –9, 128 – 82 048	MaxPooling – 1, 200 – 0
MaxPooling – 3, 128 – 0	Flatten – 200 – 0
Flatten – 384 – 0	Dense – 128 – 40 200
Dense – 128 – 49 280	Softmax – 3 – 603
Dense – 3 – 387	

Abbildung 6-13: Struktur des CNN und des LSTM zur Sentimentanalyse.

6.2.8 Expertenfindung

Ein wesentlicher Bestandteil der Verortung ist die Follower-Liste. Durch sie soll ermöglicht werden, Nutzer einzubeziehen, die bereits ein LIW gepostet haben, welches sich eindeutig auf Rostock bezieht. Anschließend wird eine Verweildauer von drei Tagen für jeden Nutzer in der Follower-Liste angenommen. Die Verweildauer ist aus dem statistischen Jahrbuch der durchschnittlichen Aufenthaltsdauer eines Touristen entnommen (drei Tage). Daneben können Nutzer auch als Einwohner Rostocks klassifiziert werden. Dies geschieht, sobald ein Nutzer drei Mal in einem Abstand der jeweils größer als drei Tage ist, eine Nachricht veröffentlicht, die sich eindeutig auf Rostock bezieht. Da die Anzahl der Personen, denen gefolgt werden kann, auf 5 000 beschränkt ist, werden die als Bewohner angenommenen Personen bei einer Inaktivität von einem Monat wieder entfernt. Dadurch sollen zum einen Fehler im Fall von Langzeiturlaubern vermieden sowie inaktive Nutzer ausgeschlossen werden.

Zusätzlich findet über die Followerliste eine grobe Filterung von Spam statt. Wenn ein Nutzer innerhalb von drei Minuten fünf beliebige Nachrichten verschickt, bekommt er einen Strike. Bei wiederum fünf Strikes wird der Nutzer geblacklistet, d. h., dass keinerlei Nachrichten mehr in die Verortung einfließen. Die zugehörigen Werte haben sich im Test als verlässlich erwiesen. Daneben wurden noch per Hand Accounts mit pornografischen Inhalten geblacklistet, da diese ebenfalls nur dazu dienen, Werbung zu verteilen und deren Nachrichten es mit möglichst vielen Hashtags und Ortsangaben darauf anlegen, gefunden zu werden.

Die Follower-Liste beschreibt letztlich also eine Methodik der Expertenfindung, wobei im konkreten Fall ein Experte eine Person ist, die sich in Rostock verorten lässt (ZHANG et al. 2007). Der Nutzen liegt darin, dass jeder Post dieses Nutzers innerhalb dieser Zeit auch mit den nicht-einzigartigen Orten des Gazetteers verglichen wird. Somit lässt sich eine größere Anzahl an Nachrichten verorten, womit das Problem der kleinen Datenbasis adressiert wird.

6.2.9 Gazetteer-Matching

Das Gazetteer-Matching stellt das Herz der Anwendung dar, da dadurch die einzelnen Nachrichten geokodiert werden. In den Matching-Algorithmus fließen alle vorher generierten Informationen ein. Grundsätzlich findet ein schichtweiser Vergleich des Nachrichtentextes mit dem Gazetteer statt, wobei die Lokationen, welche keine eindeutige Übereinstimmung benötigen, mit Hilfe von Skip-Grams und dem Dice-Koeffizienten extrahiert werden (vgl. Kap. 6.2.2, Kap. 3.4.3.2). Als Schwellenwert (Treshold) wurde eine Ähnlichkeit nach Dice von 90 % gewählt, da so einfache Tippfehler abgefangen, jedoch keine sinnentfernten Wörter erkannt werden.

Es findet sich allerdings auch eine Vielzahl von Einträgen im Gazetteer, die nicht nur einem einzigen Ort zuordenbar sind. Dazu zählen beispielsweise Einzelhandelsketten wie Lidl oder Rewe, deren Filialen über das gesamte Stadtgebiet verteilt sind. Um hier bei einer Erwähnung eine korrekte Geokodierung durchzuführen, kommen zwei Strategien zur Anwendung. Die erste nutzt die Polygonelemente, d. h. Bereiche im Stadtgebiet. Wird ein solches erkannt, werden für den weiteren Vergleich nur noch die Elemente innerhalb dieses Bereichs herangezogen. Dadurch lässt sich bei multiplen Objekten i. d. R. schon eine Einschränkung auf einige wenige bzw. ein einzelnes Objekt durchführen. Werden dennoch mehrere gefunden, wird die Nachricht zu allen gefundenen Orten zugewiesen. Die zweite Methode ist, insofern kein Polygonobjekt gefunden wird, eine eindeutige Lokation zu Hilfe zu nehmen. Wird ein in Rostock einzigartiger Ort gefunden (z. B. „Ich bin im Lidl bei der OSPA-Arena“), wird das nächstgelegene Element verwendet.

Da es in Rostock Lokationen gibt, die einer gesonderten Behandlung bedürfen, werden diese über eine eigene Liste abgefangen. Dies gilt beispielsweise für den „Alter Strom“, eine Straße in Warnemünde, um sicherzustellen, dass nicht alle elektrizitätsbezogenen Meldungen in Warnemünde verortet werden.

6.3 Front-End und Visualisierung

Durch die Nutzung der KOGGE-GDI wird der Twittermonitor in das Projekt eingebettet und die redundante Entwicklung einer Infrastruktur vermieden. Vor allem GeoServer ist dabei von Bedeutung. Über diesen werden mit Hilfe einer Datenbankverbindung die Tweets als WMS, als auch als WFS geliefert. Daneben lassen sich über GeoServer zeitbasierte Layer bereitstellen (GEOSERVER 2014). Dies ist vor allem für die Darstellung von Bedeutung, da so direkt aus den Datumswerten eine entsprechende temporale Abfolge erzeugt werden kann, welche sich schließlich abbilden lässt.

Des Weiteren werden über den für KOGGE bereitgestellten Server die zugehörigen Websites gehostet. Auch hier war der wesentliche Grund, dass bereits eine Infrastruktur verfügbar ist, die mit relativ geringem Aufwand angepasst und genutzt werden kann.

Das Front-End des Twittermonitors bildet eine HTML-Seite, welche die verorteten und abgespeicherten Tweets in einer Webmap visualisiert (Abbildung 6-14). Auf die Oberfläche von GeoNetwork wurde bewusst verzichtet, da diese für den konkreten Anwendungsfall nicht die gewünschten Features beinhaltet. Hierfür wurde das Framework Bootstrap verwendet, das es ermöglicht, unabhängig von der Bildschirmgröße eine korrekte Darstellung der Seite zu gewährleisten (BOOTSTRAP 2018).

Direkt unter dem Kopfsegment befindet sich ein Schieberegler (Timeslider), der es dem Benutzer ermöglicht, ein Datumsintervall, innerhalb dessen sich die dargestellten Tweets befinden, einzustellen. Die beiden Enden des Balkens, den man innerhalb des Sliders

verschieben kann, stellen das Anfangs- bzw. das Enddatum des gewünschten Intervalls dar. Beim Aufruf des Twittermonitors wird der aktuelle Tag abgefragt und der Balkenbereich auf die letzten 14 Tage eingestellt. Der Balken kann sowohl im Ganzen verschoben werden als auch beide Enden getrennt voneinander. Zudem wurden eine Minimum- sowie eine Maximumbreite festgelegt (zwei Wochen bzw. ein Tag). Der zeitbezogene Schieberegler wurde über das Open Source JavaScript Plugin *jQRangeSlider*[46] realisiert. Über eine CSS-Datei ist es möglich, den Stil des Sliders entsprechend anzupassen.

Die Webmap, auf der die Tweet-Lokalisationen dargestellt sind, wurde mit OpenLayers 3 (OL3) sowie GeoServer umgesetzt (GRATIER et al. 2015, GEOSERVER 2014). Dazu wird durch GeoServer auf die PostgreSQL-Tabelle mit den verorteten Tweets zugegriffen. An dieser Stelle ist anzumerken, dass hierfür die benötigten Daten für die Darstellung in einer separaten Datentabelle gespeichert werden. Diese beinhaltet nur die Tweets der letzten 14 Tage, da so gelöschte Tweets nicht mehr angezeigt werden. Ein permanentes Löschen kommt aufgrund des Datenumfangs und der dem gegenüberstehenden Twitter Rate Limits nicht in Frage. Zudem ist der Zugriff Passwortgeschützt. Des Weiteren sind auf GeoServer verschiedene Stile in Form von SLDs je nach Thematik des Tweets hinterlegt. Die Tweets werden als WMS in die Karte integriert, wobei die dargestellten Punkte durch die vom Timeslider ausgewählte Zeitspanne begrenzt werden. Um dies zu realisieren, wird die Möglichkeit genutzt, WMS-Anfragen an GeoServer direkt mit einem Query für die Zeitbegrenzung zu senden (GEOSERVER 2014). Auf eine Darstellung via WFS wird aus Performanzgründen verzichtet. Für eine bessere Visualisierung von Hotspots wird auf die Möglichkeit von OpenLayers 3, Heatmaps aus WMS-Daten zu erzeugen, zurückgegriffen (OPENLAYERS 3 2017). Durch diese Funktion sind der zeitliche Verlauf und die Ausprägung diverser Hot-Spots sehr gut erkennbar.

Über zwei Buttons am rechten Rand der Karte lassen sich aktuelle Trends bzw. Themen anzeigen. Im Themen-Overlay ist es möglich, die einzelnen Themen zu selektieren und sich die zugehörigen Tweets sowohl auf der Karte als auch den Inhalt der letzten zehn Tweets anzeigen zu lassen. Hierzu wird neben dem WMS noch ein WFS mit diesen Tweets hinzugefügt. Dargestellt werden sie über die Twitter eigene JavaScript Bibliothek, da so gelöschte Tweets nicht mehr sichtbar sind. Für eine bessere Navigation und Visualisierung wird zudem die Nachricht, die in dem Fenster ausgewählt wurde, auch auf der Karte hervorgehoben. Beim Klick auf ein einzelnes Thema wird die Nachrichtenverteilung über den ausgewählten Zeitraum eingeblendet. Dadurch ist es möglich, sich auf einzelne Ausschnitte mit einem besonders hohen Nachrichtenaufkommen zu konzentrieren. Neben dem Text werden auch geteilte Bilder angezeigt. Dieses Feature soll es ermöglichen, bei potentiellen Ereignissen sich sofort einen Überblick über Ort und Lage zu verschaffen. Die Trends sollen dazu dienen, aktuelle Schlüsselthemen zu identifizieren und zu lokalisieren.

.

[46] http://ghusse.github.io/jQRangeSlider/index.html

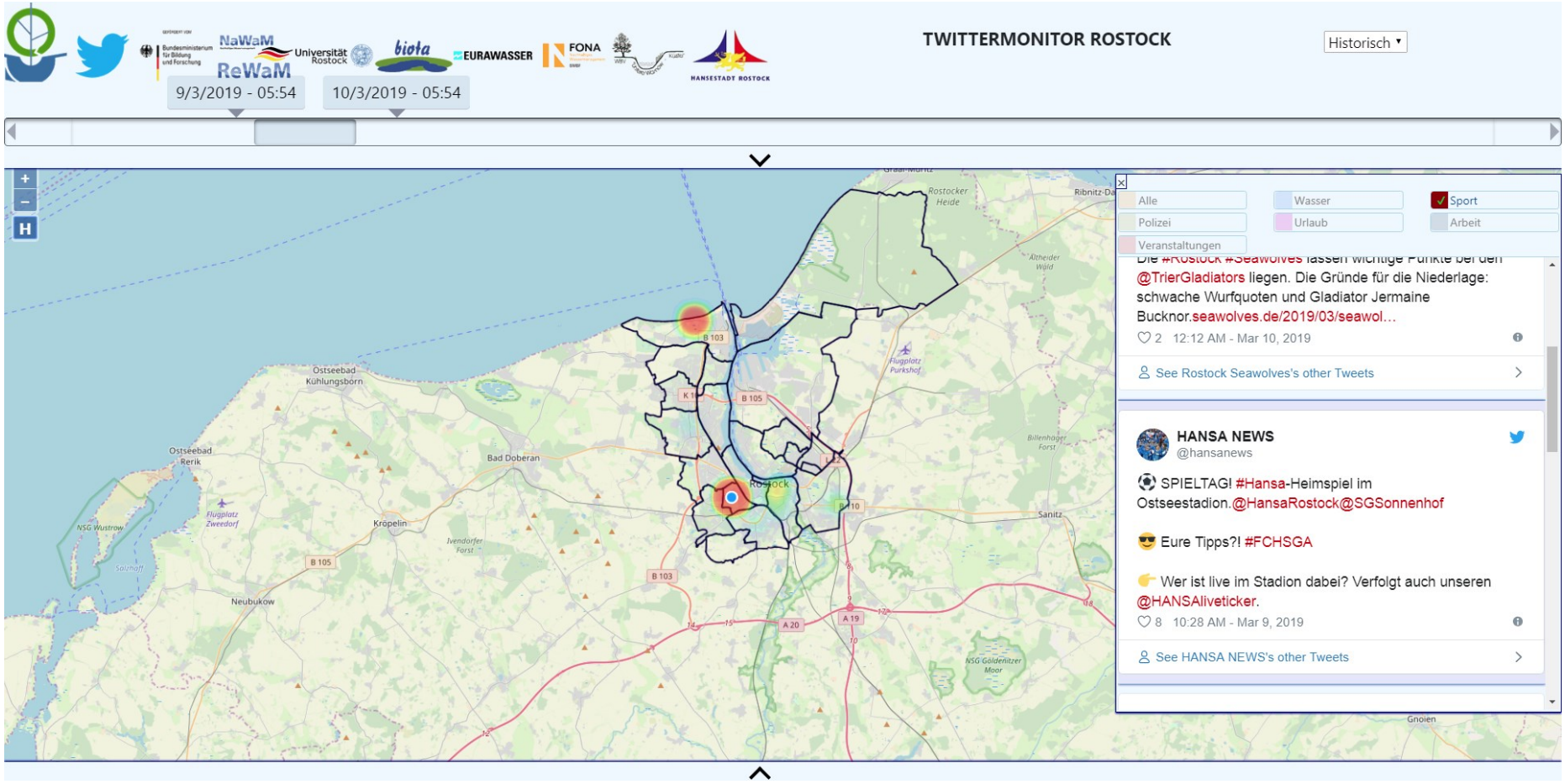


Abbildung 6-14: Screenshot des Twittermonitors.

7 Ergebnisse

„Was wir Ergebnisse nennen, ist nur der Anfang.“

Ralph Waldo Emerson (1803 - 1882), US-amerikanischer Geistlicher, Lehrer, Philosoph und Essayist

7.1 Allgemeine statistische Eigenschaften

Für die Auswertung der Tweets soll auf einen beschränkten Zeitraum von acht Wochen im letzten Sommer (06.08.2018 - 30.09.2018; KW 32 - KW 39) Bezug genommen werden. Insgesamt umfasst der Beobachtungszeitraum damit 56 Tage, da jedoch zwei Tage keine Erfassung erfolgte (29.09 - 30.09.2018) bleibt ein tatsächlicher Beobachtungszeitraum von 54 Tagen. In diesem Zeitraum liegen mehrere Veranstaltungen und Ereignisse (Hanse Sail, 26 Jahre Rostock Lichtenhagen, Rechtsextreme Ausschreitungen in Chemnitz mit häufigem Bezug zu Rostock, Saisonbeginn in der dritten Fußballliga), wodurch eine große Vielfalt an Nachrichten zu unterschiedlichen Themen und mit unterschiedlichen Stimmungen erfasst werden konnte.

Insgesamt sind in diesem Beobachtungszeitraum 29 771 Tweets aufgelaufen. Von diesen waren 60 % (17 896) nur auf Rostock als Ganzes bezogen, 12 % (3 662) bezogen sich auf Gazetteer-Elemente der Ebene 1, 10 % auf Ebene 2 (3 081) und 17 % (5 132) auf Ebene 3. Das entspricht im Mittel unter Berücksichtigung von Datenlücken 557 Nachrichten pro Tag. Das Minimum lag bei 199 Tweets pro Tag (15.08.2018). Dies rührt allerdings daher, dass bedingt durch Auswertungen zur Hanse Sail eine Datenlücke bei den Nachrichten von 17 Uhr bis 6 Uhr des Folgetages existiert. Daher ist der 28.09. mit 221 Tweets als nachrichtenärmster Tag anzunehmen. Das Maximum liegt bei 1 681 Nachrichten (22.09.2018), wohl durch die Großdemonstration in Rostock gegen die AfD verursacht (NDR 2018). Die Standardabweichung der täglichen Nachrichten beträgt 269 Tweets pro Tag.

Wesentlich für die Auswertung ist die Ermittlung der Genauigkeit. Hierzu wurden nach dem Zufallsprinzip zufällig 1 047 Tweets ausgewählt und ihnen eine Lokation zugeordnet. Damit die Verhältnisse weiterhin korrekt sind, wurden jeweils anteilig die Nachrichten nach ihrer Genauigkeitsebene ausgewählt, d. h. 643 Nachrichten bezogen sich nur auf Rostock, 116 auf Stadtteile, 93 auf Straßen sowie 195 auf POIs und exakte Gebäude. Dabei gibt es mehrere Möglichkeiten der Fehlklassifikation:

1. Identifikation einer falschen Lokation
2. Nicht-Identifikation einer genaueren Lokation
3. Nicht-Identifikation einer weiteren Lokation
4. Fehlende Lokation im Gazetteer, aber korrekte Einordnung

Das heißt, dass Nachrichten bedingt durch den Zuordnungsalgorithmus auch teilweise falsch bzw. richtig verortet sein können. In Tabelle 7-1 sind die jeweiligen Genauigkeiten je nach Ebene sowie die Anzahl der jeweils fehlerhaft zugeordneten Nachrichten dargelegt. Aus den Daten wird deutlich, dass mit zunehmender räumlicher Genauigkeit (d. h. Nummer der Ebene) die Gesamtgenauigkeit abnimmt. Werden noch 95 % der Nachrichten korrekterweise auf Stadtebene verortet, liegt der Wert bei POI’s und Gebäuden nur noch bei 55 %. Wesentliche Ursache für diese eklatante Abnahme der Zuordnungsgenauigkeit ist die falsche Verortung von Nutzern, deren hauptsächlicher Bezugspunkt mit Rostock angegeben ist. So sind 79 der 87 der mit Fehlertyp 1 identifizierten Nachrichten der Ebene

3 durch das Folgen bestimmter Nutzer verschuldet. Die hierbei am häufigsten verorteten Lokationen sind die Staatsanwaltschaft sowie der Hbf (Hauptbahnhof). Die Ursache ist hierbei wohl in diversen Polizeimeldungen respektive Verspätungsmeldungen von Zügen zu finden, die sich ebenfalls häufig auf Rostock bezogen haben. Dadurch wurden die zugehörigen Nutzer fälschlicherweise als Rostocker angenommen. Gleichzeitig sind dank der Followerliste lediglich zwölf Nachrichten zusätzlich auf dritten Ebene korrekt verortet worden. Daher kann der Nutzen hier durchaus in Frage gestellt werden.

Es ist auffällig, dass der Fehlertyp 1 bei den Tweets der Ebene 0 nicht vorkommt. Ursache dafür ist, dass hier keine fehlerhafte Zuordnung stattfinden kann, weil die Struktur des Filteralgorithmus keine falsche Zuordnung nur auf Stadtebene zulässt. Nachrichten, die zwar auf Rostock bezogen waren, aber nicht identifiziert worden sind, sind zudem nicht erfassbar, da hierfür der gesamte Twitter-Nachrichtenstream im Zeitraum zum Vergleich herangezogen und händisch verortet werden müsste. Es ist ersichtlich, dass der Algorithmus in der Lage ist, das Gros der Nachrichten komplett zu analysieren und auf der höchstmöglichen Tiefe zu verorten (Fehlertyp 2 und 3). Gleichzeitig wird aber auch deutlich, dass bei einem verbesserten Gazetteer die Genauigkeit auf Ebene 0 auf 98 % und auf Ebene 1 auf 79 % erhöht werden könnte. Daraus wird die besondere Bedeutung eines umfassenden und aktuellen Gazetters für die Genauigkeit der Verortung deutlich.

Tabelle 7-1: Lokationsgenauigkeit.

Art	Korrekt	Fehler-typ 1	Fehler-typ 2	Fehler-typ 3	Fehler-typ 4	Gesamt	Genau-igkeit
Alle Tweets	870	129	19	2	27	1 047	0.83
Tweets Ebene 0	609	0	13	0	21	643	0.95
Tweets Ebene 1	88	16	6	0	6	116	0.76
Tweets Ebene 2	66	26	0	1	0	93	0.71
Tweets Ebene 3	107	87	0	1	0	195	0.55

7.2 Spatio-Temporale Verteilung

Großer Vorteil der erhobenen Daten ist, dass diese sich spatio-temporal sehr gut visualisieren lassen. Dadurch treten Hotspots, Trends und einzelne Ereignisse räumlich in der Hansestadt Rostock heraus. Aus diesem Grund ist im Folgenden über einen Zeitraum von insgesamt acht Wochen (06.08.2018 - 30.09.2018; KW 32 - KW 39) für jede Woche eine gesonderte Kerndichte-Karte (Kernel Density*)* erstellt worden. Auf jeder dieser Karten ist die Dichte der Tweets, d. h. die Anzahl an Nachrichten pro km² dargestellt. Einbezogen wurden hierfür nur die Tweets, die mindestens auf Stadtteilgenauigkeit verortbar waren. Folglich entstehen dabei Hotspots in den Zentren häufig erwähnter Stadtteile. Daneben verdeutlichen sich aber auch singuläre Ereignisse. Um die Informationen weiter zu aggregieren, sind die Tweets für den gesamten Untersuchungszeitraum der Ebene eins und tiefer mit Hilfe anamorpher Karten auf Stadtbereichsgenauigkeit zusammengefasst worden. Die Dichtekarten sind im Anhang dargestellt.

Des Weiteren wurden die Nachrichten hinsichtlich ihrer zeitlichen Verteilung analysiert, da sich so Rückschlüsse auf bestimmte Ereignisse ziehen lassen. Es ist außerdem wichtig zu berücksichtigen, welche Zeit ein Ereignis von seinem Eintreten bis zu seiner Rezeption

auf Twitter benötigt. Um zu zeigen, welche Möglichkeiten sich raum-zeitlich kombiniert bieten, wird auf das Einzelereignis der AfD-Demonstration am 22.09.2018 Bezug genommen.

7.2.1 Temporale Verteilung

In Abbildung 7-1 sind die Tweets pro Tag und Thematik im Beobachtungszeitraum dargestellt. Dabei ist trotz einer mehrstündigen Datenlücke auch der 15.08.2018 für eine bessere Darstellung im Graphen abgebildet. Aus der Verteilung der einzelnen Nachrichten zu den einzelnen Themen werden vor allem einzelne Ereignisse deutlich erkennbar. Besonders sticht das Wochenende um den 22.09.2018 ins Auge, aber auch die Hanse Sail wird durch einen deutlichen Anstieg der Tweets im Themenbereich Veranstaltung gut sichtbar. Daneben zeigen sich die Fußballspiele sehr gut im Wochenverlauf durch den regelmäßigen Anstieg der Tweets zur Thematik Sport. Besonders fällt dabei das DFB-Pokalspiel zwischen Rostock und Stuttgart am 19.08.2018 auf. Dies spiegelt sich auch im Tagesgang der Nachrichten wieder (Abbildung 7-3). 14 Uhr respektive 21 Uhr steigt der Anteil der sportbezogenen Tweets an. Dies ist auf die jeweiligen Anstoßzeiten zurück zu führen. Auch das Großereignis der AfD-Demonstration am 22.09.2018 in der Rostocker Innenstadt ist deutlich am jeweiligen Nachrichtenaufkommen zu erkennen. Dies ist sowohl am Wochengang (Abbildung 7-2) als auch am Tagesgang (Abbildung 7-1) erkennbar. Es steigen dabei die Nachrichten mit Bezug zum Thema Sicherheit bzw. Veranstaltungen auf über 300 Tweets respektive 200 Tweets pro Tag an.

Die Tweets zur Thematik Urlaub und Arbeit folgen weder in der absoluten Anzahl der Nachrichten pro Tag bzw. pro Woche noch im Tagesgang einem besonderen Verlauf. Am Tagesgang der Nachrichten (Abbildung 7-3) wird vor allem deutlich, dass das Minimum der Twitteraktivität mit Bezug zu Rostock mit 4.8 Tweets pro Stunde bei 4 Uhr früh, das Maximum mit 32.6 Tweets pro Stunde bei 15 Uhr liegt.

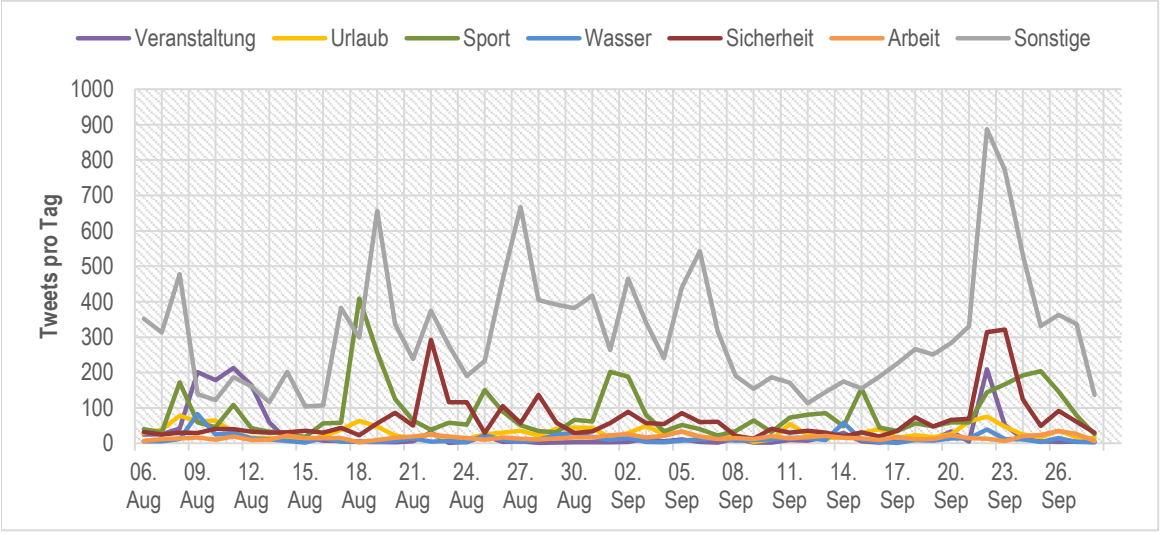


Abbildung 7-1: Absolute Anzahl der Tweets pro Tag im Untersuchungszeitraum (06.08.2019 – 30.09.2018).

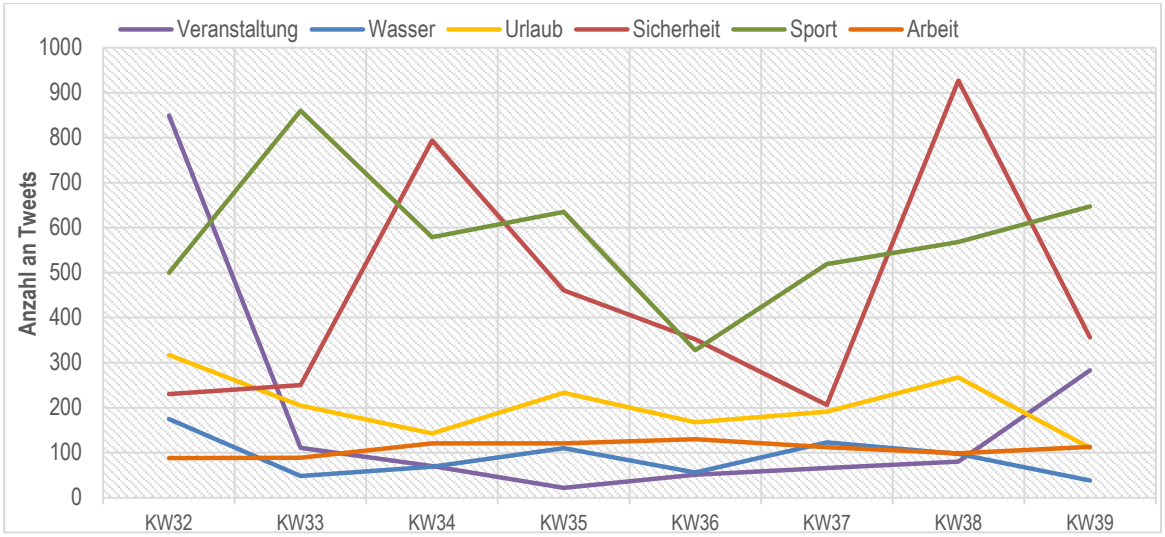


Abbildung 7-2: Absolute Anzahl der Tweets pro Woche im Untersuchungszeitraum (06.08.2018 – 30.09.2018).

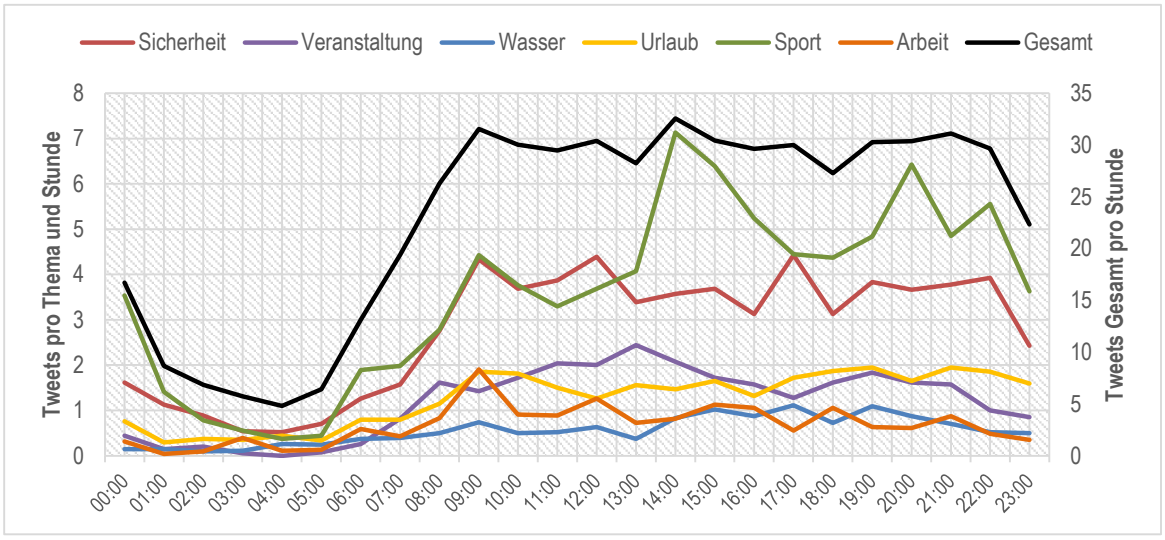


Abbildung 7-3: Mittlerer Tagesgang der Tweets im Untersuchungszeitraum (06.08.2018 - 30.09.2018).

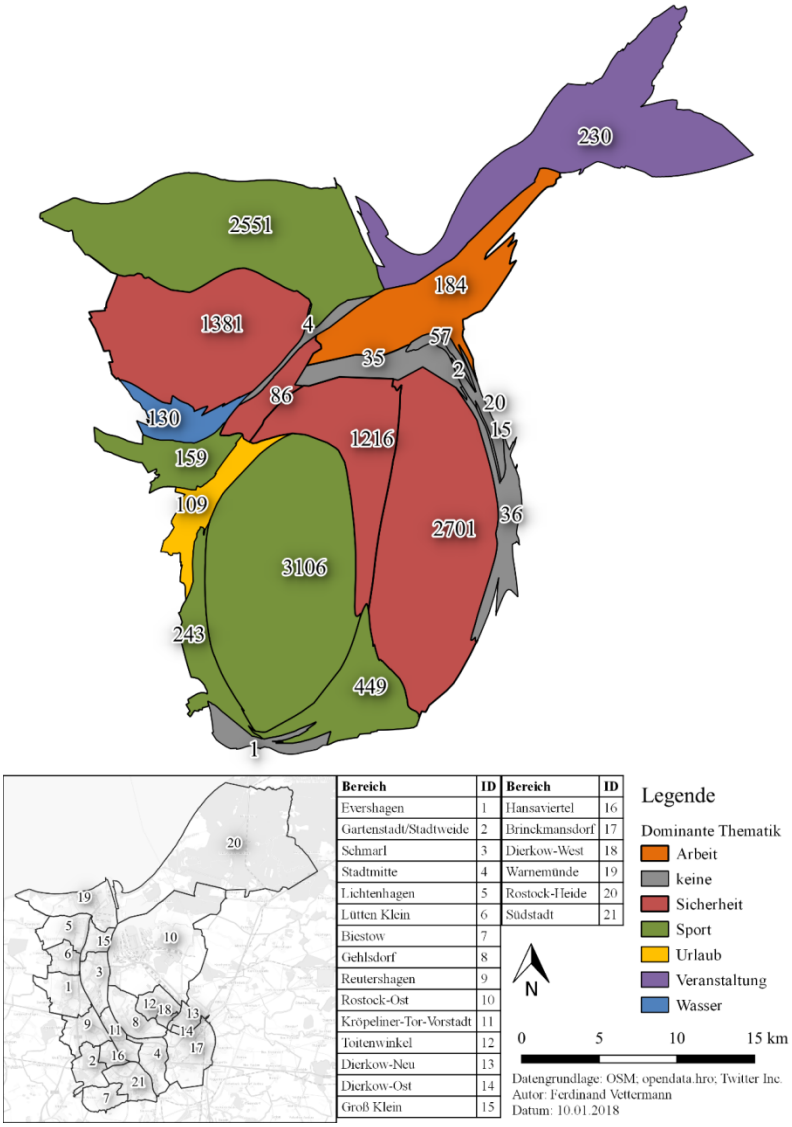
7.2.2 Räumliche Verteilung

Um die räumliche Verteilung der Nachrichten darzustellen, bietet sich eine anamorphe Darstellung mittels des QGIS-Plugins cartogram3 (20 Iterationen) der einzelnen Stadtbereiche der Hansestadt Rostock an. Diese Darstellung wird sowohl auf Basis der absoluten Anzahl der Nachrichten, als auch auf Basis der Follower, die einem jeweiligen Nutzer respektive seine Tweets direkt zu sehen bekommen, erstellt. Die Darstellung wird außerdem um die mit der pro Stadtbereich dominanten Thematik erweitert, sodass sich dadurch ein Bild von der Rezeption der Stadt Rostock auf Twitter zeichnen lässt.

Die Darstellung (Abbildung 7-4) der räumlichen Verteilung der absoluten Nachrichtenzahlen im Untersuchungszeitraum zeigt sehr deutlich, dass auf die Neubaugebiete sowohl im Osten als auch im Westen der Hansestadt kaum Bezug genommen wird. Einzige Ausnahme stellt hier Lichtenhagen dar. Eine große Aufmerksamkeit wird zudem dem Zentrum Rostocks als auch dem Szeneviertel KTV zuteil. Überraschend erscheint hier, dass die

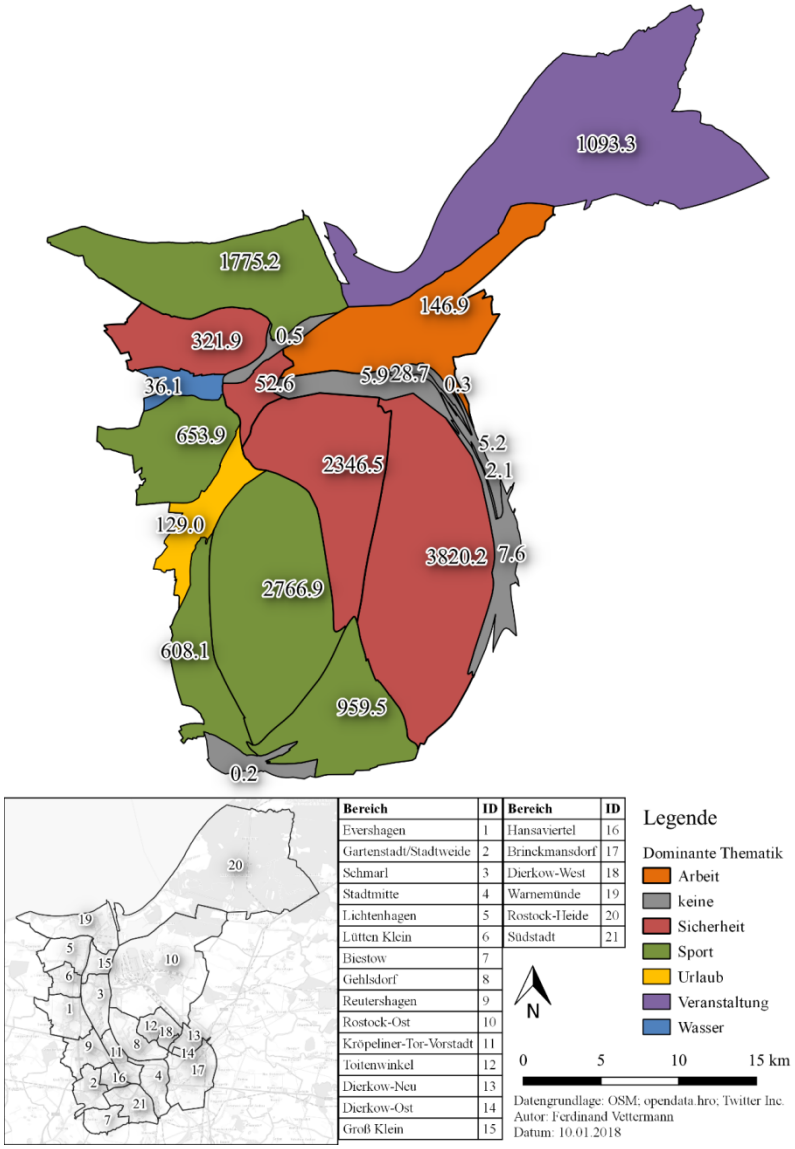
Tweets vor allem der Thematik Sicherheit zugeordnet werden. Besonders herauszuheben ist das Hansaviertel, welches bedingt durch das dort zu verortende Stadion der Stadtbereich ist, zu dem auf Twitter am häufigsten Bezug genommen wird.

Vergleichend dazu sind in Abbildung 7-5 die Stadtbereiche und deren dominierende Themen anhand der Summe der Followerzahlen anamorph dargestellt. Hierbei wird ersichtlich, dass zwar viel über Lichtenhagen getwittert wird, der gesamte Einfluss aber deutlich geringer ist. Im Gegenzug ist die Reichweite der Tweets bezüglich der Rostocker Heide als auch der Südstadt als deutlich größer einzustufen. Die östlichen Stadtteile sowie die Neubaugebiete im Westen finden allerdings weiter wenig Beachtung. Eine exakte Aufschlüsselung der einzelnen Anteile an den gesamten Nachrichten als auch der gesamten Followerzahl ist im Anhang zu finden.



Bereich	ID	Bereich	ID
Evershagen	1	Hansaviertel	16
Gartenstadt/Stadtweide	2	Brinckmansdorf	17
Schmarl	3	Dierkow-West	18
Stadtmitte	4	Warnemünde	19
Lichtenhagen	5	Rostock-Heide	20
Lütten Klein	6	Südstadt	21
Biestow	7		
Gehlsdorf	8		
Reutershagen	9		
Rostock-Ost	10		
Kröpeliner-Tor-Vorstadt	11		
Toitenwinkel	12		
Dierkow-Neu	13		
Dierkow-Ost	14		
Groß Klein	15		

Abbildung 7-4: Anamorphe Darstellung der Stadtbereiche auf Basis der aufsummierten Tweets der Ebene eins und tiefer im Untersuchungszeitraum (06.08.2018 - 30.09.2018). Anzahl der Tweets je Stadtbereich dargestellt.



Bereich	ID	Bereich	ID
Evershagen	1	Hansaviertel	16
Gartenstadt/Stadtweide	2	Brinckmansdorf	17
Schmarl	3	Dierkow-West	18
Stadtmitte	4	Warnemünde	19
Lichtenhagen	5	Rostock-Heide	20
Lütten Klein	6	Südstadt	21
Biestow	7		
Gehlsdorf	8		
Reutershagen	9		
Rostock-Ost	10		
Kröpeliner-Tor-Vorstadt	11		
Toitenwinkel	12		
Dierkow-Neu	13		
Dierkow-Ost	14		
Groß Klein	15		

Abbildung 7-5: Anamorphe Darstellung der Stadtbereiche auf Basis der aufsummierten Follower je Tweet der Ebene eins und tiefer im Untersuchungszeitraum (06.08.2018 - 30.09.2018). Anzahl der Follower je Stadtbereich in 10 tsd. dargestellt.

Betrachtet man die Tweets hinsichtlich der Bevölkerung in den einzelnen Stadtbereichen wird dieses Bild bestätigt. Mit 0.042 Tweets pro Einwohner und Woche ist die Nachrichtendichte im Hansaviertel am höchsten, gefolgt von Warnemünde (0.038 Tweets pro Einwohner und Woche) und Rostock Ost (0.018 Tweets pro Einwohner und Woche). Auch in den Innenstadtgebieten wird im Bezug zur Einwohnerzahl häufig getwittert. Deutlich unterrepräsentiert sind hingegen erneut die Neubaugebiete mit Ausnahme von Lichtenhagen (0.012 Tweets pro Einwohner und Woche). Alle Ergebnisse sind noch einmal im Anhang aufgeschlüsselt.

Die Bevölkerungszahl der Stadtbereiche korreliert dabei nicht mit den Tweets pro Person (r^2 0.0237; Pearson -0.15), das Durchschnittsalter und die Anzahl an Tweets sind schwach positiv korreliert (r^2 0.06; Pearson 0.24).

7.2.3 Sentimentanalyse

Neben der räumlichen Verteilung der Nachrichten ist auch deren Zuordnung zu den jeweiligen Stimmungen von Bedeutung. Insgesamt sind die Tweets zur Hansestadt Rostock im Mittel mit 61 % eher positiv zu werten. Dafür wurden die Tweets der Ebene 1 und ihr Sentiment jedem Stadtbereich zugeordnet. Neben der gesamten Zuordnungswahrscheinlichkeit sind dazu die Zuordnungen des Sentiments nach den Wahrscheinlichkeiten bereits vorselektiert worden. Konkret bedeutet das, dass alle Tweets, bei denen ein positives Sentiment respektive negatives Sentiment mit > 95 %, > 90 %, > 85 % und 80 % zugeordnet werden konnte, entsprechend mit 1 (positiv), 0.5 (neutral) und 0 (negativ) bestimmt worden ist. Das Ergebnis ist in Tabelle 7-2 dargestellt.

Als besonders positiv treten Dierkow-West (87 % positiv) und Biestow (70 % positiv) hervor. Allerdings ist zu beachten, dass hier jeweils nur zwei respektive ein Tweets klassifiziert worden sind, weshalb die Aussagekraft anzuzweifeln ist. Betrachtet man nur die Stadtbereiche, in denen mindestens 1 000 Tweets verortet worden sind, sticht insbesondere Warnemünde (66 % positiv) sowie die KTV (63 % positiv) heraus. Tweets mit Bezug zu Lichtenhagen scheinen im Gegensatz eher ein leicht negatives Sentiment (49 % positiv) zu besitzen. Dieses Bild zeigt sich auch bei der Vorselektion der positiven und negativen Tweets. Vor allem mit Bezug zu Warnemünde und zum Hansaviertel scheinen die Tweets bei einer Zuordnungswahrscheinlichkeit von 85 % bzw. 80 % mehrheitlich positiv auszufallen. Setzt man die Schwelle für die Zuordnung höher, sind keine klaren Tendenzen mehr zu erkennen.

Zur Visualisierung wurde auf Basis von Moran‘s I mittels ArcMap eine Hotspot-Analyse durchgeführt (WESTERHOLT et al. 2016, CLIFF & ORD 1969). Eingang fanden hier nur die Tweets, welche mindestens auf Ebene 1 verortet worden sind. Dabei ist zum einen die Verteilung von Hot- (positives Sentiment) und Coldspots (negatives Sentiment) anhand der jeweiligen Wahrscheinlichkeiten der Zuordnung in Abbildung 7-6 dargestellt. Zum anderen sind auch für die jeweiligen vorselektierten Zuordnungswahrscheinlichkeiten die Hot- und Coldspots berechnet und dargestellt worden. Die Ergebnisse sind für eine Zuordnungswahrscheinlichkeit von 95 % in Abbildung 7-7, von 90 % in Abbildung 7-8, von 85 % in Abbildung 7-9 und von 80 % in Abbildung 7-10 visualisiert.

Aus den Abbildungen wird deutlich, dass Warnemünde in jedem Fall als Hotspot zu bewerten ist. Daneben ist der nördliche Teil der Innenstadt und der Stadthafen sowohl bei der Analyse mittels der Gesamtwahrscheinlichkeit als auch bei einer Zuordnungswahrscheinlichkeit > 90 % als Hotspot zu identifizieren. Über > 95 % Zuordnungswahrscheinlichkeit ist das Gebiet als nicht signifikant zu betrachten. Gleichzeitig ist ein bedeutender Coldspot in Rostock-Lichtenhagen sowie im südlichen Bereich der Stadtmitte und der Südstadt zu verorten.

Deutlich differenzierter gestaltet sich die Zuordnung im Bereich des Hansaviertels. Hier wird bei der Zuordnung nach Wahrscheinlichkeiten ein Hotspot, bei einer Zuordnungswahrscheinlichkeit > 95 % kein signifikanter Hot- oder Coldspot und bei einer Zuordnungswahrscheinlichkeit von 90 % respektive 85 % ein Coldspot angenommen.

Tabelle 7-2: Zuordnung des Sentiments je Stadtbereich - > 0.5 eher positiv, <0.5 eher negativ.

Stadtbereich	Zuordnungswahr-scheinlichkeit Gesamt	Zuordnungswahr-scheinlichkeit > 95 %	Zuordnungswahr-scheinlichkeit > 90 %	Zuordnungswahr-scheinlichkeit > 85 %	Zuordnungswahr-scheinlichkeit > 80 %	Anzahl Tweets
Schmarl	0.68	0.51	0.58	0.61	0.64	86
Stadtmitte	0.57	0.50	0.51	0.53	0.55	2 552
Südstadt	0.61	0.51	0.53	0.57	0.61	431
Toitenwinkel	0.69	0.52	0.58	0.63	0.67	57
Lütten Klein	0.68	0.51	0.54	0.58	0.63	123
Reutershagen	0.69	0.51	0.54	0.60	0.64	109
Rostock-Heide	0.65	0.50	0.55	0.57	0.61	218
Rostock-Ost	0.64	0.51	0.55	0.59	0.61	168
Groß Klein	0.59	0.50	0.50	0.50	0.50	4
Hansaviertel	0.62	0.50	0.52	0.55	0.60	2 808
Kröpeliner-Tor-Vor-stadt	0.63	0.51	0.52	0.55	0.59	1 134
Lichtenhagen	0.49	0.50	0.49	0.50	0.50	1 376
Dierkow-West	0.87	0.50	0.50	0.75	1.00	2
Evershagen	0.64	0.51	0.55	0.58	0.61	158
Gartenstadt/Stadt-weide	0.57	0.52	0.53	0.55	0.57	241
Gehlsdorf	0.62	0.50	0.49	0.51	0.54	35
Biestow	0.70	0.50	0.50	0.50	0.50	1
Brinckmansdorf	0.66	0.51	0.53	0.54	0.58	36
Dierkow-Neu	0.57	0.50	0.53	0.55	0.58	20
Dierkow-Ost	0.62	0.50	0.50	0.57	0.63	15
Warnemünde	0.66	0.51	0.54	0.58	0.62	2 385

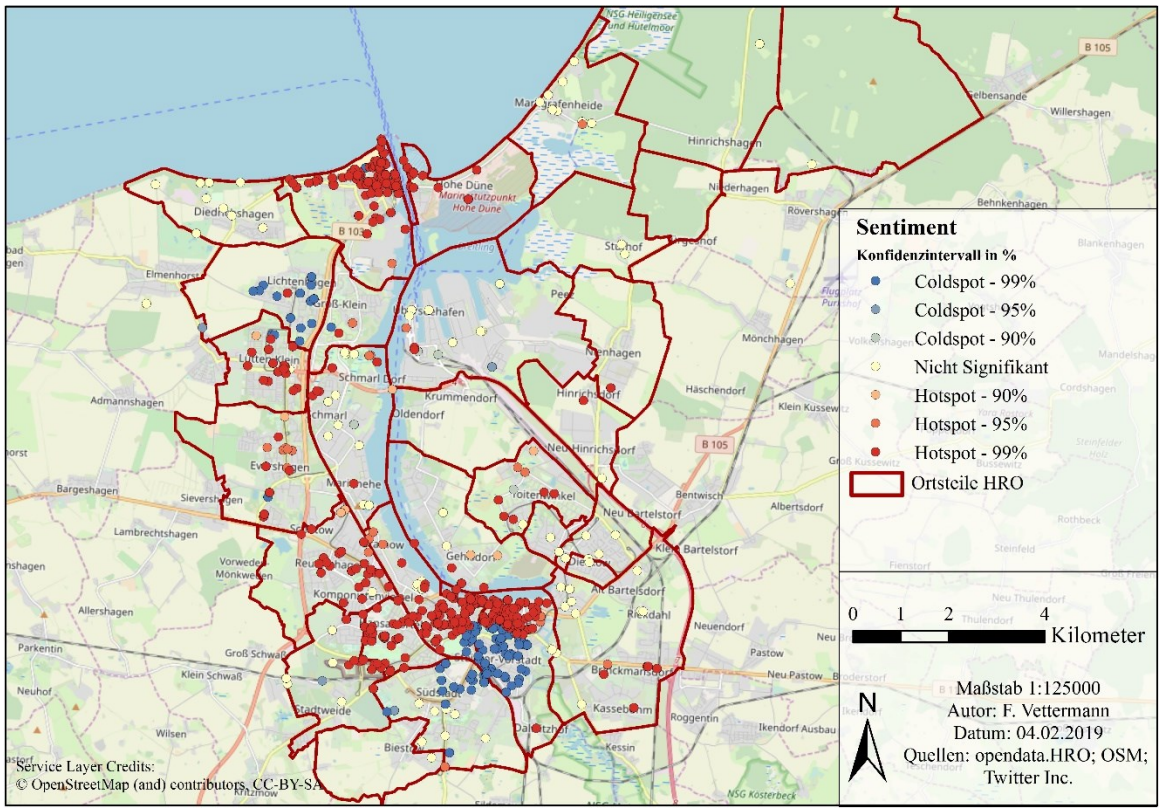


Abbildung 7-6: Signifikante Hot- (positives Sentiment) und Coldspots (negatives Sentiment) berechnet mittels Morans I aus der Gesamtwahrscheinlichkeit der Sentimentzuordnung.

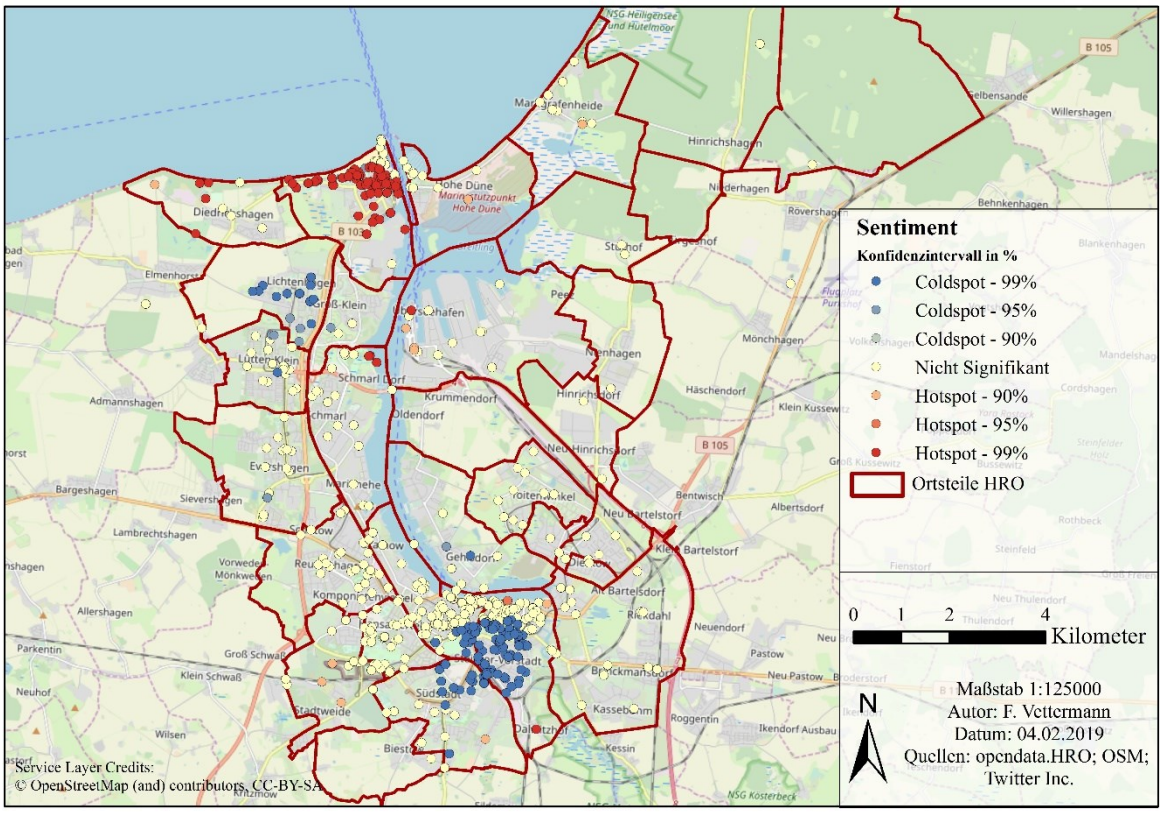


Abbildung 7-7: Signifikante Hot- (positives Sentiment) und Coldspots (negatives Sentiment) berechnet mittels Morans I aus der Zuordnungswahrscheinlichkeit von 95 % zu positiven bzw. negativen Tweets.

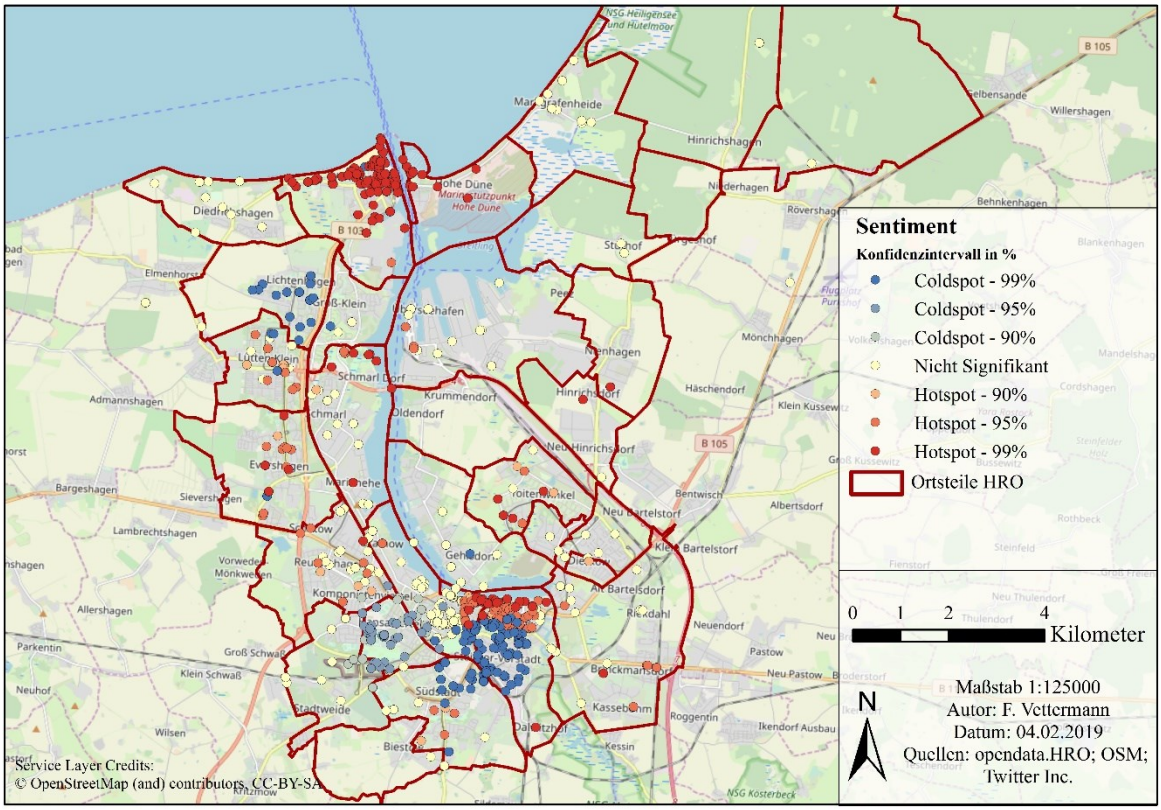


Abbildung 7-8: Signifikante Hot- (positives Sentiment) und Coldspots (negatives Sentiment) berechnet mittels Morans I aus der Zuordnungswahrscheinlichkeit von 90 % zu positiven bzw. negativen Tweets.

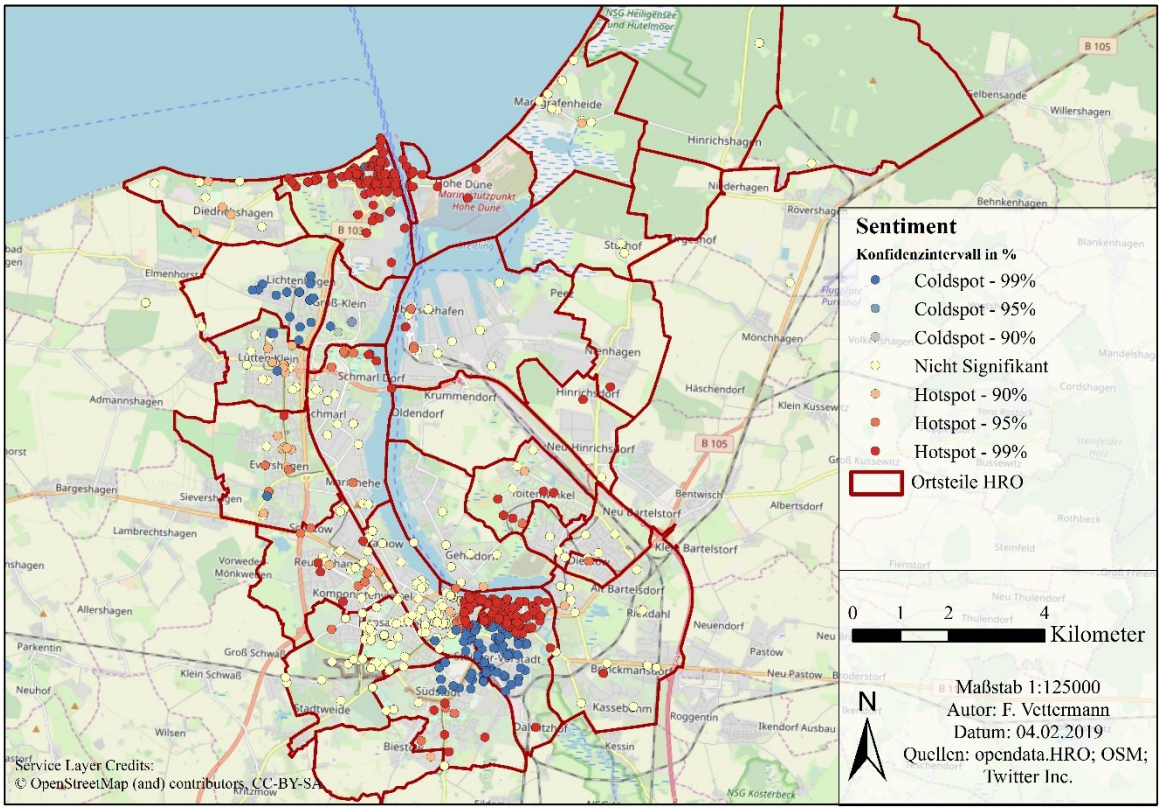


Abbildung 7-9: Signifikante Hot- (positives Sentiment) und Coldspots (negatives Sentiment) berechnet mittels Morans I aus der Zuordnungswahrscheinlichkeit von 85 % zu positiven bzw. negativen Tweets.

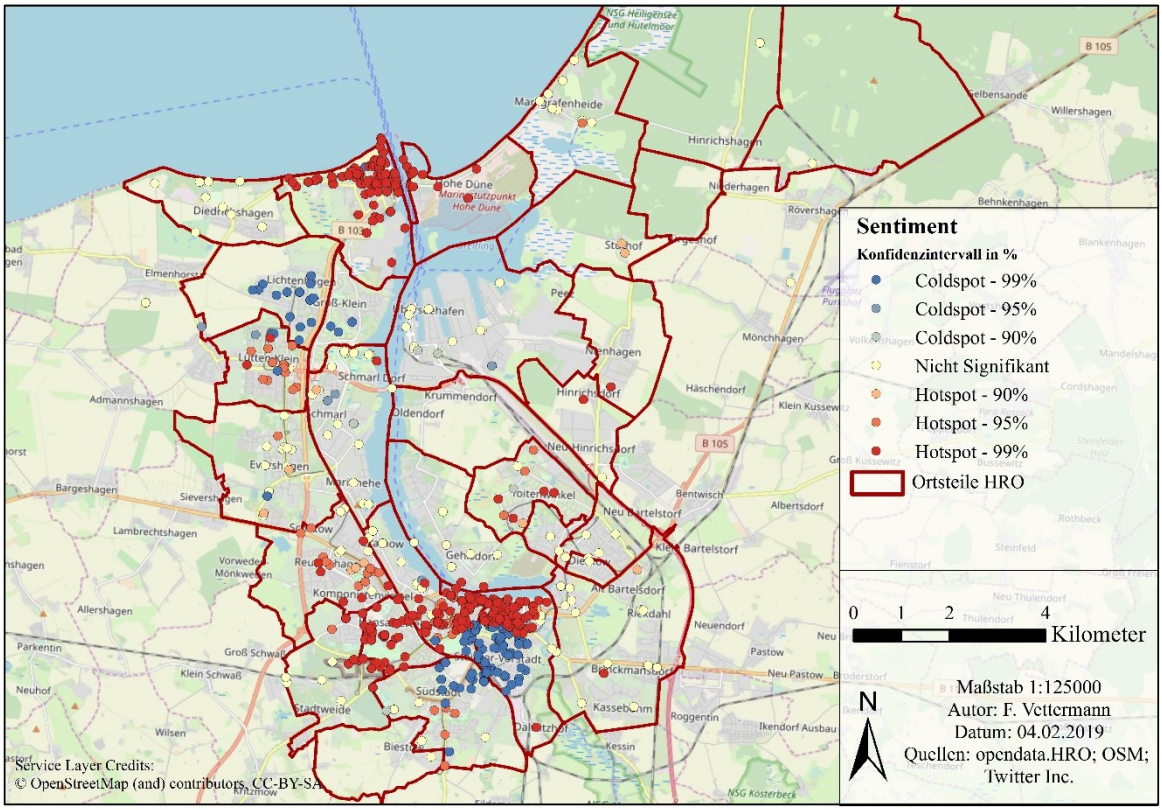


Abbildung 7-10: Signifikante Hot- (positives Sentiment) und Coldspots (negatives Sentiment) berechnet mittels Morans I aus der Zuordnungswahrscheinlichkeit von 80 % zu positiven bzw. negativen Tweets.

Tabelle 7-3: Mittleres Sentiment je Thema.

Thema	Mittleres Sentiment
Sicherheit	0.58
Wasser	0.62
Urlaub	0.62
Sport	0.63
Veranstaltung	0.70
Arbeit	0.70

Neben der räumlichen Verteilung ist nachfolgend das Sentiment in Bezug zu den einzelnen Themen dargestellt (Tabelle 7-3). Daraus geht hervor, dass Tweets mit Bezug zur Thematik Sicherheit deutlich negativer ausfallen als Tweets mit Bezug zu Veranstaltungen und Arbeit. Gerade bei der Thematik Arbeit überrascht das positive Sentiment, ist jedoch vermutlich darin begründet, dass Jobangebote häufig mit positiven Attributen untermalt sind.

7.2.4 Kombinierte raum-zeitliche Analyse

Für eine raum-zeitliche Analyse wurden zum einen die Tweets des gesamten Untersuchungszeitraums in vier Sechs-Stunden-Abschnitte eingeteilt (0 – 6 Uhr; 6 – 12 Uhr; 12 – 18 Uhr; 18 – 0 Uhr). Anschließend wurde eine Kerndichteschätzung aller Tweets ab der Ebene 1 mittels ArcMap durchgeführt (Suchradius 500 m, 10 m Auflösung; vgl. SILVERMAN 1986 für die Beschreibung des Kernels). In Abbildung 7-11 ist klar zu erkennen, dass das

Nachrichtenaufkommen in den Nachtstunden deutlich geringer ausfällt als in den späteren Zeitfenstern. Die Hotspots beschränken sich auf das Ostseestadion, den Hauptbahnhof, Lichtenhagen und Warnemünde. Bis Mittag zeichnen sich hingegen schon deutlichere Hotspots am Alten Strom in Warnemünde sowie in der Rostocker Innenstadt ab (Abbildung 7-12). Am Nachmittag steigt die Nachrichtendichte weiter an (Abbildung 7-13), um schließlich in den Abendstunden ihren Höhepunkt zu erreichen (Abbildung 7-14). Vor allem das Ostseestadion tritt in der zweiten Tageshälfte als Hotspot in Erscheinung. In Warnemünde hingegen nimmt die Nachrichtendichte gegen Abend ab, verteilt sich aber weiter entlang der Strandpromenade. In der Innenstadt wandert der Hotspot vom Neuen Markt in Richtung des Doberaner Platzes und damit in die KTV.

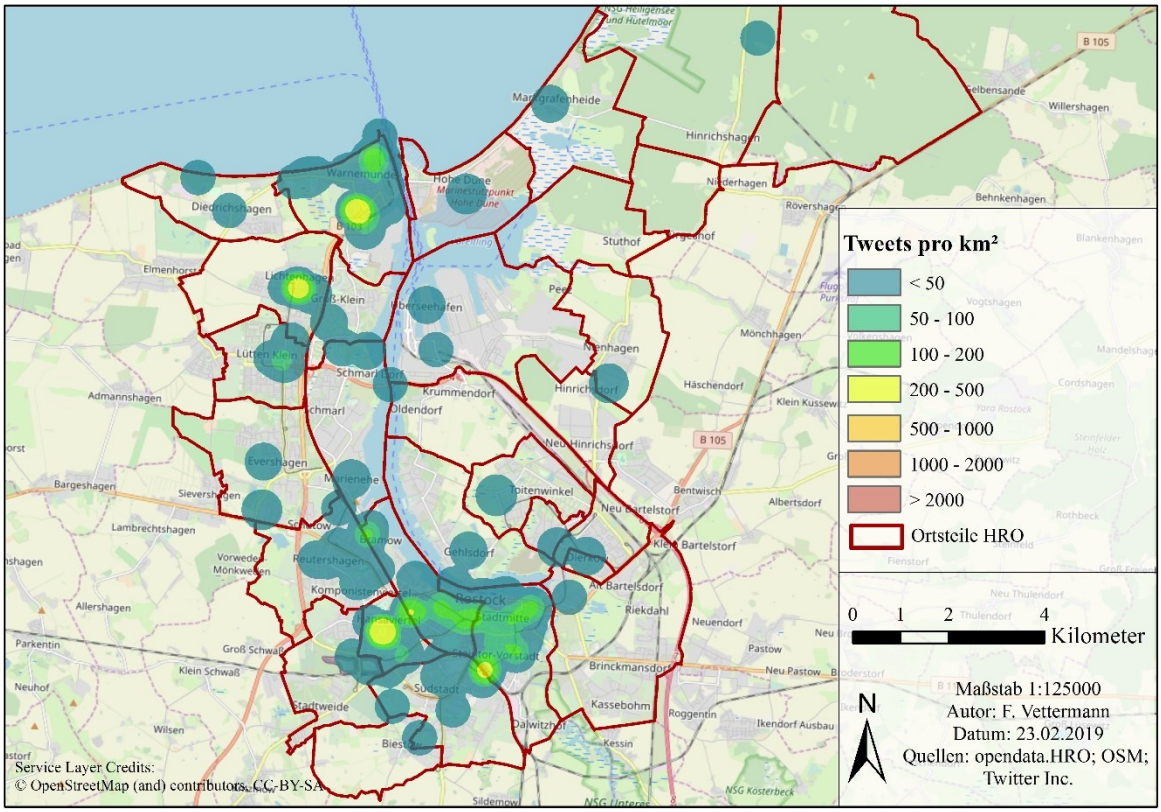


Abbildung 7-11: Kernel Density anhand aller Tweets im Untersuchungszeitraum (06.08.2018 - 30.09.2018) von 0 - 6 Uhr.

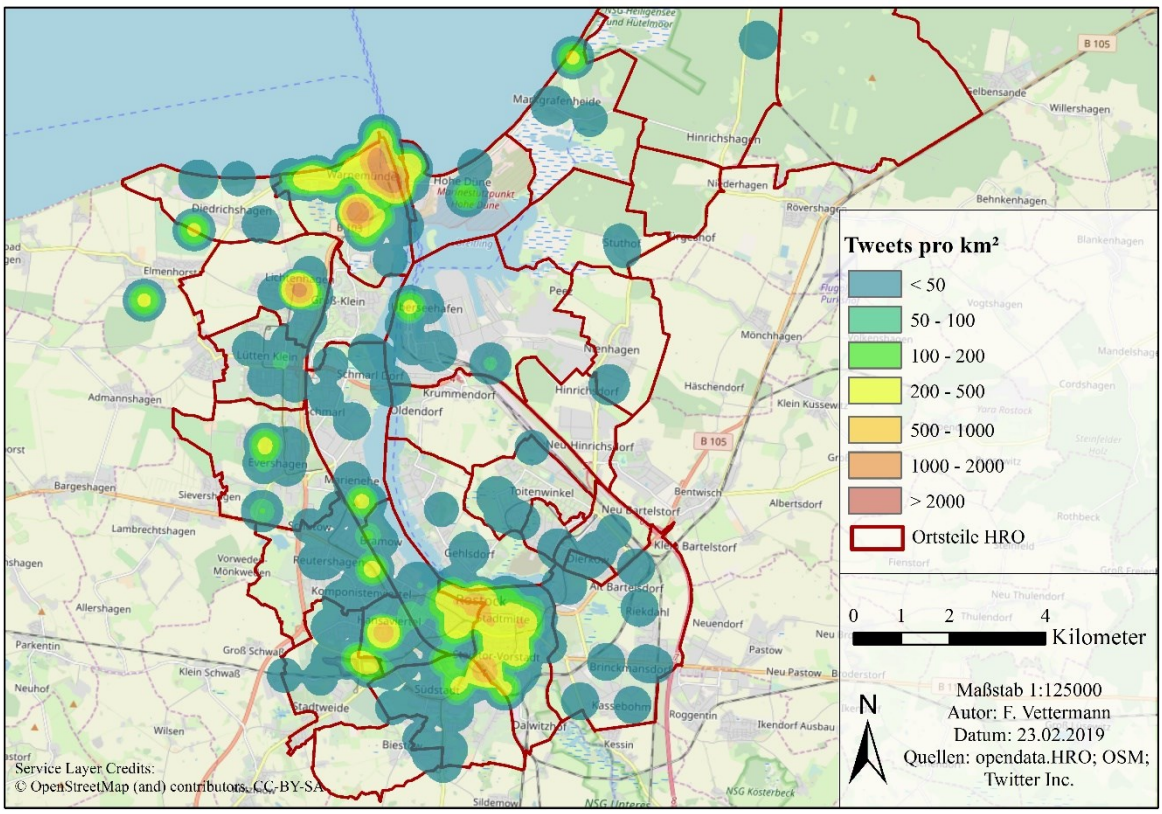


Abbildung 7-12: Kernel Density anhand aller Tweets im Untersuchungszeitraum (06.08.2018 - 30.09.2018) von 6 - 12 Uhr.

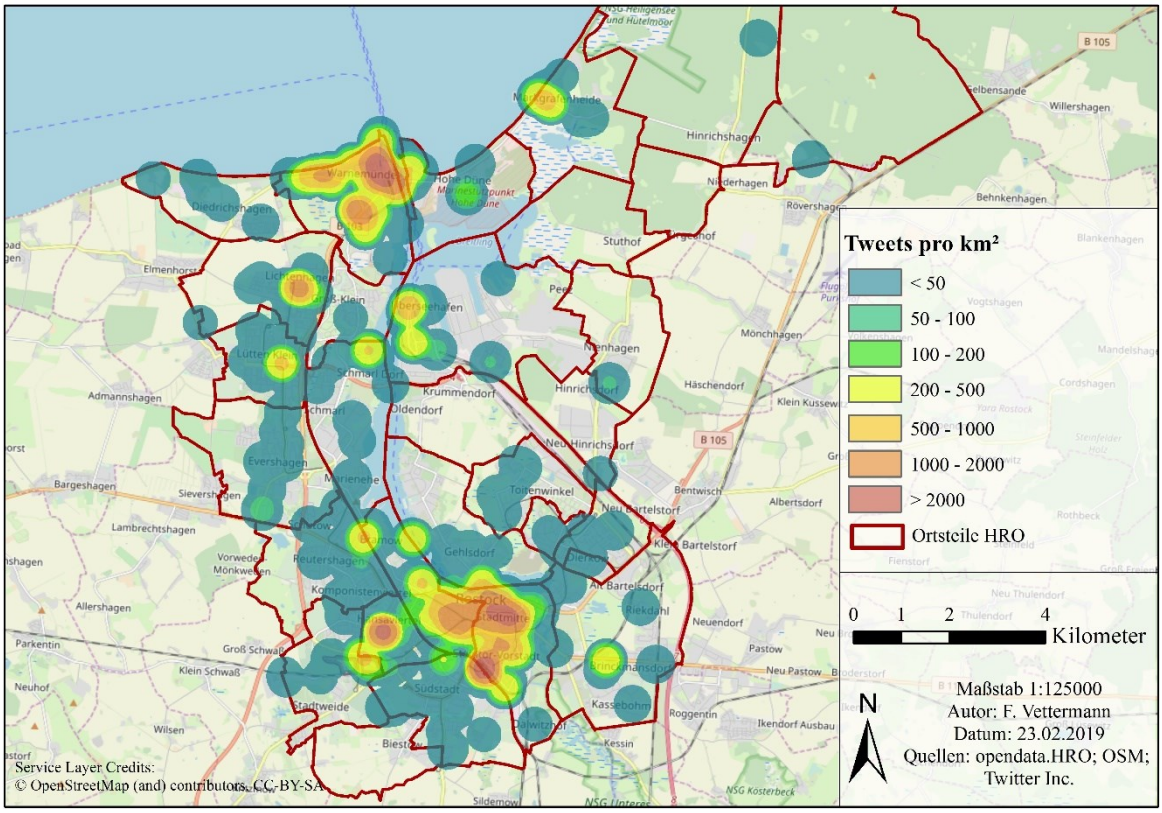


Abbildung 7-13: Kernel Density anhand aller Tweets im Untersuchungszeitraum (06.08.2018 - 30.09.2018) von 12 - 18 Uhr.

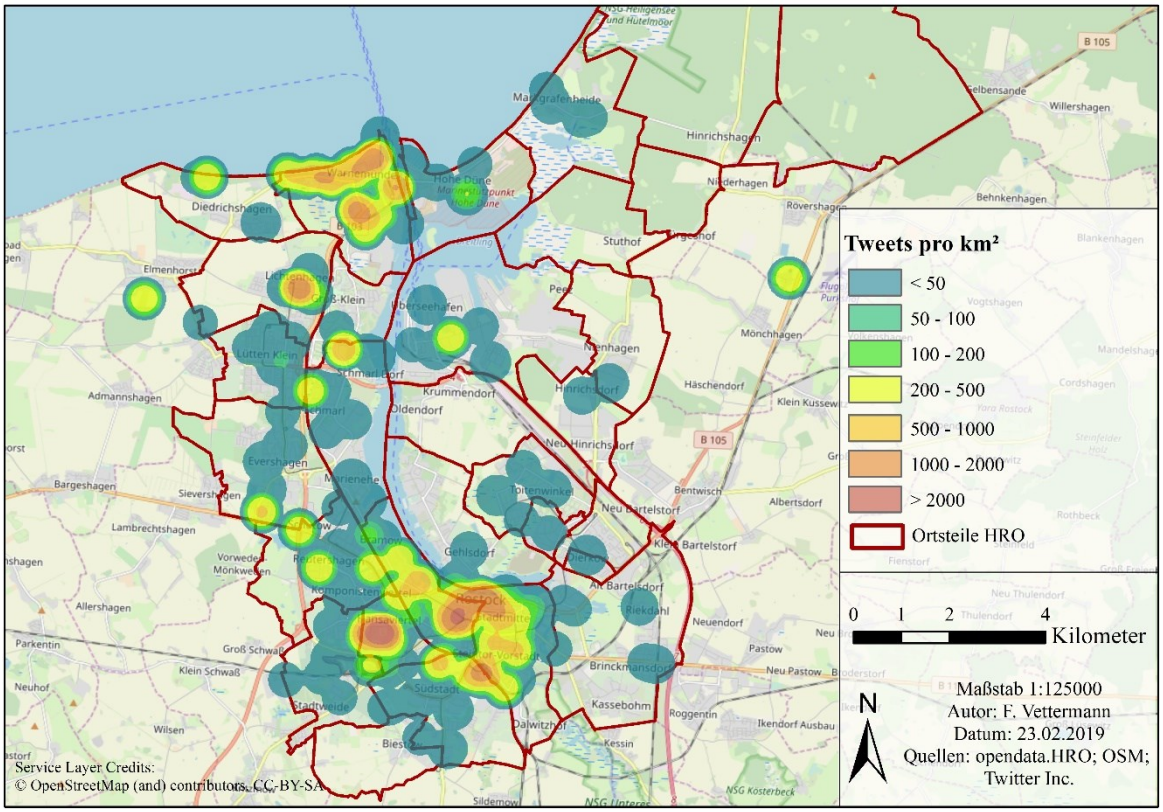


Abbildung 7-14: Kernel Density anhand aller Tweets im Untersuchungszeitraum (06.08.2018 - 30.09.2018) von 18 - 0 Uhr.

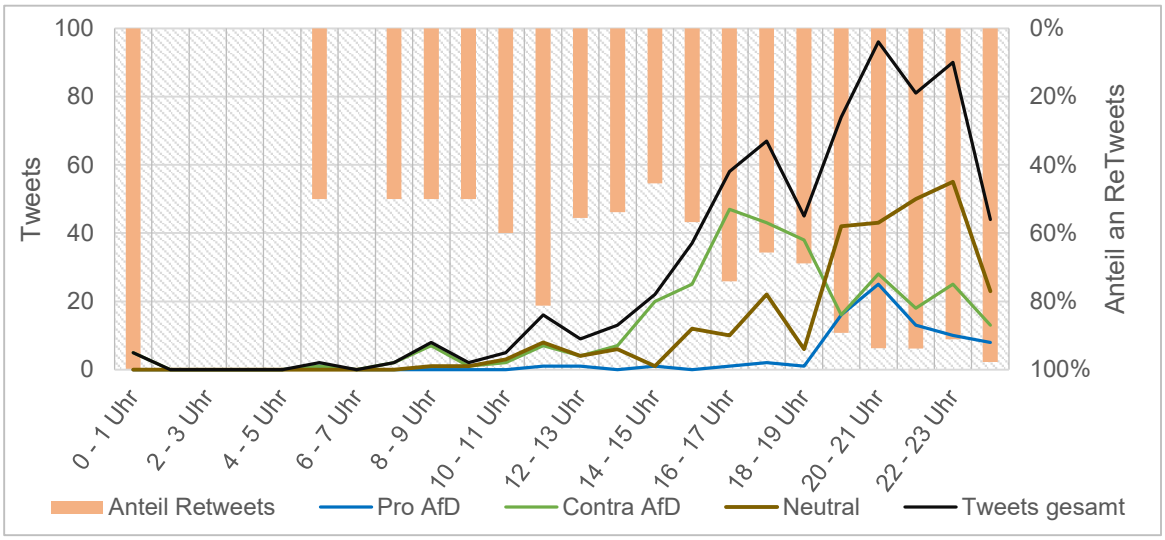


Abbildung 7-15: Zeitlicher Verlauf der Tweets mit Bezug zur AfD-Demonstration am 22.09.2018.

Die gezeigte spatio-temporale Aufschlüsselung ist allerdings recht grob. In der Kartographie ist die hochaufgelöste Darstellung der Zeit ein vieldiskutiertes Problem. Daher bietet es sich an, auf ein besonderes Ereignis zurückzugreifen: nämlich die AfD-Demonstration am 22.09.2018. Zum ersten kann hier ein zeitlicher Verlauf einzelner Ereignisse direkt nachvollzogen werden, zum zweiten ist die Analyse durch ein entsprechend hohes Nach-

richtenaufkommen gestützt. Und zum dritten bezieht sich die Analyse auf eine aktuell intensiv diskutierte Thematik, nämlich die Beziehung zwischen der neuen Rechten und den Sozialen Medien.

Hierfür wurden alle Nachrichten vom 22.09.2018 mittels folgender Schlagwörter durchsucht und anschließend händisch den Positionen pro AfD, contra AfD oder neutral zugeordnet:

afd, demo, rechts, links, höcke, antifa, hro2209, nazis, aufzug, blockade, vermummt

Die Auswahl der Schlagworte fand auf Basis der manuellen Sichtung der Tweets vom 22.09.2018 statt. Insgesamt umfasst die Abfrage 676 Nachrichten. Ab ca. 9 Uhr steigt dabei das Nachrichtenaufkommen kontinuierlich über den Tag an, bis es nach 20 Uhr seinen Höhepunkt erreicht (Abbildung 7-15). Während der Vorbereitungen und dem Demonstrationsgeschehen selbst (die Gegenveranstaltungen begannen 14 Uhr, 18 Uhr fand die Rede von Björn Höcke auf Seiten der AfD statt) ist der Anteil an Retweets bei etwas über 50 % anzusiedeln. Erst mit dem Ende der Demonstrationen werden ab 19 Uhr vermehrt Nachrichten geretweetet und der Anteil steigt auf über 90 % an. Auffällig ist, dass erst mit Ende der Demonstrationen die Nachrichten deutlich häufiger pro AfD ausfallen, während des Demonstrationsgeschehens ist der überwiegende Anteil der Nachrichten gegen die Positionen der AfD gerichtet. Hierbei wird vor allem der Tweet „Was die @tagesschau verschweigt […]“ geteilt. Betrachtet man nur die parteiischen Tweets, sind 20 % pro AfD, 80 % contra AfD positioniert. Dies ist sicherlich auch mit der jeweiligen Teilnehmeranzahl zu begründen. 700 Demonstranten (15 %) waren auf Seite der AfD vorzufinden, ca. 4 000 auf der Gegenseite (85 %) (NDR 2018).

Insgesamt sind 104 Nachrichten auf der Ebene zwei oder drei mit Bezug zu den Demonstrationen zwischen 14 und 20 Uhr verortet worden. Zur Darstellung wurde aus diesen ein Raum-Zeit-Würfel (Zeitschritt: 30 min, Distanz-Intervall 50 m) erstellt und entsprechend dargestellt (Abbildung 7-16). Dabei symbolisiert die Höhe den Zeitraum, der Ort den jeweiligen Bezugspunkt der Tweets sowie die Farbe und Größe der Würfel die Anzahl der Nachrichten im jeweiligen Zeitraum.

Es wird deutlich, dass sich der Verlauf der Demonstration sehr gut anhand einzelner Tweets verfolgen lässt. So startet die Gegendemonstration 14 Uhr am Doberaner Platz (1), führt weiter am Kröpeliner-Tor (2) entlang zum Uni-Platz (3), wo sich zwei Demos verbinden. Schließlich endet sie am Steintor (4), wo sie nur wenige Meter von der AfD-Demonstration getrennt ist. Zeitgleich werden Barrikaden an der Deutschen Med (5) errichtet. Durch diese wird der Demonstrationszug der AfD zur Umkehr gezwungen und findet schließlich seinen Abschluss am Neuen Markt (6).

Betrachtet man die einzelnen Tweets, wird die Reaktionsgeschwindigkeit der Nutzer auf Twitter bzw. die Geschwindigkeit, mit der neue Informationen geteilt werden, deutlich. Dabei ist die Polizei Rostock über ihren Twitter Account sehr bemüht, die Öffentlichkeit auf aktuellem Stand zu halten. Insgesamt wurden durch die Polizei elf Tweets mit Bezug zur Demonstration geteilt, wovon kein einziger ein retweet war, d. h., dass es sich höchstwahrscheinlich um originären Content gehandelt hat. In Abbildung 7-17 sind dabei die Accounts dargestellt, die den meisten originären Inhalt geteilt haben. Neben der Polizei ist hier auch der als weitgehend neutral einzuordnende Account „Isabel Lerch“ zu finden. Darüber hinaus finden sich mit „Rostock Hilft“ und „Rostock Nazifrei“ noch zwei Institutionen, die aktiv an der Organisation der Demonstrationen beteiligt waren und über Twitter versuchen, Demonstranten zu mobilisieren. Allerdings wird ersichtlich, dass auch ein großer Anteil individueller Nutzer die Demonstrationen zu beobachten und neueste Informationen zu diesen

zu teilen scheinen. Des Weiteren zeigt sich, dass das Gros der Nutzer nur zwei oder einen originären Tweet bezüglich abgesetzt hat.

Am besten lässt sich die Geschwindigkeit der Rezeption der Ereignisse mittels des Livetickers der Ostseezeitung[47] ermitteln. Hierzu sind in Abbildung 7-18 vergleichend die Meldungen verschiedener Ereignisse des Demonstrationstages aufgetragen. Es wird deutlich, dass Twitter z. T. deutlich schneller als der Liveticker der Ostseezeitung ist. So ist die Meldung, dass die Deutsche Med erklommen wird, bereits 14.24 Uhr auf Twitter und erst 15.56 Uhr bei der Ostseezeitung zu finden. Bei den meisten anderen Meldungen beträgt der Zeitunterschied nur wenige Minuten, teilweise ist auch hier Twitter schneller. Nicht betrachtet wurden Meldungen, die bei der Ostseezeitung eindeutig ersichtlich direkt aus Twitter heraus gemeldet worden sind. Dies unterstreicht allerdings die Bedeutung des Sozialen Netzwerks hinsichtlich der Informationsverbreitung und verdeutlicht dessen Schnelligkeit.

Abbildung 7-16: Raum-zeitlicher Verlauf der Demonstrationen am 22.09.2018.

[47] http://www.ostsee-zeitung.de/Mecklenburg/Rostock/LIVE-Ticker-Bjoern-Hoecke-tritt-bei-AfD-Demo-in-Rostock-auf-Fuenf-Gegendemos-angekuendigt

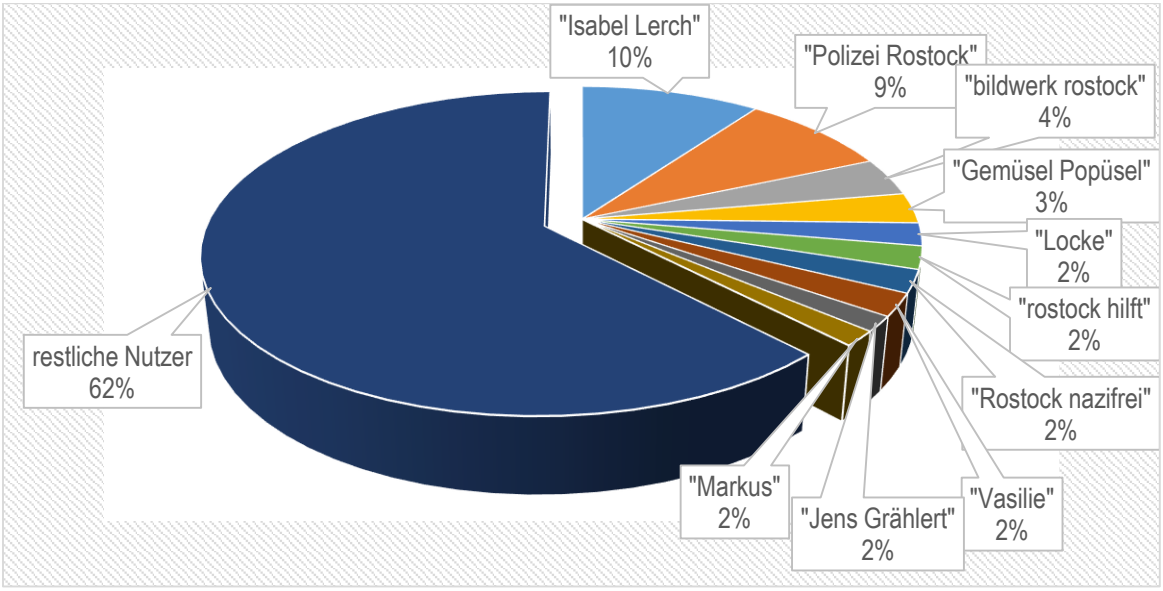


Abbildung 7-17: Anteil der zehn Accounts mit den meisten originären Tweets.

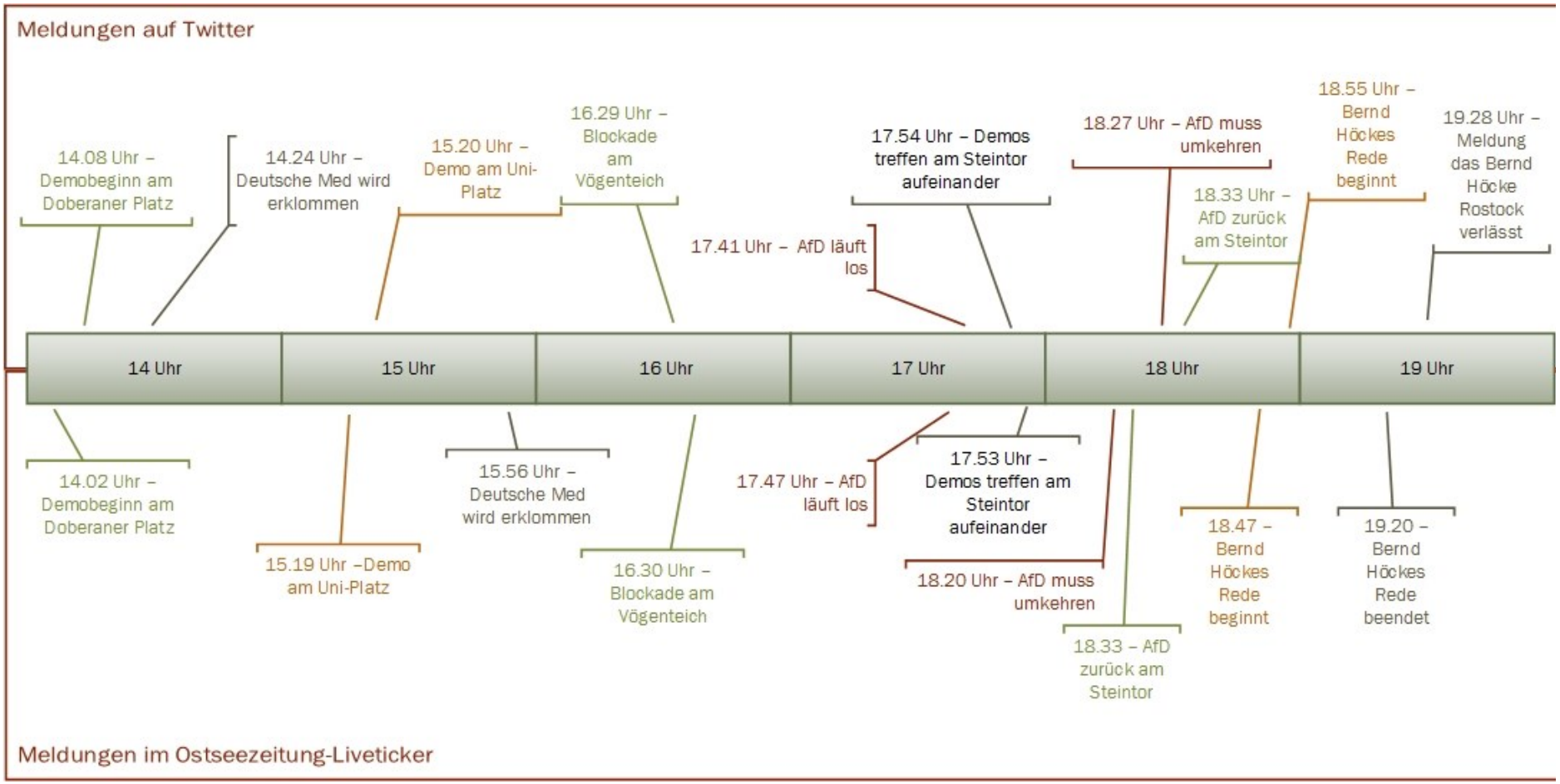


Abbildung 7-18: Rezeption und offizielle Meldung verschiedener Ereignisse der Demonstrationen am 22.09.2018.

7.3 Trending Topics

7.3.1 TF-IDF

Neben der Zuordnung der Tweets zu vordefinierten Themen sind auch die Schlagworte ermittelt worden, die im Untersuchungszeitraum getrendet sind. Bedingt durch die vorhergehende Textkorrektur sind die Trends jedoch teilweise schwer zuzuordnen. Aus diesem Grund sind nur eindeutig zuordenbare und aussagekräftige Trends in Tabelle 7-4 dargestellt.

Aus den Daten lassen sich sehr eindeutig dominante Themen in den jeweiligen Kalenderwochen ablesen. Im gesamten Zeitraum fallen die Schlagworte „Domscheit“ (Bezug zu einer vermissten Frau[48]), der Archivtag in der Hansestadt Rostock, aber beispielsweise auch der Unfall eines Flixbusses auf der A19[49] und die bereits genannte AfD-Demonstration im September auf. In den Schlagworten der einzelnen Wochen spiegeln sich die wesentlichen Ereignisse ebenfalls direkt wider. So findet die Hanse Sail in der KW 32 entsprechende Berücksichtigung, genauso wie die Fußballspiele des FC Hansa Rostock gegen Stuttgart (KW 33) bzw. gegen Nürnberg (KW 37) oder der Anschlag auf aserbaidschanische Studenten in der KW 35.

Neben den Trends sind in Abbildung 7-19 die in Rostock diskutierten Themen zum Überblick als Wordcloud dargestellt. In die Wordcloud flossen alle vorverarbeiteten Tweets des Untersuchungszeitraumes ein. Dies gibt einen schnellen Überblick über die auf Twitter diskutierten Themen und kann gut als Überblick dienen. Auf eine Darstellung je Stadtbereich wurde verzichtet, da hier die Gesamtzahl an verwertbaren Nachrichten deutlich geringer wäre und bestimmte Themen, die nur Rostock als Ganzes betreffen, nicht mehr dargestellt werden würden.

[48] https://www.tag24.de/nachrichten/junge-frau-vermisst-rostock-tiffany-fehlt-jede-spur-zeugen-melden-polizei-oeffentlichkeit-suche-798174
[49] https://www.rbb24.de/panorama/beitrag/2018/08/flixbus-unfall-schweden-berlin-busfahrer.html

Abbildung 7-19: Wordcloud der diskutierten Themen in der Hansestadt Rostock im Untersuchungszeitraum.

Tabelle 7-4: Mittels TF-IDF ermittelte Trends im Untersuchungszeitraum, manuell nach Aussagekraft gefiltert.

Kalenderwoche	Trend	TF-IDF-Score
Gesamt	'domscheit'	7.56
	'aquarelle'	7.35
	'bürgerschaft'	7.19
	archivtag rostock'	6.92
	'treffen rostock'	6.92
	'flixbus'	6.92
	'hambacherforst'	6.92
	afd demo rostock'	6.91
	'demosamstag rostock'	6.91
KW 32	'brandstifter'	6.97
	'regen'	6.97
	'rostock feuerwerk hansesail'	6.87
	'rostock kleinkind sterben'	6.87
	'rostock mädchen sterben'	6.87
	'feine sahne'	6.87
	'hansa liga'	6.87
	sail freitag'	6.87
	'stau rostock'	6.87
	'segelboot'	6.87
KW 33	'hansa rostock stuttgart'	6.74
	'berlin autobahn verunglückt'	6.66
	'drittligist hansa rostock'	6.66
	'nähe rostock reisebus'	6.66
	'rostock richtung berlin'	6.66
	'vfb stuttgart verliert'	6.66
	'graben gefahren'	6.66
	'km stau'	6.66
	'rostock dfbpokal'	6.66
	'rostock ewolves'	6.66
KW 34	rostock runde dfbpokal'	6.78
	'berlin richtung rostock'	6.78
	'fc nürnberg'	6.78
	'afdfch'	6.78
	'aidamar'	6.78
	'beach'	6.78

Kalenderwoche	Trend	TF-IDF-Score
KW 35	'baden'	6.86
	'studenten aserbaidschan'	6.79
	verlässt hansa'	6.79
	'chemnitzer'	6.79
	'erinnern'	6.79
KW 36	'hochzeit'	7.03
	'vfb'	7.03
	'alleoderkeiner allesaufrausch niemalsebbe'	6.91
	'mord'	6.91
	'tribsees'	6.91
	'waldorfschule'	6.80
	'fc hansa rostock'	6.80
KW 37	'exhibitionistische handlungen warnemünde'	6.69
	'noafd wirsindmehr rostock'	6.69
	'rostock widerstände polizeibeamte'	6.69
	'fc nürnberg'	6.69
	'löwen rostock'	6.69
	ostseestadion hansa'	6.69
KW 38	'rostock hinhocken'	6.94
	'fans'	6.80
	'liga'	6.74
	'arppe'	6.69
	'aufzug'	6.69
	'blockadetraining'	6.69
	'seawolves'	6.69
KW 39	'jena'	6.93
	'münster tor dadashov'	6.81
	'afdprotestzug rostock'	6.81
	'demokratie archive'	6.81
	'juden afd'	6.81
	'kirche moschee'	6.81
	'rostock archivtag'	6.81
	'zoo rostock'	6.81

7.3.2 LDA

Da die Trend-Identifikation via TF-IDF für Twitter nur durchschnittliche Ergebnisse liefert, wurde für Tweets des Untersuchungszeitraums ein LDA-Modell erstellt und ausgeführt. Grundsätzlich ist dies mittels der Python-Bibliotheken gensim und Scikit-Learn möglich. Aufgrund des größeren Funktionsumfangs werden die Themen mittels der gensim LDA ermittelt, im Anhang sind dazu allerdings vergleichend auch Ergebnisse der Scikit-LDA dargestellt.

Die optimale Anzahl an Klassen wurde anhand der Kohärenz der Klassen sowie mit Hilfe der Bibliothek pyLDAvis ermittelt (SIVERT & SHIRLEY 2014). In Abbildung 7-20 ist die Kohärenz zwischen den Themen in Abhängigkeit von der Themenanzahl dargestellt. Daraus wird deutlich, dass die höchsten Werte bei vier (0.54) respektive sieben (0.56) Klassen zu finden sind. Dies bedeutet, dass in diesen beiden Fällen die Themen die meisten unterschiedlichen Wörter aufweisen. Warum die Kohärenz allerdings bei fünf respektive sechs Themen so schlecht ausfällt, ist nicht nachzuvollziehen. Zudem wird deutlich, dass zwölf Themen ebenfalls einen guten Wert bei der Kohärenz liefern.

Da die Entscheidung zu einer optimalen Klassenzahl schwer fällt, sind die Daten anhand einer t-Distributed Stochastic Neighbour Embedding Interpolation (t-SNE) respektive einer Singular Value Decomposition (SVD) verarbeitet und dargestellt worden (ROSSANT 2015, LESKOVEC et al. 2014). Beispielhaft sei hier auf die SVD Interpolation der verwendeten LDA auf Basis der Python-Bibliothek gensim mit sieben respektive zwölf Klassen in Abbildung 7-21 verwiesen. Alle weiteren Scatterplots sind für eine genauere Betrachtung im Anhang zu finden.

Im Wesentlichen wird deutlich, dass sich die einzelnen Themen zu großen Teilen überschneiden. Zudem werden erneut die besseren Ergebnisse durch gensim bestätigt. Allerdings liefern die reinen Scatterplots noch keine Aussagen bezüglich der Themen selbst. Aus diesem Grund ist in Abbildung 7-22 jedes Thema und die zehn Schlagwörter mit der höchsten Wichtung dargestellt. Bei der dargestellten Verteilung mit sieben Klassen sind die einzelnen Themen gut trennbar, obwohl es einige Überlappungen gibt. Bei zwölf Klassen ändert sich das sehr deutlich, die Themen sind nicht mehr wirklich auseinander zu halten. Daher empfiehlt sich eine LDA-Klassifikation auf Basis von sieben zu differenzierenden Themen. Die gute Separierbarkeit wird auch an den Gewichten der Wörter der einzelnen Themen ersichtlich.

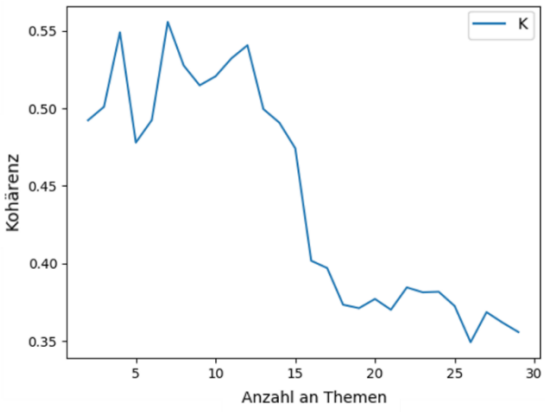


Abbildung 7-20: Anzahl an Themen und deren Kohärenz mittels gensim LDA.

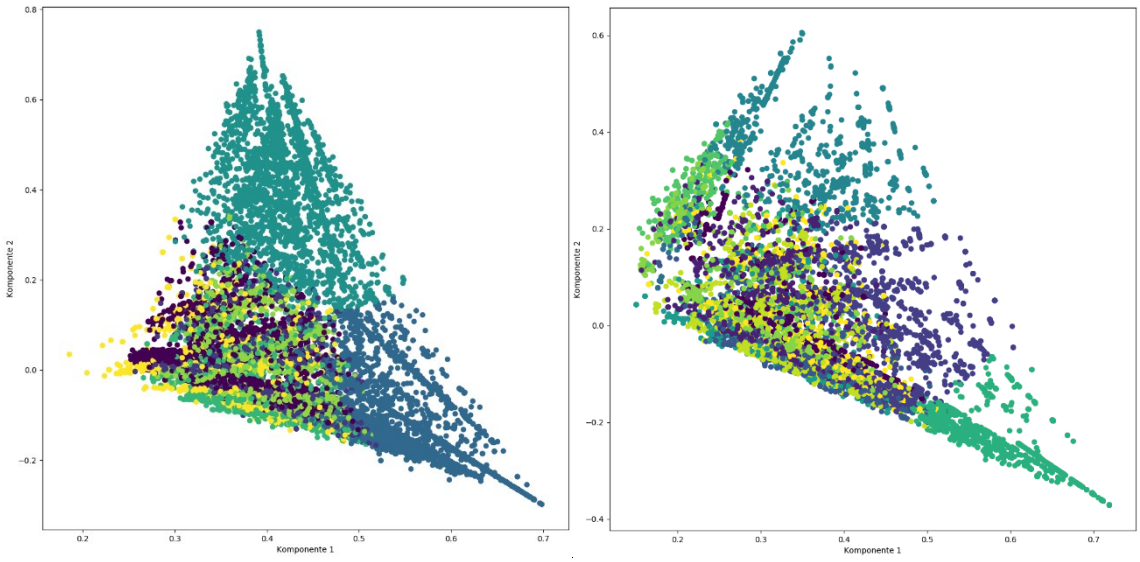


Abbildung 7-21: SVD-Interpolation der LDA mittels gensim auf Basis von sieben Klassen (links) und zwölf Klassen (rechts).

Nachfolgend sind auch die wöchentlichen Themen auf Basis von sieben Klassen berechnet worden. Beispielhaft ist die Kalenderwoche 38 mit der AfD-Demonstration dargestellt (Tabelle 7-5). Eine vollständige Auflistung aller Themen pro Woche befindet sich im Anhang der Arbeit. Aus den Daten wird deutlich, dass die Themen sich extrem überschneiden. Daher rühren auch die niedrigen Gewichte der einzelnen Themen. Andererseits wird ersichtlich, dass das dominante Thema in der KW 38 die AfD-Demonstration in Rostock war.

Tabelle 7-5: Ergebnis der LDA für die KW 38.

Thema	Schlagwörter	Max. Gewicht
1	afd, rostock, sitzblockade, hro, umkehren	0.235
2	höcke, rostock, protesten, antifaschisten, umdrehen	0.145
3	rostock, afddemo, fckafd, illegalen, demonstration	0.143
4	rostock, demonstranten, straßen, zdf, tagesschau	0.109
5	Rostock, ib, ibler, ibmitgliedern, exnpdler	0.068
6	demo, tausenden, gingen, grüße, fahnen	0.068
7	noafd, rostock, fcknzs, platzierung, ipunktes	0.127

Abbildung 7-22: Top-Wörter nach Wichtung und Anzahl der einzelnen Themen auf Basis der gensim LDA für den gesamten Untersuchungszeitraum (06.08.2018 - 30.09.2018).

7.4 Netzwerkdarstellung

Neben der Abbildung der Ergebnisse in der physischen Welt sind auch die Darstellung und Berechnung und deren Beziehungen untereinander im Sozialen Netzwerk von Twitter selbst von Bedeutung. Dafür werden die Knoten (twitternde Accounts) und Kanten (Wer folgt wem?) entsprechend visualisiert.

Dazu wurden hier sowohl die Accounts mit Bezug zur Hansestadt Rostock im gesamten Untersuchungszeitraum (Abbildung 7-23) als auch die Accounts mit Bezug zur AfD-Demonstration am 22.09.2018 (Abbildung 7-24) ausgewählt und entsprechend verarbeitet und dargestellt.

In Abbildung 7-23 sind dabei auf Basis der OpenOrd Berechnung alle Accounts, welche Tweets mit Bezug zu Rostock abgesetzt haben und deren Beziehung untereinander mittels Gephi dargestellt worden (MARTIN et al. 2011). Zur besseren Differenzierung sind die einzelnen Cluster farblich hervorgehoben worden, wobei durch die entsprechenden Verbindungen Mischungen entstehen. Die Größe der Knoten entspricht dabei der Anzahl an den ein- und ausgehenden Kanten und stellt damit letztlich die Bedeutung dar. Allerdings ist anzumerken, dass die Größe keine Verbindung mit der Anzahl an Tweets im Untersuchungszeitraum aufweist.

Im Graph sind deutlich vier Cluster zu identifizieren. Das erste große Cluster umfasst vor allem Accounts, die in Beziehung mit der AfD zu sehen sind (blau). Dazu zählt beispielsweise Björn Höcke. Demgegenüber steht ein Bereich, der vor allem der linken Szene Rostocks zuzuordnen ist. Hierzu zählt beispielsweise Feine Sahne Fischfilet oder auch Rostock Hilft (rot). Im rechten Bereich der Abbildung hingegen sind vor allem Accounts mit Bezug zum Fußball zu finden. Dazu zählt auch eine ganze Reihe an Twitter-Accounts die in Verbindung mit dem VfB Stuttgart stehen. Die dem FC Hansa Rostock nahe stehenden Knoten sind hingegen eher in der Mitte des Graphen zu finden. Die Thematik erstreckt sich dabei sehr weitreichend, d.h., dass diese auch eine entsprechende Bedeutung für eine Vielzahl an Twitter-Nutzern in Rostock zu haben scheint.

Oberhalb der AfD-nahen Accounts findet sich zudem ein Cluster, welches die türkische Gemeinde zu umfassen scheint (grün). Direkt daneben sind zudem lose verbundene Knoten zu finden, die keiner eindeutigen Thematik oder Bevölkerungsgruppe zugeordnet werden können. Im Zentrum des Graphen, jedoch deutlich stärker an die politische Linke angebunden, finden sich zudem die großen Nachrichtenportale. An den Rändern sind durch einzelne, unverbundene Punkte Accounts dargestellt, die keinerlei Verbindung zu anderen Twitter-Nutzern mit Bezug zu Rostock im Untersuchungszeitraum aufwiesen.

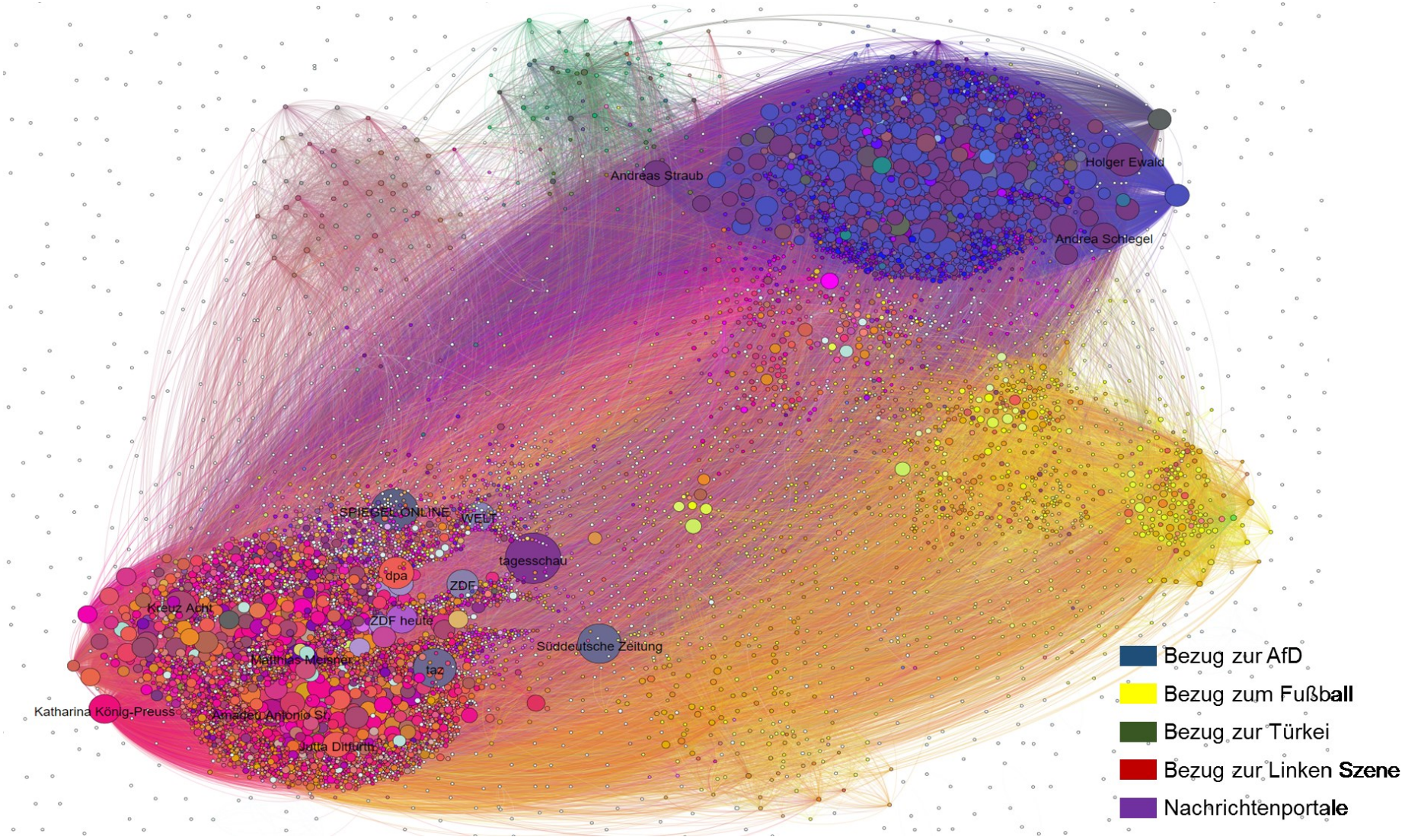


Abbildung 7-23: Twitternetzwerk aller Accounts mit Bezug zur Hansestadt Rostock im Untersuchungszeitraum (06.08.2018 - 30.09.2018).

Aus dem mittels des ForcedAtals 2 berechneten Netzwerk der User mit Bezug zur AfD-Demonstration werden sehr deutlich drei große Fraktionen sichtbar, wobei zwei enger verbunden sind (Abbildung 7-24) (JACOMY et al. 2014). Zum einen ist dies das Netzwerk der AfD-Anhänger (blau). Diese scheinen untereinander gut vernetzt zu sein, die Bindung mit den AfD-Gegnern (rot) findet allerdings nur über die herkömmlichen Medien (grün) statt. Vor allem auf den „#wasdietagesschauverschweigt“-Hashtag sei an dieser Stelle verwiesen. Aber auch dem Spiegel oder dem NDR wird von den Anhängern der AfD gefolgt, obgleich die Bindung dieser Medien an die Gegner der AfD wohl deutlich größer ist. Zwischen den beiden Lagern selbst scheinen hingegen kaum Beziehungen zu existieren.

Das Netzwerk, welches sich gegen die AfD richtet, scheint in sich sehr eng verknüpft. Dies reicht von den Mitorganisatoren der Gegendemonstrationen Rostock Nazifrei bis zur Punkband Feine Sahne Fischfilet. Erstaunlich ist, dass wohl auch enge Beziehungen zum als prinzipiell neutral zu wertenden Twitteraccount der Polizei Rostock bestehen. Zwischen dieser selbst und den Anhängern der AfD hingegen werden kaum Beziehungen sichtbar.

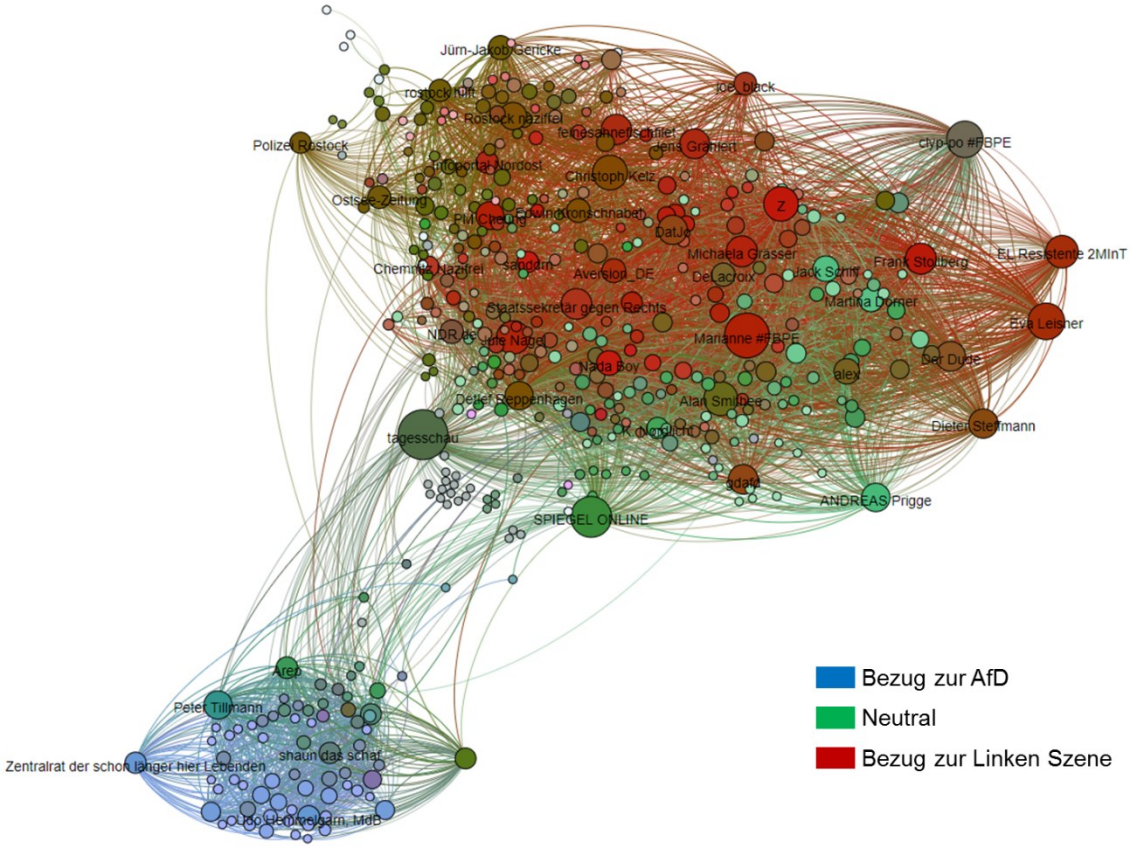


Abbildung 7-24: Netzwerkdiagramm der Accounts mit Bezug zur AfD-Demonstration am 22.09.2018 in Rostock. Die Größe und Farbintensität der Punkte steht für die Anzahl an Followern innerhalb des Netzes.

8 Diskussion

„Nicht Sieg sollte der Zweck der Diskussion sein, sondern Gewinn."

Joseph Joubert (1754 - 1824), französischer Schriftsteller

8.1 Verortung

Die Verortung von Nachrichten aus Sozialen Netzen steht seit einiger Zeit im Fokus der Wissenschaft. Die größte Herausforderung während dieser Arbeit war die Übertragung einer vor allem in der englischsprachigen Welt angesiedelten Wissenschaftsdomäne auf den deutschsprachigen Raum (ZHENG et al. 2018, ATEFEH & KHREICH 2015).

Obwohl verschiedene Arbeiten gezeigt haben, dass hohe Nachrichtenzahlen in eher ländlichen Gebieten eine Herausforderung darstellen (FUCHS et al. 2013), konnte im Rahmen dieser Arbeit dargelegt werden, dass es möglich ist, auf einem kleinen Maßstab eine große Anzahl an Tweets zu verorten, zu kategorisieren und zu analysieren. Bereits VETTERMANN et al. (2018) haben in einem deutlich kürzeren Zeitraum zeigen können, dass sich im Schnitt 348 Nachrichten pro Tag mit Bezug zu Rostock verorten ließen. Den Fokus auf die Ableitung der Lokalisation aus den Metadaten sowie dem Text des Tweets zu stützen, ist auch hier erfolgreich. Vor allem die deutliche Steigerung der Auflösung der Nachrichten im Vergleich zu ähnlichen Arbeiten wie z. B. von SCHULZ et al. (2013) ist hervorzuheben.

Allerdings müssen die Ergebnisse durchaus kritisch betrachtet werden. Im Vergleich zu VETTERMANN et al. (2018) sowie VETTERMANN et al. (2017b) konnte der Anteil der hochgenau lokalisierbaren Nachrichten auf über 17 % gesteigert werden. Dazu kommen 10 % der Nachrichten, die sich auf Straßenebene verorten ließen. Andererseits lassen sich nun auch klare Aussagen zur Genauigkeit treffen, welche bei den Verortungen der Ebene drei mit lediglich 55 % sehr schlecht ausfällt. Die Ursachen sind hier in der Expertenfindung zu suchen. Ein Problem stellen Spam und Werbung dar. Hierbei werden, z. T. in unregelmäßigen Abständen, entsprechende Werbeangebote durch wechselnde Accounts veröffentlicht, in deren Hashtags häufig eine ganze Reihe von Städten zu finden sind. Dadurch werden die Accounts als Einwohner, also als Experten, angenommen. Erst ein fortschrittlicherer Spam-Filter könnte hier Abhilfe schaffen (INUWA-DUTSE et al. 2018). Ein weiteres Problem können Nachrichtenportale darstellen. Auch hier werden häufig Nachrichten mit Bezug zu Rostock via Twitter geteilt, gerade bei besonderen Ereignissen wie z. B. der Hanse Sail. Folglich werden die entsprechenden Accounts ebenfalls als in Rostock verortbar angenommen und Nachrichten in Rostock verortet, die sich eigentlich auf andere Städte beziehen.

Die Gesamtgenauigkeit von 83 % wird vor allem durch die hohe Genauigkeit der nur auf Rostock bezogenen Nachrichten gestützt. Dieser Wert ist jedoch mit Vorsicht zu betrachten, da die Nachrichten, die sich zwar auf Rostock bezogen, aber keine Berücksichtigung fanden, sich nicht in der Genauigkeit widerspiegeln. Ursache ist, dass kein gesamter Twitter-Stream deutscher Tweets einbezogen und kontrolliert werden kann. Andererseits ist ein Vergleich nur mit geokodierten Tweets, die Bezug zu Rostock haben, ebenfalls nicht möglich, da deren Anteil zu gering ist. Daher lassen sich diesbezüglich keine belastbaren Genauigkeiten mit Bezug zum gesamten Twitter-Stream ermitteln, wie dies beispielsweise bei SCHULZ et al. (2013) durchgeführt wird. Nichtsdestotrotz lässt sich die im Vorhinein aufgestellte These 1, dass sich Nachrichten mit ausreichender Dichte generieren lassen,

bejahen, obgleich die Dichte mit der Integration weiterer Sozialer Netze wie z. B. Instagram weiter erhöht werden könnte. Mit 500 Mio. DAUs ist das Potential hier ebenfalls riesig (STATISTA 2018). Im Vergleich mit BUSCHBAUM et al. (2017) wird die Performanz des hier vorgestellten Verfahrens noch deutlicher. So gelang es BUSCHBAUM et al. (2017) über einem Zeitraum von sechs Monaten lediglich knapp 10 000 Nachrichten auf dem Maßstab von Stadtteilen zu verorten.

Gleiches gilt für These 2, dass sich eine Genauigkeit auf der Skale einzelner Straßen erreichen lässt. Hier konnte sogar gezeigt werden, dass ein nicht unerheblicher Anteil an Nachrichten sich sogar auf kleinerer Skale verorten ließ, wobei dies auf Kosten der Genauigkeit geht. ZHENG et al. (2018) zeigen hierbei 18 Arbeiten auf, bei denen auf Basis von Tweets versucht wurde, POIs oder Koordinaten aus selbigen zu bestimmen. Allerdings findet sich in deren Übersicht nur eine Arbeit, bei der das Ground Truthing, wie hier, mittels manueller Zuordnung durchgeführt wurde (HAHMANN et al. 2014). Dies erscheint auch als ein großer Nachteil der gezeigten Methode. Hierfür müssten die Daten über einen längeren Zeitraum gesammelt werden, um einen ausreichenden Anteil an mit GNSS bestimmten Lokationen zu erhalten (ZHENG et al. 2018). Aus diesen ließe sich dann die Abweichung zwischen GNSS-Koordinate und Verortung quantifizieren (SCHULZ et al. 2013).

Der große Unterschied fast aller Arbeiten ist jedoch, dass diese sich auf deutlich größere Maßstäbe beziehen und daher im lokalen Raum keine so genaue Verortung, wie sie in dieser Arbeit gezeigt wird, durchgeführt werden kann. So zeigen HAHMANN et al. (2014) zwar, wie ihr Algorithmus Tweets einzelnen POI-Kategorien, wie z. B. Bahnhöfen oder Supermärkten, zuordnen kann, Probleme treten hier bedingt durch den Algorithmus aber beispielsweise in Gebieten mit einer hohen Dichte an Geschäften oder Supermärkten auf. Eine eindeutige Zuordnung ist hier bei HAHMANN et al. (2014) nicht mehr gegeben. Großer Vorteil des dort vorgestellten Verfahrens ist allerdings die Unabhängigkeit von einem statischen Gazetteer. Demgegenüber versuchen sich FLATOW et al. (2015) an einer hochgenauen Lokalisation, allerdings mit Bezugspunkt New York und einer Datenbasis von 14.5 Mio. Tweets in einem Zeitraum von zwei Jahren. Zudem steht hier ebenfalls die Zuordnung von N-Grammen zu bestimmten Lokationen im Vordergrund, weniger die praktische Lokalisierung. Solch ein Verfahren kann allerdings für die Region Rostock mit ihrer eher geringen Rezeption auf Twitter nicht durchgeführt werden.

Fazit ist, dass das Verfahren des Naive Gazetteer Matchings auf Basis eines auf lokale Gegebenheiten angepassten Gazetteers sehr gute Ergebnisse erzielen kann. Großes Problem bleibt allerdings die Übertragbarkeit, da der Gazetteer hierzu entsprechend auf andere Orte angepasst und die Daten von OSM recht aufwendig nachbearbeitet werden müssten. Hier könnte das von HAHMANN et al. (2014) vorgestellte Verfahren eine gute Ergänzung darstellen. Dennoch lässt sich These 3 bestätigen, da das Ziel, die Zuordnungen mittels eines angepassten Gazetteers zu verbessern, erreicht wurde. Dies haben schon die Arbeiten von VETTERMANN et al. (2017b) sowie VETTERMANN et al. (2018) gezeigt. Allerdings wären noch weitere Verbesserungen möglich, wie beispielsweise die Integration von Abkürzungen (GELERNTER & BALAJI 2013).

8.2 Themenklassifikation

Aus den Daten lässt sich ein gutes Bild bestimmter Themen sowie der Wichtigkeit einzelner Stadtteile ziehen. Allerdings gibt es hier relativ wenige vergleichbare Arbeiten. So liefern LONGLEY & ADNAN (2015) mit ihrer sozio-demografischen Twitteranalyse eine vergleichbare Analyse bezüglich Londons. Vor allem die Identifikation raum-zeitlicher Hotspots mittels des Shannon Diversity Index könnte für diese Arbeit eine weitere Möglichkeit

sein, die Stadt Rostock in ihrer Rezeption in den sozialen Netzen hervorzuheben. Allerdings können auch LONGLEY & ADNAN (2015) auf eine deutlich höhere Datendichte zurückgreifen, bedingt durch die Stadtgröße aber auch der allgemeinen Aktivität auf Twitter (FUCHS et al. 2013). BUSCHBAUM et al. (2017) haben mit Hilfe mehrerer Sozialer Netze (Facebook, Instagram, Twitter) versucht, die Rezeption einzelner Frankfurter Stadtteile darzustellen. Allerdings wurde hierbei nur auf Hashtags geachtet, nicht jedoch auf explizite Themen und das Verfahren zur Datengewinnung unterscheidet sich zusätzlich deutlich.

Die Rezeption der einzelnen Stadtteile der Hansestadt deckt sich im Wesentlichen mit ihrer Bedeutung. Bekannt, vor allem bei Touristen, ist Warnemünde. Daneben spielt sich das soziale Leben v. a. in der KTV sowie im Stadtzentrum ab, das Hansaviertel hingegen ist insbesondere als Austragungsort für die Spiele des FC Hansa Rostock bekannt (HANSESTADT ROSTOCK 2018). Das negativ belastete Bild der Stadt Rostock bedingt durch Lichtenhagen 1992, spiegelt sich ebenfalls auf Twitter wider, nicht zuletzt bedingt durch die in den Untersuchungszeitraum fallenden Krawalle in Chemnitz. Hierbei stellten auch einige Nachrichtenportale eine Verbindung zu den Ereignissen in Rostock her, insbesondere da diese am Jahrestag des Ereignisses von Rostock-Lichtenhagen stattfanden (HECKMANN 2018, KIPPER 2018, KÜHNEL 2018).

Die Zuordnung der Trends zeigt, dass das Verfahren auf Basis von TF-IDF zwar gut auf Basis von Textdokumenten funktioniert, bei kurzen Texten allerdings deutliche Schwächen aufweist (GHOSH & DESARKAR 2018). Aus diesem Grund wird TF-IDF in den meisten Fällen mit überwachten Verfahren wie Naive Bayes (LI et al. 2018), SVM (PARILLA-FERRER et al. 2014) oder NN (KHOSLA et al. 2017) kombiniert. Ursächlich dafür sind mehrere Faktoren: Zum einen wird der Text durch die Aufbereitung (Entfernung von Stoppwörtern, Zahlen, etc.) zwar besser verarbeitbar, der Interpretation der jeweiligen Trends ist dies allerdings nicht zuträglich. Lediglich mit Vorwissen sind die Trends entsprechend interpretierbar. Eine Möglichkeit wäre, hier nur noch auf Hashtags Bezug zu nehmen.

Aufgrund der durchwachsenen Performanz des TF-IDF-Verfahrens wurde im Rahmen dieser Arbeit noch LDA auf Basis der Python Bibliothek gensim verwendet. Die Differenzierung der Themen und Trends ist aber auch hier schwierig. Allerdings wird anhand des Verfahrens auch klar erkennbar, dass die Themen sich deutlich überschneiden. Eine Möglichkeit wäre, die Anzahl an Themen weiter zu reduzieren, wie dies beispielsweise OSTROWSKI (2015) durchgeführt hat. Andererseits würde wohl auch eine Vergrößerung der Datenbasis oder eventuell eine Lemmatisierung Verbesserungspotential bieten. Letztere ist jedoch im Deutschen nur schwer verwendbar, da alle getesteten Verfahren nicht die gewünschten Ergebnisse geliefert haben. Lediglich eine Integration von weiteren textbasierten Analyseverfahren, wie N-Gramme oder TF-IDF könnte hier vielleicht eine Verbesserung bieten (HUANG et al. 2017). Gerade im Bezug zu den wöchentlichen Trends mittels LDA verschärft sich dieses Problem weiter, da die Datenbasis weiter verringert wird. Es wird außerdem deutlich, dass Rostock als Ortsbezeichnung in der Majorität der Tweets vorkommt. Hier könnte eine Vorabfilterung eventuell für eine weitere Verbesserung sorgen. Nichtdestotrotz lassen sich sehr deutlich die jeweils diskutierten Themen identifizieren.

Die Verfahrensweise der LDA ist nicht unbedingt als reine Trend-Identifikation zu bezeichnen, sondern entspricht vielmehr dem Topic Finding (GHOSH & DESARKAR 2018). Letztlich überschneiden sich aber beide Bereiche (OSTROWSKI 2015). Da die Identifikation von Themen in Echtzeit auf die Tweets angewendet wird, kann im Rahmen dieser Arbeit durchaus weiter von Trend-Identifikation gesprochen werden, obgleich eben Verfahren des Topic-Findings genutzt werden (BENHARDUS 2013).

Durch das einfache Matching der Tweets zu einem oder mehreren vordefinierten Themen lässt sich zumindest ein Überblick über dominante Themen als auch in Echtzeit eine direkte thematische Zuordnung durchführen. Das Verfahren ist jedoch deutlich einfacher, als dies beispielsweise in GHOSH & DESARKAR (2018) oder PARILLA-FERRER et al. (2014) dargelegt ist. Allerdings reicht es für die Anforderung, eine grobe Klassifikation im Twittermonitor zu gewährleisten und sich einen Überblick über die bereits vorselektierten Themen und deren Rezeption auf Twitter zu verschaffen. Gerade in Kombination mit der Sentimentanalyse kann hier ein Mehrwert geschaffen werden. Zum einen wird es so möglich, negativ belastete Tweets direkt in Verbindung mit Ereignissen zu erkennen. D. h., dass beispielsweise die Schäden einer Sturmflut sich von einem Urlauber, der baden geht, unterscheiden lassen. Evident wird diese Möglichkeit bereits an den deutlich positiveren Stimmungen mit Bezug zur Thematik Veranstaltungen gegenüber der Thematik Sicherheit. Dennoch sind die Unterschiede relativ gering, was sicherlich u. a. auf das grobe Klassifikationsverfahren der Themen zurück zu führen ist.

Gerade wenn man sich die höher aufgelösten Daten (Ebene 1 und tiefer) hinsichtlich ihrer thematischen Zuordnung betrachtet, werden die Schwachstellen deutlich. So ist die Nachrichtenzahl pro Stadtbereich z. T. zu gering, um Aussagen über dominante Themen oder gar Wünsche der Bevölkerung treffen zu können.

Sinnvoll ergänzt werden kann diese Schwachstelle über die Visualisierung der Beziehungen der Accounts untereinander. Bei der Analyse des Netzwerkes, sowohl beim Einzelereignis der AfD-Demonstration als auch über den gesamten Untersuchungszeitraum wird deutlich, dass sich die wesentlichen Themen in Form der vernetzten Accounts niederschlagen. Ganz besonders deutlich wird dies anhand des DFB-Pokalspiels zwischen dem FC Hansa Rostock und dem VfB Stuttgart. Im Rahmen des Fußballspiels lässt sich ein komplettes Netz identifizieren, was im Bezug zum VfB Stuttgart zu stehen scheint. Gleichzeitig wird aber auch deutlich, dass die AfD-Demonstrationen offenbar den Diskurs im Zeitraum dominierten, da sich eine Vielzahl an links- als auch rechtsgerichteten Nutzern identifizieren ließ, die untereinander sehr gut vernetzt sind. Gerade hinsichtlich der Rechtspopulisten wird dies deutlich, da diese eine in sich sehr eng vernetzte Gruppe, jedoch mit wenig Kontakt zu den anderen Gruppen darzustellen scheinen (FIEDLER et al. 2017).

Nichtdestotrotz lassen sich über die Trend- und Themenfindung Diskussionsschwerpunkte identifizieren, die die Twitternutzer mit Bezug zu Rostock interessieren und diskutieren. These 4 kann somit teilweise bestätigt werden.

Vor einem ähnlichen Problem standen auch BUSCHBAUM et al. (2017). Hier kommt letztlich wieder die Problematik der relativ geringen Nachrichtendichte abseits von Hotspots zum Tragen (FUCHS et al. 2013). So findet man beispielsweise keine Rezeption des in den lokalen Medien recht intensiv diskutierten Projektes Groß-Biestow auf Twitter (SCHWERINER VOLKSZEITUNG 2016). Andererseits liegt in diesem Fall die Diskussion schon etwas zurück. Im Falle von konkreten Aussagen zur Stadtplanung lassen sich aus den Daten im Untersuchungszeitraum dennoch keine Aussagen bestimmen. Folglich muss These 5 (mit Hilfe von Sozialen Netzwerken lassen sich konkrete Handlungsempfehlungen im Bereich der kommunalen Planung aussprechen) in der vorliegenden Arbeit eher abgelehnt werden obgleich sich in einem längerem Untersuchungszeitraum ein anderes Bild zeigen kann.

8.3 Sozio-kulturelle Hotspots

In einer Region, in der insbesondere der Tourismus eine wesentliche wirtschaftliche Komponente darstellt, ist das Image von besonderer Wichtigkeit. Dazu können eine Vielzahl von Maßnahmen durchgeführt werden, die versuchen, das Bild einer Region oder eines Ortes nachhaltig zu prägen, z. B. die Darstellung Thüringens als Grünes Herz Deutschlands (KÖLLNER 2016). Allerdings sagen diese Kampagnen selbst noch nichts über die tatsächliche Wahrnehmung von Orten aus (BUSCHBAUM et al. 2017).

Hier können die Methoden der Sentimentanalyse sowie der Hotspotanalyse Möglichkeiten bieten, eine weitere Datengrundlage neben Umfragen, wie beispielsweise bei MEHL et al. (2017), zu schaffen.

Eng mit den oben angesprochenen Punkten ist der Fokus bezüglich der Identifikation von sozio-kulturellen Hotspots verbunden. Hier liefern die Ergebnisse in Verbindung mit der Sentimentanalyse ein sehr gutes Bild der Hansestadt Rostock. Klar hervor tritt das touristische Zentrum Warnemünde (HANSESTADT ROSTOCK 2018). Unterstützt wird der touristische Aspekt durch die Zuordnung der Themen. So bezogen sich neben der Thematik Sport 13 % aller in Warnemünde verorteten Nachrichten auf das Thema Urlaub. Gleichzeitig erscheint aber die sportliche Bedeutung etwas überraschend. Diese scheint wohl den zahlreichen Segel-, Beachsoccer- oder sonstigen Sportveranstaltungen in Warnemünde geschuldet zu sein. Zudem sind in ganz Rostock über 53 000 Menschen Mitglied in einem Sportverein (Stand 2018) (HANSESTADT ROSTOCK 2018). Folglich ist Sport für die Einwohner Rostocks von großer Bedeutung.

Betrachtet man sich die lokale, räumliche Verteilung der Tweets, wird deutlich, dass neben dem reinen Bezug auf den Stadtbereich (auf der Heatmap in Zentrum Warnemündes) vor allem auf den Alten Strom sowie auf die Bereiche der Strandpromenade vom Teepott bis zum Hotel Neptun häufig Bezug genommen wird. Die touristische Attraktivität schlägt sich aber auch deutlich im physischen Stadtbild an den zahlreichen Besuchern nieder.

Neben Warnemünde ist v. a. die Bedeutung des Hansaviertels herauszuheben. Dies ist hier konkret dem FC Hansa Rostock geschuldet. Mit über 11 000 Mitgliedern ist der Fußballclub zudem der mit Abstand größte Sportverein Rostocks (HANSESTADT ROSTOCK 2018). Allein im Hansaviertel sind knapp 21 % aller Tweets, welche sich mindestens auf Ebene 1 verorten, nur der Thematik Sport und wohl hauptsächlich dem Fußball zuzuschreiben. Die Reichweite ist über den Untersuchungszeitraum mit in Summe knapp 2.5 Mio. Followern enorm. Diese Reichweite wird von keiner anderen Thematik auch nur annähernd erreicht. Zusätzlich ist das Ostseestadion in der höher aufgelösten Verteilung immer ein Hotspot. Dadurch wird die große Bedeutung des Fußballs für die Rezeption der Hansestadt unterstrichen und das Ostseestadion kann als einer der bedeutendsten sozio-kulturellen Hotspots der Hansestadt gewertet werden. Allerdings ist zusätzlich anzumerken, dass durch das Weiterkommen des FC Hansa Rostock im DFB-Pokal gegen den VfB Stuttgart eine zusätzliche, mediale Aufmerksamkeit erzeugt worden ist, die in einem anderen Untersuchungszeitraum so nicht gegeben gewesen wäre. Bestätigt wird dies auch durch die Vernetzung der einzelnen Nutzer.

Zu erwartende sozio-kulturelle Hotspots der Hansestadt sind die KTV sowie die Innenstadt der Hansestadt. Dies schlägt sich auch in der Verteilung der Nachrichten nieder. So bezogen sich im Untersuchungszeitraum über 1 200 respektive 2 700 Tweets auf die beiden Stadtbereiche. Auch die räumliche Verteilung innerhalb der Stadtbereiche zeigt Hotspots. Dazu zählt der Stadthafen, der auch neben der Hanse Sail häufig Erwähnung findet (vgl.

VETTERMANN et al. 2017b). Neben den klassischen POI's Kröpeliner Tor oder Steintor findet im Bereich der Innenstadt insbesondere der Hauptbahnhof eine große Rezeption. Allerdings ist hier anzumerken, dass ein Bot häufig Tweets hinsichtlich Verspätungen oder Aufzügen absetzt. Interessant ist zudem, dass Bereiche um den Doberaner Platz sowie um den Neuen Markt das Gros des Interesses auf sich lenken. Das Interesse an der KTV lässt sich vor allem durch viele Bars und Restaurants sowie Läden und Veranstaltungsorte erklären. Diese werben häufig über Twitter, finden aber auch entsprechend Berücksichtigung durch die Onlinecommunity. Dies wird daran deutlich, dass hier verhältnismäßig viele Tweets mit Bezug zum Thema Arbeit zu finden sind. Erstaunlich ist allerdings, dass die Stadtbereiche durch einen hohen Anteil an Tweets mit Bezug zur Thematik Sicherheit auffallen. 49 % der Tweets in der KTV und immerhin 16 % der Tweets im Stadtbereich Stadtmitte sind dieser Thematik zuzuordnen. Erklärt werden könnte dies mit dem hohen Anteil an den gesamten Straftaten in Rostock. So sind 2017 16 % aller Straftaten in der Stadtmitte und 11 % in der KTV verübt wurden (HANSESTADT ROSTOCK 2018). Über Feuerwehreinsätze etc. gibt es leider keine flächenscharfen Daten. Allerdings lässt sich der Überhang auch durch die AfD-Demonstration in der Rostocker Innenstadt und für die KTV durch die Verortung der Polizeiinspektion Rostock in der KTV erklären. Durch diese werden zahlreiche Kriminalfälle, Fahndungsmeldungen etc. via Twitter geteilt und entsprechend dort verortet.

Der vierte und letzte große Hotspot der in den Ergebnissen deutlich wird, ist Lichtenhagen. Dabei spielt zum einen das Erinnern an die Ereignisse vom August 1992 eine Rolle. So fiel in den Untersuchungszeitraum der 26. Jahrestag der Brandanschläge und Ausschreitungen rund um die ehemalige Aufnahmestelle und ein Wohnheim für Asylbewerber (OSTSEEZEITUNG 2018). Zum anderen wird zu den damaligen Ereignissen eine Verbindung zu den rechtsextremen Ausschreitungen in Chemnitz Ende August 2018 gezogen. Dies spiegelt die Themenzuordnung (43 % Sicherheit) als auch der Coldspot bei der Sentimentanalyse wider. Damit bleibt Rostock in der Wahrnehmung auf Twitter weiterhin sehr eng mit dem damaligen Pogrom verbunden.

Im Rahmen der Netzwerkdarstellung ließ sich allerdings relativ überraschend eine fünfte soziale Gruppe identifizieren, die in der Hansestadt zumindest eine Rolle zu spielen scheint: die Türkische Gemeinde. Allerdings spiegeln sich deren Nachrichten nicht räumlich wieder, dennoch scheint sie einen gewissen sozio-kulturellen Einfluss in der Hansestadt zu besitzen.

Interessant ist die Verschiebung der Hotspots im Tagesgang. Die Attraktivität Warnemündes scheint nachmittags am höchsten und am späten Abend zu verflachen. Dafür verteilen sich die Besucher weiträumiger im Stadtviertel. In der Innenstadt ist ähnliches zu beobachten. Die Verschiebung vom Neuen Markt in Richtung Westen lässt sich vor allem durch die lebhafte Gaststättenszene in der KTV erklären. Die Menschen pendeln also vom wirtschaftlichen ins kulturelle Zentrum der Hansestadt.

Die These 6, dass touristisch oder kulturell-sozial relevante Orte sich via Twitter identifizieren lassen, lässt sich folglich bestätigen, auch wenn es wünschenswert wäre, dies in einer noch besseren Auflösung und in kleineren Zeitfenstern zu ermöglichen. Allerdings ist hier weiterhin das Problem der Datendichte gegeben, dem im konkreten Fall beispielsweise mit der Integration weiterer Sozialer Netze wie Instagram entgegen gewirkt werden könnte.

8.4 Ereignisrezeption und Soziale Netzwerke

Eines der Hauptaugenmerke der Arbeit war darauf gerichtet, zu prüfen, in wie weit sich einzelne Ereignisse in ihrer Rezeption auf Twitter niederschlagen. Ursprünglich lag hierbei der Fokus vor allem auf hydrologischen Extremereignissen. Diese fanden jedoch im konkreten Untersuchungszeitraum nicht statt. Allerdings konnten VETTERMANN et al. (2017a) zeigen, dass diese sich durchaus auf Twitter widerspiegeln. Eine solche Ereignisrezeption macht sich vor allem die Hazardforschung zu Nutze, da so direkt Information zu bestimmten Gefahren und Hotspots gewonnen werden kann (RESCH et al. 2018, COMUNELLO et al. 2016, FUCHS et al. 2013). Das wurde in dieser Arbeit am Beispiel der Tweets zur AfD-Demonstration am 22.09.2018 untersucht. Neben dem raum-zeitlichen Verlauf lag der Fokus auf der im Falle von Einzelereignissen wichtigen Rezeptionsgeschwindigkeit der Ereignisse auf Twitter. Letztere ist v. a. für die Koordination von Einsatzkräften sowie für die Bevölkerungsinformation von großer Bedeutung (TERPSTRA et al. 2012).

Gerade an der Demonstration der AfD schien ein großes mediales Interesse zu bestehen Dies spiegelt sich nicht nur in der Berichterstattung der lokalen Medien, sondern zusätzlich auf der Rezeption auf Twitter wider. Besonders deutlich wird dies an den stark ansteigenden Nachrichtenzahlen, wie sie häufig bei Ereignissen dieser Art auftreten (vgl. TERPSTRA et al. 2012). Gleichzeitig zeigt Twitter, welches Potential sich im Falle von Großereignissen bietet. So lieferten eine Vielzahl von Accounts Informationen über den aktuellen Stand der Demonstration. Des Weiteren fand ein Teil der Demonstrationsorganisation via Twitter statt. So konnten einzelne Gruppen direkt auf bestimmte Ereignisse hingewiesen und in bestimmte Richtungen gelenkt werden. Solche Informationen sind letztlich für die Sicherheitsdienste von entscheidender Bedeutung, da sich daraus Brennpunkte lokalisieren und der Bedarf an Einsatzkräften ableiten lassen (TERPSTRA et al. 2012). Durch die frei verfügbaren Informationen kann so eventuell auf diskutierte Methoden der großflächigen Überwachung jeglicher privater Kommunikation verzichtet, oder sie zumindest eingeschränkt bzw. konkretisiert werden (SPIEGEL ONLINE 2011).

Vor allem die hohe Geschwindigkeit der Rezeption einzelner Ereignisse ist dabei von Bedeutung. So existieren mehrere Arbeiten und Projekte[50], die sich genau dies zu Nutze machen wollen, beispielsweise zur Ereignisdetektion in Katastrophenfällen, oder auch zur Identifikation von Hotspots bei bestimmten Ereignissen (z. B. Erdbeben) in naher Echtzeit (RESCH 2017, COMUNELLO et al. 2016). Dies bedeutet, dass sich bei konkreten Ereignissen in kurzen Zeiträumen entsprechende Handlungsempfehlungen ableiten lassen. Das zeigt sich neben dem 22.09.2018 vor allem daran, dass publikumswirksame Einzelereignisse auch in wöchentlichen Trends deutlich zutage treten. Hier dominieren Themen wie einzelne Fußballspiele, Unfälle oder Konzerte.

Neben der reinen Rezeption der Ereignisse und deren Zuordnung im Raum können auch die Positionen einzelner Nutzer von Bedeutung sein. Gerade in Verbindung mit einer Analyse und Visualisierung der mit den jeweiligen Nutzern verbundenen Sozialen Netze lassen sich tiefgreifende Aussagen über selbige treffen. Großer Vorteil ist hier, dass nicht nur eine Analyse des Sozialen Netzes als solches stattfindet, sondern stets, im Gegensatz zu FIEDLER et al. (2017), ein Bezug zur physischen Umwelt hergestellt werden kann. Denn schließlich kann eine Einordnung bestimmter Ereignisse in einen sozio-kulturellen Rahmen gerade für die Einsatzplanung von Sicherheitskräften von großer Relevanz sein. Interkulturell ist dies von besonderer Wichtigkeit, um beispielsweise eine Eskalation zu vermeiden (JACOBSEN 2011). Damit kann These 7, die Identifikation von einzelnen Events

[50] https://colabis.de/

und die Visualisierung des Interesses der Bevölkerung daran in Sozialen Netzen vollumfänglich bestätigt werden.

9 Zusammenfassung und Ausblick

„Wer enden kann zur rechten Zeit, der hat Vergnügen lange Zeit."

Carl Peter Fröhling (1933), deutscher Germanist und Philosoph

Im Rahmen dieser Arbeit konnte eine Methode entwickelt werden, die in der Lage ist, Tweets in lokalem Maßstab zu verorten, vorgegebenen Themen zuzuweisen und hinsichtlich ihrer Trends und Stimmungen zu analysieren. Außerdem konnte gezeigt werden, dass die sozialen Netzwerke in der realen Welt ihre Spiegelbilder im Netz haben. Durch diese Verknüpfung können wertvolle Informationen aus den Sozialen Netzen generiert und in die analoge Welt übertragen werden. Damit bietet die Arbeit eine vollumfängliche Analyse der Hanse- und Universitätstadt Rostock auf Twitter.

Im verhältnismäßig kurz gefassten Untersuchungszeitraum von sechs Wochen (06.08.2018 - 30.09.2018) konnte der entwickelte Algorithmus 29 771 Nachrichten mit Bezug zur Hansestadt filtern und auf verschiedenen Skalen verorten. Das Verfahren war in der Lage, mit Hilfe des mehrschichtigen Gazetteer-Matchings insgesamt 27 % aller Nachrichten mindestens auf der Skale von Straßen bei einer Gesamtgenauigkeit von 83 % zu verorten. Allerdings geht dieser hohe Anteil auf Kosten der Genauigkeit des Algorithmus bedingt durch die Expertenfindung. Für eine Verbesserung des Matchings müsste also hier angesetzt werden.

Aus der hochgenauen Verortung ergeben sich vielfältige Möglichkeiten. Diese umfassen die Identifikation besonders beliebter und attraktiver Orte sowie von eher unattraktiven Orten. Daraus lassen sich potentiell konkrete Aussagen und Empfehlung hinsichtlich des Stadtmarketings ableiten, um die Attraktivität für den Tourismus, die Bevölkerung sowie die Wirtschaft gleichermaßen steigern zu können. Allerdings hat die Arbeit auch gezeigt, dass im konkreten Anwendungsfall bei der Sentimentklassifikation klassische Methoden des maschinellen Lernens, konkret die SVM, praktikabler sind als Neuronale Netze.

Daneben wird durch die Identifikation von Trends respektive Themen via TF-IDF und LDA die Möglichkeit gegeben, die Nachrichten zu kategorisieren und bestimmte Ereignisse zu identifizieren. Allerdings lassen sich nicht, wie erhofft, konkrete Handlungsempfehlungen hinsichtlich der Stadtplanung aus den Tweets ableiten.

Des Weiteren ist es möglich, auf Basis von Tweets in Echtzeit entscheidungsunterstützend zu wirken. Dies kann im Rahmen von Katastrophen und Extremereignissen geschehen. Diverse Arbeiten haben bereits gezeigt, dass Twitter mit seinem hohen Datenaufkommen in der Lage ist, schnell Informationen für Rettungsdienste oder den Katastrophenschutz bereitzustellen. Auch hier konnte die schnelle Reaktionszeit verifiziert werden, wobei dies anhand eines gesellschaftlichen Ereignisses, der Demonstration der AfD in Rostock am 22.09.2018, abgebildet worden ist. Folglich lassen sich aus der spatio-temporalen Kombination der Daten in Echtzeit Handlungsempfehlungen für Sicherheitskräfte ableiten.

Alles in allem konnten die Ziele der Arbeit erreicht werden, obgleich weiteres Verbesserungspotential besteht, was jedoch nicht mehr Eingang in diese Arbeit gefunden hat. Dazu zählt die Integration weiterer Sozialer Netze, um die Datenbasis zu vergrößern. Dies wird bei vielen kommerziellen Produkten im Rahmen der Analyse von Sozialen Medien, z. B. um Reichweiten von Werbekampagnen zu ermitteln, bereits so gehandhabt. Der bestehende Algorithmus könnte hierfür einfach um einen Harvester beispielsweise für Instagram oder Flickr erweitert werden.

Weiter oben wurde zudem die Problematik der Genauigkeit auf der kleinsten Skale bedingt durch das Verfahren zur Expertenfindung angesprochen. Hier besteht deutliches Verbesserungspotential. So werden viele Nachrichtenportale etc. fälschlicherweise als Einwohner der Hansestadt kategorisiert. Eine Einbindung eines weiteren Blacklistings könnte hierbei helfen, dies zu verhindern.

Ein großer Schwachpunkt des Verfahrens ist der Gazetteer selbst und die mit ihm verbundene Übertragbarkeit auf andere Städte oder Regionen. Andere Arbeiten in dieser Richtung haben versucht, die Verortung mit Hilfe von Webservices durchzuführen. Leider ist dies im Rahmen der gewünschten Auflösung nicht möglich, da bestimmte Ortsbezeichnungen nicht in OSM o. ä. zu finden sind. Allerdings wäre eine Kombination möglich, bei der die Webservices um eine lokale Komponente ergänzt werden.

Der letzte Punkt, welcher einer zukünftigen Überarbeitung bedarf, ist die Themenzuordnung. Hier würde sich ein Verfahren des maschinellen Lernens empfehlen, um die Zuordnung zu verbessern. Die genannten Punkte könnten also zukünftig das Verfahren weiter verbessern und erweitern und u. U. Teil weiterer Arbeiten und Veröffentlichungen sein.

10 Literatur

ABADI, M., AGARWAL, A., BARHAM, P., BREVDO, E., CHEN, Z., CITRO, C., CORRADO, G.S., DAVIS, A., DEAN, J., DEVIN, M., GHEMAWAT, S., GOODFELLOW, I., HARP, A., IRVING, G., ISARD, M., JIA, Y., JOZEFOWICZ, R., KAISER, L., KUDLUR, M., LEVENBERG, J., MANÉ, D., MONGA, R., MOORE, S., MURRAY, D., OLAH, C., SCHUSTER, M., SHLENS, J., STEINER, B., SUTSKEVER, I., TALWAR, K., TUCKER, P., VANHOUCKE, V., VASUDEVAN, V., VIÉGAS, F., VINYALS, O., WARDEN, P., WATTENBERG, M., WICKE, M., YU, Y. & ZHENG, X. (2015): TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems.

AGARWAL, A., XIE, B., VOVSHA, I., RAMBOW, O. & PASSONNEAU, R. (2011): Sentiment Analysis of Twitter Data. In: NAGARAJAN, M. & GAMON, M. (Hrsg.). Proceedings of the Workshop on Languages in Social Media. Stroudsburg: Association for Computational Linguistics, 30–38.

AGGARWAL, C.C. (Hrsg.) (2011): Social Network Data Analytics. New York, Dordrecht, Heidelberg, London: Springer.

AGGARWAL, C.C. & WANG, H. (2011): Text Mining in Social Networks. In: AGGARWAL, C.C. (Hrsg.). Social Network Data Analytics. New York, Dordrecht, Heidelberg, London: Springer, 353–378.

ALFONSECA, E. & MANANDHAR, S. (2002): Improving an Ontology Refinement Method with Hyponymy Patterns. In: EUROPEAN LANGUAGE RESOURCES ASSOCIATION (Hrsg.). Third International Conference on Language Resources and Evaluation, 235–239.

AL-RFOU, R. (2015): Python Bindings for Compact Language Detector 2, https://github.com/aboSamoor/pycld2/blob/master/README.rst (Stand: 2015-03-03) (Zugriff: 2016-08-26).

ARBEITSKREIS ARCHITEKTUR (2013): Architektur der Geodateninfrastruktur Deutschland: Ziele und Grundlagen.

ARNOLD, M., SCHWARZWÄLDER, B., BEER-TÓTH, K., ZBINDEN, M. & BAUMGART, K. (2009): Mehrwert naturnaher Wasserläufe: Untersuchung zur Zahlungsbereitschaft mit besonderer Berücksichtigung der Erschließung für den Langsamverkehr. Umwelt-Wissen 09/12. Bern: Bundesamt für Umwelt.

ASLAM, S. (2018): Instagram by the Numbers: Stats, Demographics & Fun Facts, https://www.omnicoreagency.com/instagram-statistics/ (Stand: 2018-09-17) (Zugriff: 2019-03-22).

ATEFEH, F. & KHREICH, W. (2015): A Survey of Techniques for Event Detection in Twitter. – Computational Intelligence 31, 1, 132–164.

BARBIER, G. & LIU, H. (2011): Data Mining in Social Media. In: AGGARWAL, C.C. (Hrsg.). Social Network Data Analytics. New York, Dordrecht, Heidelberg, London: Springer, 327–352.

BASTIAN, M., HEYMANN, S. & JACOMY, M. (2009): Gephi: An Open Source Software for Exploring and Manipulating Networks. In: ADAR, E., HURST, M., FININ, T., GLANCE, N., NICOLOV, N. & TSENG, B. (Hrsg.). Proceedings of the Third International ICWSM Conference. Menlo Park: AAAI Press, 361–362.

BENHARDUS, J. (2013): Streaming Trend Detection in Twitter. – International Journal of Web Based Communities 9, 1, 122–139.

BERGMANN, N. (2009): Volkszählung und Datenschutz: Proteste zur Volkszählung 1983 und 1987 in der Bundesrepublik Deutschland. Hamburg: Diplomica Verlag.

BERNET, M. (2010[1]): Social Media in der Medienarbeit: Online-PR im Zeitalter von Google, Facebook und Co. Wiesbaden: Springer Fachmedien.

BESBINAR, B., SARIGIANNIS, D. & SMEROS, P. (o. J.): Tweet Sentiment Classification. Lausanne.

BIKEL, D.M., MILLER, S., SCHWARTZ, R. & WEISCHEDEL, R. (1997): Nymble: a High-Performance Learning Name-finder. In: GRISHMAN, R. (Hrsg.). Proc. Conference on Applied Natural Language Processing. San Francisco: ACL, 194–201.

BILL, R. (2016): Grundlagen der Geo-Informationssysteme. Berlin, Offenbach: Herbert Wichmann Verlag.

BILL, R., LORENZEN-ZABEL, A. & HINZ, M. (2018): Offene Daten für Lehre und Forschung in raumbezogenen Studiengängen: OpenGeoEdu. – GisScience, 1, 32–44.

BIRD, S., KLEIN, E. & LOPER, E. (2009): Natural Language Processing with Python. Sebastopol: O'Reilly.

BIRZER, M. (2015): So geht Bürgerbeteiligung: Eine Handreichung für die kommunale Praxis. Texte der KommunalAkademie 7. Bonn: Friedrich-Ebert-Stiftung, KommunalAkademie.

BKG (2016): Geoportal DE, http://www.geoportal.de/DE/GDI-DE/gdi-de.html?lang=de (Stand: 2016) (Zugriff: 2017-03-01).

BLEI, D.M., NG, A.Y. & JORDAN, M.I. (2003): Laten Dirichlet Allocation. – Journal of Machine Learning Research, 3, 993–1022.

BLOCK BY BLOCK (2018): Block by Block, https://blockbyblock.org/ (Stand: 2018) (Zugriff: 2018-01-25).

BMUB & BFN (2016): Naturbewusstsein 2015: Bevölkerungsumfrage zu Natur und biologischer Vielfalt.

BODENDORF, F. (2003): Daten- und Wissensmanagement. Springer-Lehrbuch. Berlin, Heidelberg: Springer.

BOND, R., FARISS, C.J., JONES, J.J., KRAMER, D.I., MARLOW, C., SETTLE, J.E. & FOWLER, J.H. (2012): A 61-Million-Person Experiment in Social Influence and Political Mobilization. – Nature 489, 7415, 295–298.

BOOTSTRAP (2018): Bootstrap, https://getbootstrap.com/ (Stand: 2018) (Zugriff: 2018-09-11).

BOYD, d.m. & ELLISON, N.B. (2007): Social Network Sites: Definition, History, and Scholarship. – Journal of Computer-Mediated Communication 13, 1, 210–230.

BRABHAM, D.C. (2012): The effectiveness of crowdsourcing public participation in a planning context. – First Monday 17, 12.

BRABHAM, D.C. (2013): Crowdsourcing. The MIT Press essential knowledge series. Cambridge, London: The MIT Press.

BROWNING, L. (2015): We sent men to the moon in 1969 on a tiny fraction of the data that's in the average laptop, http://www.businessinsider.de/mind-blowing-growth-and-power-of-big-data-2015-6?r=US&IR=T (Stand: 2015-06-09) (Zugriff: 2017-03-06).

BUSCHBAUM, K., BLITZ, A., REITHMEIER, C. & KANWISCHER, D. (2017): Hashtags und Raumkonstruktionen: Eine explorative Studie zum Potential von digitalen Methoden zur Analyse raum-zeitlicher Daten in sozialen Medien. – GisScience, 4, 115–125.

CESIUM CONSORTIUM (2017): Cesium, https://cesiumjs.org/ (Stand: 2017) (Zugriff: 2017-06-20).

CHARALABIDIS, Y., N. LOUKIS, E., ANDROUTSOPOULOU, A., KARKALETSIS, V. & TRIANTAFILLOU, A. (2014): Passive crowdsourcing in government using social media. – Transforming Government: People, Process and Policy 8, 2, 283–308.

CHOLLET, F. (2015): Keras, https://keras.io (Stand: 2019-03-17) (Zugriff: 2019-03-23).

CIELIEBAK, M., DERIU, J.M., EGGER, D. & UZDILLI, F. (2017): A Twitter Corpus and Benchmark Resources for German Sentiment Analysis. In: LUN, W.K. & CHENG, T.L. (Hrsg.). Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. Valencia, 45–51.

CIVITAS (2015): The use of social media to involve citizens in urban mobility projects and city planning.

CLIFF, A.D. & ORD, J.K. (1969): The Problem of Spatial Autocorrelation. In: SCOTT, A. (Hrsg.). Studies in Regional Science. London: Pion, 25–55.

CLOUD SECURE ALLIANCE (2013): Expanded Top Ten Big Data Security and Privacy Challenges, https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Expanded_Top_Ten_Big_Data_Security_and_Privacy_Challenges.pdf (Stand: April 2012) (Zugriff: 2017-03-06).

COHN, J.P. (2008): Citizen Science: Can Volunteers Do Real Research? – BioScience 58, 3, 192–197.

COLYER, A. (2016): The amazing power of word vectors, https://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/ (Stand: 2016-04-21) (Zugriff: 2018-02-22).

COMUNELLO, F., PARISI, L., LAUCIANI, V., MAGNONI, F. & CASAROTTI, E. (2016): Tweeting after an earthquake: User localization and communication patterns during the 2012 Emilia seismic sequence. – Annals of Geophysics 59, 5.

CRAMPTON, J.W., GRAHAM, M., POORTHUIS, A., SHELTON, T., STEPHENS, M., WILSON, M.W. & ZOOK, M. (2013): Beyond the geotag: Situating 'big data' and leveraging the potential of the geoweb. – Cartography and Geographic Information Science 40, 2, 130–139.

CREIGHTON, J.L. (2005): The Public Participation Handbook: Making better decisions through citizen involvement. San Francisco: Jossey-Bass.

CRESCI, S., CIMINO, A., ORLETTA, F.D. & TESCONI, M. (2015): Crisis Mapping During Natural Disasters via Text Analysis of Social Media Messages. In: WANG, J., CELLARY, W., WANG, D., WANG, H., CHEN, S.-C., LI, T. & ZHANG, Y. (Hrsg.). Web Information Systems Engineering - WISE 2015: 16th International Conference, Miami, FL, USA, November 1-3, 2015, Proceedings, Part II. LNCS sublibrary. SL 3, Information systems and applications, incl. Internet/Web, and HCI 9419. Cham: Springer, 250–258.

CYBENKO, G. (1989): Approximation by superpositions of a sigmoidal function. – Mathematics of Control Signals and Systems 2, 2, 303–314.

DELANEY, K.J. (2005): Yahoo Acquires Flickr Creator, http://www.wsj.com/articles/SB111136815551984786 (Stand: 2005-03-20) (Zugriff: 2016-04-29).

DERIU, J., LUCCHI, A., LUCA, V. de, SEVERYN, A., MÜLLER, S., CIELIEBAK, M., HOFMANN, T. & JAGGI, M. (2017): Leveraging Large Amounts of Weakly Supervised Data for Multi-Language Sentiment Classification. In: BARRETT, R., CUMMINGS, R., AGICHTEIN, E. & GABRILOVICH, E. (Hrsg.). International World Wide Web Conference Committee. Perth, 1045–1052.

DESHWAL, A. & SHARMA, S.K. (2016): Twitter Sentiment Analysis using Various Classification Algorithms. In: SHUKLA, B., KHATRI, S.K. & KAPUR, P.K. (Hrsg.). 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions): September 7-9, 2016 venue: Amity University Uttar Pradesh, Noida, India. Piscataway, NJ: IEEE, 251–257.

DIE GRÜNEN (2016): Leitfaden zur Bürgerbeteiligung in Rostock entwickeln, https://gruene-fraktion-rostock.de/news/news-detail/article/leitfaden_zur_buergerbeteiligung_in_rostock_entwickeln/ (Stand: 2016-02-02) (Zugriff: 2017-02-24).

DITTMANN, A. (2016): Soziale Netzwerke: Liste von A-Z für Deutschland, https://www.axeldittmann.de/blog/soziale-netzwerke-liste-von-a-z-11711 (Stand: 2016-03-11) (Zugriff: 2018-01-26).

DOAN, A., RAMAKRISHNAN, R. & HALEVY, A.Y. (2011): Crowdsourcing systems on the World-Wide Web. – Communications of the ACM 54, 4, 86–96.

DOBUSCH, L. (2014): Digitale Zivilgesellschaft in Deutschland: Stand und Perspektiven 2014.

EIMEREN, B. & RIDDER, C.-M. (2001): Trends in der Nutzung und Bewertung der Medien 1970 bis 2000. – Media Perspektiven 2001, 11, 538–553.

EKBAL, A. & BANDYOPADHYAY, S. (2008): Bengali Named Entity Recognition using Support Vector Machine. In: SANGAL, R., SHARMA, M.D. & SINGH, A.K. (Hrsg.). Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages. Hyderabad: Asian Federation of Natural Language Processing, 51–58.

ELLISON, N.B., STEINFIELD, C. & LAMPE, C. (2007): The Benefits of Facebook "Friends: Social Capital and College Students' Use of Online Social Network Sites. – Journal of Computer-Mediated Communication 12, 4, 1143–1168.

ESRI (o.J.): Einführung in ArcGIS, http://resources.arcgis.com/de/help/getting-started/articles/026n00000014000000.htm (Stand: o.J.) (Zugriff: 2018-01-30).

EUROPEAN COMISSION (2017): INSPIRE: Infrastructure for spatial information in Europe, http://inspire.ec.europa.eu/ (Stand: 2017-03-01) (Zugriff: 2017-03-01).

FACEBOOK INC. (2016): Facebook Places, https://www.facebook.com/places/ (Stand: 2016) (Zugriff: 2016-04-11).

FACEBOOK INC. (2018): Annual Report 2018.

FACEBOOK INC. (2019): Facebook Q4 2018 Results.

FAKE, C. (2008): Cofounder Flickr. In: LIVINGSTON, J. (Hrsg.). Founders at Work, 257–264.

FAZ (2014): Google verabschiedet sich von Google+, http://www.faz.net/aktuell/wirtschaft/netzwirtschaft/google/online-netzwerk-google-verabschiedet-sich-von-google-12909732.html (Stand: 2014-04-25) (Zugriff: 2016-04-29).

FELDMAN, R. & DAGAN, I. (1995): Knowledge Discovery in Textual Databases. In: FAYYAD, U. & UTHURUSAMY, R. (Hrsg.). KDD'95 Proceedings of the First International Conference on Knowledge Discovery and Data Mining. Montreal, Canada: AAAI, 112–117.

FIEDLER, M., KREIL, M., LEHMANN, H., REUTER, M. & ROST, L.C. (2017): So twittert die AfD, https://digitalpresent.tagesspiegel.de/afd (Stand: 2017-04-01) (Zugriff: 2019-03-19).

FLATOW, D., NAAMAN, M., XIE, K.E., VOLKOVICH, Y. & KANZA, Y. (2015): On the Accuracy of Hyper-local Geotagging of Social Media Content. In: CHENG, X., LI, H., GABRILOVICH, E. & TANG, J. (Hrsg.). WSDM 2015 Eighth ACM International Conference on Web Search and Data Mining: Shanghai, China. New York: ACM, 127–136.

FLICKR (2016): Über Flickr, https://www.flickr.com/about (Zugriff: 2016-04-29).

FREES, b. & KOCH, W. (2018): ARD/ZDF-Onlinestudie 2018: Zuwachs bei medialer Internetnutzung und Kommunikation. – Media Perspektiven 2018, 9, 398 -413.

FUCHS, G., ANDRIENKO, N., ANDRIENKO, G., BOTHE, S. & STANGE, H. (2013): Tracing the German centennial flood in the stream of tweets. In: PFOSER, D. & VOISARD, A. (Hrsg.). GEOGROWD 2013: Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information. New York: ACM, 31–38.

GANDOMI, A. & HAIDER, M. (2015): Beyond the hype: Big data concepts, methods, and analytics. – International Journal of Information Management 35, 2, 137–144.

GATTRINGER, K. & TURUCEK, I. (2018): ma 2018 Audio – Konvergenzwährung für Radio und Online-Audio. – Media Perspektiven 2018, 9, 438–450.

GDI-DE (2015[3]): Geodatendienste im Internet: Ein Leitfaden.

GELERNTER, J. & BALAJI, S. (2013): An algorithm for local geoparsing of microtext. – GeoInformatica 17, 4, 635–667.

GELERNTER, J. & MUSHEGIAN, N. (2011): Geo-parsing Messages from Microtext. – Transactions in GIS 15, 6, 753–773.

GEONETWORK OPENSOURCE (2014): GeoNetwork User Guide Release 2.10.4-0.

GEOSERVER (2014): GeoServer User Manual: Release 2.5.x.

GHAHRAMANI, Z. (2001): An Introduction to Hidden markov Models and Bayesian Networks. – Internation Journal of Pattern Recognition and Artificial Intelligence 15, 1, 9–42.

GHOSH, S. & DESARKAR, M.S. (2018): Class Specific TF-IDF Boosting for Short-text Classification: Application to Short-texts Generated During Disasters. In: WWW '18 Companion Proceedings of the The Web Conference 2018 // The Web Conference 2018: Companion of the World Wide Web Conference WWW2018 April 23-27, 2018, Lyon, France. [Geneva, Switzerland]: International World Wide Web Conferences Steering Committee, 1629–1637.

GIERSBERG, F. & LEIBIGER, j. (2019): Mediennutzung in Deutschland 2018: VAUNET-Mediennutzungsanalyse.

GO, A., BHAYANI, R. & HUANG, L. (2009): Twitter Sentiment Classification using Distant Supervision Processing. Stanford.

GOLDHAHN, D., ECKART, T. & QUASTHOFF, U. (2012): Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In: CALZOLARI, N., CHOUKRI, K., DECLERCK, T., DOGAN, M.U., MAEGAARD, B., MARIANI, J., MORENO, A., ODIJK, J. & PIPERIDIS, S. (Hrsg.). Proceedings of the 8th International Language Ressources and Evaluation. Istanbul, 759–765.

GOMAA, W.H. & FAHMY, A.A. (2013): A Survey of Text Similarity Approaches. – International Journal of Computer Applications 68, 13, 13–18.

GONTRUM, J. & SCHEFFLER, T. (2015): Text-based Geolocation of German Tweets. In: BEIßWENGER, M. & ZESCH, T. (Hrsg.). Proceedings of the NLP4CMC 2015 Workshop at GSCL. Duisburg: In Selbstverlegung, 28–32.

GOODCHILD, M.F. (2007): Citizens as sensors: The world of volunteered geography. – GeoJournal 69, 4, 211–221.

GOODFELLOW, I., BENGIO, Y. & COURVILLE, A. (2016): Deep Learning. Cambridge, Massachusetts, London, England: MIT Press.

GOUWS, S., METZLER, D., CAI, C. & HOVY, E. (2011): Contextual Bearing on Linguistic Variation in Social Media. In: NAGARAJAN, M. & GAMON, M. (Hrsg.). Proceedings of the Workshop on Languages in Social Media 11. Stroudsburg: Association for Computational Linguistics, 20–29.

GRAHAM, M., HALE, S.A. & GAFFNEY, D. (2014): Where in the World Are You?: Geolocation and Language Identification in Twitter. – Professional Geographer 66, 4, 568–578.

GRATIER, T., SPENCER, P. & HAZZARD, E. (2015): OpenLayers 3: Beginner's Guide. Birmingham, Mumbai: Packt Publishing.

GREGORIO, F.D. & VARRAZZO, D. (2018): psycopg2, http://initd.org/psycopg/docs/ (Stand: 2018-11-09) (Zugriff: 2019-03-20).

HAHMANN, S., PURVES, R. & BURGHARDT, D. (2014): Twitter location (sometimes) matters: Exploring the relationship between georeferenced tweet content and nearby feature classes. – Journal of Spatial Information Science, 9, 1–36.

HAKLAY, M. (2013): Citizen Science and Volunteered Geographic Information: Overview and Typology of Participation. In: SUI, D., ELWOOD, S. & GOODCHILD, M. (Hrsg.). Crowdsourcing Geographic Knowledge. Dordrecht: Springer Netherlands, 105–122.

HAMMON, L. & HIPPNER, H. (2012): Crowdsourcing. – Wirtschaftsinformatik 54, 3, 165–168.

HAN, B., COOK, P. & BALDWIN, T. (2014): Text-Based Twitter User Geolocation Prediction. – Journal of Artificial Intelligence Research 49, 451–500.

HANSESTADT ROSTOCK (2013): Rostock 2025: Leitlinien zur Stadtentwicklung.

HANSESTADT ROSTOCK (2015a): OpenData.HRO: Portal für offene Daten der Hansestadt Rostock, http://www.opendata-hro.de/.

HANSESTADT ROSTOCK (2015b): Statistisches Jahrbuch Hansestadt Rostock 2015: Selbstverlegung.

HANSESTADT ROSTOCK (2016): Statistische Nachrichten: Bevölkerungsprognose bis 2035: In Selbstverlegung.

HANSESTADT ROSTOCK (2017): Arbeit am Leitfaden zur Bürgerbeteiligung beginnt: Bürgerforum am 6. März im Rathaus, http://rathaus.rostock.de/sixcms/detail.php?id=54378 (Stand: 2017-02-17) (Zugriff: 2017-02-27).

HANSESTADT ROSTOCK (2018): Statistisches Jahrbuch 2018: In Selbstverlegung.

HANSESTADT-ROSTOCK (2012): Die Hansestadt Rostock mit dem Seebad Warnemünde: Tourismuskonzeption 2022.

HECKMANN, D.-O. (2018): Da kann keiner mehr von besorgten Bürgern sprechen: Rechte Ausschreitungen in Chemnitz, https://www.deutschlandfunk.de/rechte-ausschreitungen-in-chemnitz-da-kann-keiner-mehr-von.694.de.html?dram:article_id=426596 (Stand: 2018-08-27) (Zugriff: 2019-01-24).

HEIDEMANN, J. (2010): Online Social Networks – Ein sozialer und technischer Überblick. – Informatik-Spektrum 33, 3, 262–271.

HELLER, C. (2011): Post-privacy: Prima leben ohne Privatsphäre. Beck'sche Reihe 6000. München: Beck.

HERRING, S.C., SCHEIDT, A., KOUPER, I. & WRIGHT, E. (2007): A Longitudinal Content Analysis of Weblogs: 2003 - 2004. In: TREMAYNE, M. (Hrsg.). Blogging, citizenship, and the future of media. London: Routledge, 3–22.

HILL, K. (2012): How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did, http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#5a9fdcd34c62 (Stand: 2012-02-16) (Zugriff: 2016-04-26).

HINZ, M. & BILL, R. (2018): Ein zentraler Einstiegspunkt für die Suche nach offenen Geodaten im deutschsprachigen Raum. – AGIT – Journal für Angewandte Geoinformatik, 4, 298–307.

HORBAT, A., MEYERHOFF, J., DEHNHARDT, A. & HARTJE, V. (2016): Wertschätzung für naturnahe Flusslandschaften an der Deutschen Mittelelbe. In: MAYER, M. & JOB, H. (Hrsg.). Naturtourismus - Chancen und Herausforderungen: Studien zur Freizeit- und Tourismusforschung (SFT) 12, 221–232.

HORNIK, K. (1991): Approximation Capabilities of Multilayer Feedforward Network. – Neural networks, 4, 251–257.

HOTHO, A., NÜRNBERGER, A. & PAAß, G. (2005): A brief survey of text mining. – LDV Forum 20, 1, 19–62.

HOWE, J. (2006): The Rise of Crowdsourcing, https://www.wired.com/2006/06/crowds/ (Stand: 2006-01-06) (Zugriff: 2019-03-15).

HOWE, J. (2008): Crowdsourcing: How the power of the crowd is driving the future of business. New York: Crown Business.

HUANG, L., MA, J. & CHEN, C. (2017): Topic Detection from Microblogs Using T-LDA and Perplexity. In: LV, J., ZHANG, J.H., HINCHEY, M. & LIU, X. (Hrsg.). 24th Asia-Pacific Software Engineering Conference workshops, APSECW 2017: Nanjing, China, 4-8 December 2017 Proceedings. Piscataway, NJ: IEEE, 71–77.

HÜBNER, S. (2016): Web-Service-Orchestrierung beim Aufbau einer Geodateninfrastruktur zur Integration, Prozessierung und Dissemination verschiedenster Daten. Rostock.

HÜBNER, S. & VETTERMANN, F. (2016): Erstellung eines Geodatenportals zu den Klein- und Kleinstgewässern in Rostock. In: BILL, R., ZEHNER, M., GOLNIK, A., LERCHE, T., SCHRÖDER, J. & SEIP, S. (Hrsg.). Geoinformation im Alltag: Nutzen und neue Anforderungen. Berlin: GITO-Verlag, 77–83.

HÜBNER, S., VETTERMANN, F., SEIP, C. & BILL, R. (2016): Creating a Data Portal for Small Rivers in Rostock. In: WOHLGEMUTH, V., FUCHS-KITTOWSKI, F. & WITTMANN, J. (Hrsg.). Advances and New Trends in Environmental Informatics: Stability, Continuity, Innovation. Cham: Springer International Publishing, 301–310.

HUFFAKER, D.A. & CALVER, S.L. (2005): Gender, Identity, and Language Use in Teenage Blogs. – Journal of Computer-Mediated Communication 10, 2, 1–23.

INSTAGRAM (2016): About, https://www.instagram.com/about/ (Stand: 2016) (Zugriff: 2019-03-20).

INTERNET LIVE STATS (2019): Twitter Usage Statistics, http://www.internetlivestats.com/twitter-statistics/ (Stand: 2019-02-24) (Zugriff: 2019-02-24).

INUWA-DUTSE, I., LIPTROTT, M. & KORKONTZELOS, I. (2018): Detection of spam-posting accounts on Twitter. – Neurocomputing 315, 496–511.

JACOBSEN, A. (2011): Interkulturelle Kompetenz als Methode: Der Situative Ansatz. – Soziale Probleme 23, 2, 154–173.

JACOMY, M., VENTURINI, T., HEYMANN, S. & BASTIAN, M. (2014): ForceAtlas2: A continuous graph layout algorithm for handy network visualization designed for the Gephi software. – PloS one 9, 6.

JAUERNIG, H. (2017): Ungezügelte Kurznachrichten: Was Donald Trump für Twitter wert ist, http://www.spiegel.de/wirtschaft/unternehmen/donald-trump-was-der-us-praesident-fuer-twitter-wert-ist-a-1165284.html (Stand: 2017-08-30) (Zugriff: 2019-03-20).

JIFFREY, K.G. (2006): Open Data: the time has come, https://blogs.ch.cam.ac.uk/pmr/2006/09/12/open-data-the-time-has-come/#comment-2236 (Stand: 2006-09-12) (Zugriff: 2019-02-06).

JOACHIMS, T. (1998): Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: NÉDELLEC, C. & ROUVEIROL, C. (Hrsg.). Machine LearningECML-98: ECML-98. Lecture Notes in Computer Science, Lecture Notes in Artificial Intelligence 1398. Berlin, Heidelberg: Springer, 137–142.

KELLY, R. (2014): PyEnchant, http://pythonhosted.org/pyenchant/ (Stand: 2014) (Zugriff: 2016-08-24).

KHARDE, V.A. & SONAWANE, S.S. (2016): Sentiment Analysis of Twitter Data: A Survey of Techniq. – International Journal of Computer Applications 139, 11, 5–15.

KHOSLA, P., BASU, M., GHOSH, K. & GHOSH, S. (2017): Microblog Retrieval for Post-Disaster Relief: Applying and Comparing Neural IR Models. In: KANDO, N., SAKAI, T., JOHO, H., CRASWELL, N., CROFT, W.B., RIJKE, M. de, GUO, J. & MITRA, B. (Hrsg.). SIGIR 2017 Workshop on Neural Information Retrieval (Neu-IR'17): Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval: August 7-11, 2017, Shinjuku, Tokyo, Japan. New York, USA: ACM Association for Computing Machinery.

KIM, J., VASARDANI, M. & WINTER, S. (2016): Similarity matching for integrating spatial information extracted from place descriptions. – International Journal of Geographical Information Science 31, 1, 56–80.

KIPPER, J. (2018): Hier wurde Pogromstimmung verbreitet: Gewalt in Chemnitz (Stand: 2018-08-27) (Zugriff: 2019-01-24).

KLEEMANN, F., VOß, G.G. & RIEDER, K. (2008): Un(der)paid Innovators. – Science, Technology & Innovation Studies 4, 1.

KLINGLER, W. & TURUCEK, I. (2016): Medienzeitbudgets und Tagesablaufverhalten. – Media Perspektiven 2016, 2, 98–107.

KNAUß, F. (2014): Die eingerollte Generation: Smartphone statt Freundschaft, https://www.wiwo.de/erfolg/trends/smartphone-statt-freundschaft-die-eingerollte-generation/9880500.html (Stand: 2014-05-13) (Zugriff: 2019-02-05).

KOGGE (2015): Kogge-Website: Kommunale Gewässer gemeinschaftlich entwickeln, http://kogge.auf.uni-rostock.de/ (Stand: 2017) (Zugriff: 2017-02-28).

KOLB, M. (2016): Wie Twitter-Präsident Trump Themen setzt, http://www.sueddeutsche.de/politik/neuer-us-praesident-wie-twitter-praesident-trump-themen-setzt-1.3281817 (Stand: 2016-12-07) (Zugriff: 2019-03-19).

KOLDRACK, N., VETTERMANN, F. & BILL, R. (2017): Modernes Geodatenmanagement in der Forschung. In: BILL, R., ZEHNER, M., GOLNIK, A., LERCHE, T., SCHRÖDER, J. & SEIP, S. (Hrsg.). Mit Geoinformation Planen. Berlin: GITO-Verlag, 103–110.

KÖLLNER, D. (2016): Tourismuskonzeption Thüringer Wald 2025: Markenstrategie und -architektur.

KOMMUNE21 (2014): Klarschiff.HRO rege genutzt, http://www.kommune21.de/meldung_19830_on.html (Stand: 2014-09-24) (Zugriff: 2017-02-27).

KÖRNERF, A. (2016): Rostock sucht Land für 25000 Bürger, http://www.ostsee-zeitung.de/Mehr/Meinung/Lesermeinung/Rostock-sucht-Land-fuer-25000-Buerger (Stand: 2016-09-12) (Zugriff: 2019-03-20).

KOTSIANTIS, S.B. (2007): Supervised Machine Learning: A Review of Classification Techniques. In: MAGLOGIANNIS, i., KARPOUZIS, K., WALLACE, M. & SOLDATOS, J. (Hrsg.). Frontiers in Artificial Intelligence and Applications: Real word AI systems with applications in eHealth, HCI, information retrieval and pervasive technologies. Frontiers in artificial intelligence and applications 160. Amsterdam, Washington, DC: IOS Press, 3–24.

KREIL, M. (2017): Social Bots, Fake News und Filterblasen: Therapiestunde mit einem Datenjournalisten und vielen bunten Visualisierungen.

KRESSE, W. & FADAIE, K. (2004): ISO Standards for Geographic Information. Berlin, Heidelberg: Springer.

KÜHNEL, J. (2018): Von Lichtenhagen bis Chemnitz: Hier tobte der rechte Mob, http://www.ostsee-zeitung.de/Nachrichten/Politik/Von-Lichtenhagen-bis-Chemnitz-Hier-tobte-der-rechte-Mob (Stand: 2018-08-27) (Zugriff: 2019-01-24).

KUNIAVSKY, M. & CREECH, A. (2009): Information Shadows: How Ubiquitous Computing Serializes Everyday Things. – The Serials Librarian 56, 1-4, 65–78.

KWAK, H., LEE, C., PARK, H. & MOON, S. (2010): What is Twitter, a Social Network or a News Media? In: RAPPA, M. (Hrsg.). Proceedings of the 19th International Conference on World Wide Web: WWW'10 ; Raleigh, NC, USA, April 26 - 30, 2010. New York: ACM, 591–600.

LECUN, Y., BENGIO, Y. & HINTON, G. (2015): Deep learning. – Nature 521, 7553, 436–444.

LECUN, Y., BOTTOU, L., BENGIO, Y. & HAFFNER, P. (1998): Gradient Based Learning Applied to Document Recognition. – Proceedings of the IEEE 86, 11, 2278–2324.

LEE, K., PALSETIA, D., NARAYANAN, R., PATWARY, M.M.A., AGRAWAL, A. & CHOUDHARY, A. (2011): Twitter Trending Topic Classification. In: SPILIOPOULOU, M., WANG, H., COOK, D., PEI, J., WANG, W., ZAIANE, O. & WU, X. (Hrsg.). IEEE 11th International Conference on Data Mining workshops (ICDMW), 2011: 11 Dec. 2011, Vancouver, Canada;. Piscataway, NJ: IEEE, 251–258.

LESKOVEC, J., RAJARAMAN, A. & ULLMAN, J.D. (2014[4]): Mining of massive datasets. Cambridge: Cambridge University Press.

LI, C. (2016): Social Media: An Ideal Tool for Public Participation to Promote Deliberative Democracy: The Case of Public Participation in Refugee Crisis. – International Journal of Journalism and Communication 1, 2, 36–41.

LI, H., CARAGEA, D., CARAGEA, C. & HERNDON, N. (2018): Disaster response aided by tweet classification with a domain adaptation approach. – Journal of Contingencies and Crisis Management 26, 1, 16–27.

LICHTENBERG, A. (2016): Totgesagte leben länger - Die SMS im WhatsApp-Zeitalter, http://www.dw.com/de/totgesagte-leben-l%C3%A4nger-die-sms-im-whatsapp-zeitalter/a-17451917 (Stand: 2014-02-23) (Zugriff: 2016-04-28).

LICKLIDER, J.C.R. & TAYLOR, R.W. (1968): The Computer as a communication device. – Science and Technology: For the Technical Man in Management, 21–31.

LONGLEY, P.A. & ADNAN, M. (2015): Geo-temporal Twitter demographics. – International Journal of Geographical Information Science 30, 2, 369–389.

LUI, M. & BALDWIN, T. (2012): langid.py: An Off-the-shelf Language Identification Tool. In: ZHANG, M. (Hrsg.). Proceedings of the ACL 2012 System Demonstrations. Jeju Island: Association for Computational Linguistics, 25–30.

MANNING, C.D., RAGHAVAN, P. & SCHÜTZE, H. (2008): Introduction to Information Retrieval. New York: Cambridge University Press.

MARR, B. (2014): Big Data: The 5 Vs Everyone Must Know, https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know (Stand: 2014-03-06) (Zugriff: 2017-03-06).

MARTIN, S., BROWN, W.M., KLAVANS, R. & BOYACK, K.W. (2011): OpenOrd: An open-source toolbox for large graph layout. In: WONG, P.C., PARK, J., HAO, M.C., CHEN, C., BÖRNER, K., KAO, D.L. & ROBERTS, J.C. (Hrsg.). Proceedings of SPIE: Visualization and Data Analysis 2011 7868, 786806.

MCCALLUM, A. & LI, W. (2003): Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In: DAELEMANS, W. & OSBORNE, M. (Hrsg.). Proceedings Conference on Computational Natural Language Learning. Edmonton: ACL, 188–191.

MCDONALD, D. & KELLY, U. (2012): The Value and Benefits of Text Mining: Digital Infrastructure.

MCGRATH, R. (2018): twython, https://twython.readthedocs.io/en/latest/ (Stand: 2018-12-02) (Zugriff: 2019-03-19).

MEHL, D., HOFFMANN, T.G., SCHNEIDER, M., LANGE, A., NEUPERT, BADROW, U. & WENSKE T. A. (2015): Gemeinschaftliches Handeln im kommunalen Hochwassermanagement: Das "Integrierte Entwässerungskonzept" (INTEK) der Hansestadt Rostock. – Korrespondenz Wasserwirtschaft 8, 11, 700–709.

MEHL, D., VETTERMANN, F., HOFFMANN, T.G. & BILL, R. (2017): Präferenzen für die Entwicklung kleiner urbaner Gewässer und Feuchtgebiete: Ergebnisse einer Online-Befragung. – Korrespondenz Wasserwirtschaft, 6, 340–346.

MEYERHOFF, J. (2002): Der Nutzen aus einem verbesserten Schutz biologischer Vielfalt in den Elbeauen: Ergebnisse einer Kontingenten Bewertung. In: DEHNHARDT, A. & MEYERHOFF, J. (Hrsg.). Nachhaltige Entwicklung der Stromlandschaft Elbe: Nutzen und Kosten der Wiedergewinnung und Renaturierung von Überschwemmungs-auen. Kiel: Vauk, 155–184.

MEYERHOFF, J., ANGELI, D. & HARTJE, V. (2012): Valuing the benefits of implementing a national strategy on biological diversity: The case of Germany. – Environmental Science and Policy, 23, 109–119.

MEYERHOFF, J., BOERI, M. & HARTJE, V. (2014): The value of water quality improvements in the region Berlin-Brandenburg as a function of distance and state residency. – Water Resources and Economics, 5, 49–66.

MICHEL, F. (2018): How many public photos are uploaded to Flickr every day, month, year?, https://www.flickr.com/photos/franckmichel/6855169886 (Stand: 2018-09-30) (Zugriff: 2019-03-22).

MIKOLOV, T., CHEN, K., CORRADO, G.S. & DEAN, J. (2013a): Efficient Estimation of Word Representations in Vector Space. – The Computing Research Repository 2013, 1301.3781.

MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G.S. & DEAN, J. (2013b): Distributed Representations of Words and Phrases and their Compositionality. In: BURGES, C.J.C., BOTTOU, L., WELLING, M., GHAHRAMANI, Z. & WEINBERGER, K.Q. (Hrsg.). Advances in neural information processing systems 25: 26th Annual Conference on Neural Information Processing Systems 2012 ; December 3 - 6, 2012, Lake Tahoe, Nevada, USA 2. Red Hook, NY: Curran, 3111–3119.

MILLER, C.C. (2011): Another Try by Google to Take On Facebook, http://www.nytimes.com/2011/06/29/technology/29google.html?_r=0 (Stand: 2011-06-28) (Zugriff: 2016-04-29).

MILLER, H.J. & GOODCHILD, M.F. (2015): Data-driven geography. – GeoJournal 80, 4, 449–461.

MITCHELL, T. (2008): Web-Mapping mit Open Source-GIS-Tools. Beijing: O'Reilly.

MORENO-SANCHEZ, R. (2009): The Geospatial Semantic Web: What are its Implications for Geospatial Information Users. In: CRUZ-CUNHA, M.M., OLIVIERA, E.F., TAVARES, A.J.V. & FERREIRA, L.G. (Hrsg.). Handbook of Research on Social Dimensions of Smeantic Technologies and Web Services. Hershey, New York: Information Science Reference, 588–609.

MORSTATTER, F., GAO, H. & LIU, H. (2015): Discovering Location Information in Social Media. In: LOMET, D. & ZHOU, X. (Hrsg.). Bulletin of the IEEE Computer Society Technical Committee on Data Engineering: Location-based Social Media Analysis 38: IEEE Computer Society, 4–13.

MOURA, J. & SERRÃO, C. (2015): Security and Privacy Issues of Big Data. In: ZAMAN, N., SELIAMAN, M.E., HASSAN, F.M., MARQUEZ, P.G., HASSAN, M.F. & MARQUEZ, F.P.G. (Hrsg.). Handbook of Research on Trends and Future Directions in Big Data and Web Intelligence: IGI Global, 20–52.

MOZETIČ, I., GRČAR, M. & SMAILOVIĆ, J. (2016): Multilingual Twitter Sentiment Classification: The Role of Human Annotators. – PloS one 11, 5, 1–26.

MURTHY, S.K. (1998): Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. – Data Mining and Knowledge Discovery 1998, 2, 345–389.

NANZ, P. & FRITSCHE, M. (2012): Handbuch Bürgerbeteiligung: Verfahren und Akteure, Chancen und Grenzen. Bonn: Bundeszentrale für politische Bildung.

NARR, S., HÜLFENHAUS, M. & ALBAYRAK, S. (2012): Language-Independent Twitter Sentiment Analysis. In: LWA 2012 (Hrsg.). Workshop on Knowledge Discovery, Data Mining and Machine Learning. Dortmund, 12–14.

NDR (2016): Hanse Sail: Positive Bilanz trotz Regenwetters, http://www.ndr.de/unterhaltung/events/Hanse-Sail-Positive-Bilanz-trotz-Regenwetters,hansesail896.html (Stand: 2016-08-15) (Zugriff: 2017-02-07).

NDR (2018): AfD-Demo in Rostock: 700 dafür - 4.000 dagegen, https://www.ndr.de/nachrichten/mecklenburg-vorpommern/AfD-Demo-in-Rostock-700-dafuer-4000-dagegen,rostock1100.html (Stand: 2018-09-23) (Zugriff: 2019-01-10).

NEITZ, D., KÖNIG, p. & FLACH, G. (2010): Klarschiff.HRO: Internet-Plattform zur aktiven Bürgerbeteiligung. In: FLACH, G. & SCHULTZ, J. (Hrsg.). 5. Rostocker eGovernment-Forum 2010: Wissensbasiertes eGovernment: Erschließung und Nutzung von Verhaltensweisen. Berlin: GITO-Verlag, 9–14.

NEITZ, D., SCHWARZ, S. & SCHRÖDER, M. (2013): Offene Daten der Verwaltung im Online-Portal OpenData.HRO. In: BILL, R., FLACH, G., KORDUAN, P., ZEHNER, M. & SEIP, S. (Hrsg.). GeoForum MV 2013: Neue Horizonte für Geodateninfrastrukturen - Open GeoData, Mobility, 3d-Stadt. Berlin: GITO-Verlag, 7–12.

NORVIG, P. (2016): How to Write a Spelling Corrector, https://norvig.com/spell-correct.html (Stand: 08/2016) (Zugriff: 2018-08-31).

OECD (2007): Participative web and user-created content: Web 2.0, wikis, and social networking. Paris: Selbstverlegung.

OLIVIERA, E.T. (2012): pgsimilarity, https://github.com/eulerto/pg_similarity (Stand: 2012-01-16) (Zugriff: 2019-03-19).

OPENLAYERS 3 (2017): API Docs, http://openlayers.org/en/latest/apidoc/ (Stand: 2017) (Zugriff: 2017-10-16).

O'REILLY, T. (2005): What Is Web 2.0, http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html?page=1 (Stand: 2005-09-30) (Zugriff: 2017-06-21).

O'REILLY, T. & BATTELLE, J. (2009): Web Squared: Web 2.0 Five Years On.

ORESKOVIC, A. (2015): Here's another area where Twitter appears to have stalled: tweets per day, https://www.businessinsider.com/twitter-tweets-per-day-appears-to-have-stalled-2015-6?IR=T (Stand: 2015-06-15) (Zugriff: 2019-03-22).

OSTROWSKI, D.A. (2015): Using Latent Dirichlet Allocation for Topic Modelling in Twitter. In: KANKANHALLI, M.S., LI, T. & WANG, W. (Hrsg.). 2015 IEEE International Conference on Semantic Computing (ICSC). Piscataway, NJ: IEEE, 493–497.

OSTSEEZEITUNG (2017): Tief „Axel“: Überschwemmungen in Rostock: In der Altstadt, dem Petriviertel und am Stadthafen ist die Warnow über die Ufer getreten. Häuser sind vom Hochwasser bedroht, http://www.ostsee-zeitung.de/Mehr/Bilder/Bildergalerien/Tief-Axel-Ueberschwemmungen-in-Rostock (Stand: 2017-01-06) (Zugriff: 2019-02-06).

OSTSEEZEITUNG (2018): Lichtenhagen: Neue Stele erinnert an Ausschreitungen, http://www.ostsee-zeitung.de/Nachrichten/MV-aktuell/Lichtenhagen-Neue-Stele-erinnert-an-Ausschreitungen (Stand: 2018-08-25) (Zugriff: 2019-02-24).

PARILLA-FERRER, B.E., FERNANDEZ, P.L. & BALLENA, J.T. (2014): Automatic Classification of Disaster-Related Tweets. In: International conference on Innovative Engineering Technologies (ICIET 2014). Bangkok, 62–69.

PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A. & MICHEL, V. (2011): scikit-learn: Machine Learning in Python. – Journal of Machine Learning Research, 12, 2825–2830, http://scikit-learn.org/stable/index.html.

PFAFFENBERGER, F. (2016): Twitter als Basis wissenschaftlicher Studien: Eine Bewertung gängiger Erhebungs- und Analysemethoden der Twitter-Forschung. Wiesbaden: Springer Fachmedien Wiesbaden; Imprint; Springer VS.

POSTGIS PROJECT STEERING COMMITTEE (2015): PostGIS, http://www.postgis.net/ (Stand: 2018-11-22) (Zugriff: 2019-03-19).

POSTGRESQL GLOBAL DEVELOPMENT GROUP (2017): PostgreSQL 9.6.2 Documentation.

PRABHAKARAN, S. (2018): Topic Modeling with Gensim (Python), https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/ (Stand: 2018) (Zugriff: 2019-02-13).

PYTHON SOFTWARE FOUNDATION (2017): Python, https://www.python.org/about/ (Stand: 2017) (Zugriff: 2017-03-27).

QGIS DEVELOPMENT TEAM (2019): QGIS Geographic Information System: Open Source Geospatial Foundation Project, https://qgis.org/en/site/ (Stand: 2019-03-19) (Zugriff: 2019-03-19).

QUINLAN, J.R. (1993): C4.5: Programs for Machine Learning 16. San Mateo: Morgan Kaufmann Publishers.

RECCHIA, G. & LOUWERSE, M. (2013): A Comparison of String Similarity Measures for Toponym Matching. In: SCHEIDER, S., ADAMS, B., JANOWICZ, K., VASARDANI, M. & WINTER, S. (Hrsg.). ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems: ACM SIGSPATIAL International Workshop on Computational Models of Place. Orlando: Association for Computing Machinery, Inc, 54–61.

REHUREK, R. & SOJKA, P. (2010): Software Framework for Topic Modelling with Large Corpora. In: PAUW, G. (Hrsg.). Proceedings of the Second Workshop on African Language Technology. Valletta, Malta, 45–50.

RESCH, B. (2017): Nutzergenerierte Daten für Entscheidungsunterstützung in naher Echtzeit. München.

RESCH, B., USLÄNDER, F. & HAVAS, C. (2018): Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment. – Cartography and Geographic Information Science 45, 4, 362–376.

RICHARD, M.D. & LIPPMANN, R. (1991): Neural Network Classifiers Estimate Bayesian a Posteriori Probabilities. – Neural Computation 3, 4, 461–483.

RITTER, A. (2012): Extracting Knowledge from Twitter and The Web. Washington.

RITTER, A., CLARK, S., MAUSAM & ETZIONI, O. (2011): Named Entity Recognition in Tweets: An Experimental Study. In: BARZILAY, R. & JOHNSON, M. (Hrsg.). Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh: Association for Computational Linguistics, 1524–1534.

ROSSANT, C. (2015): An illustrated introduction to the t-SNE algorithm, https://www.oreilly.com/learning/an-illustrated-introduction-to-the-t-sne-algorithm (Stand: 2015-03-03) (Zugriff: 2019-03-23).

SANTIAGO, A. (2015): The Book of OpenLayers 3: Theory & Practice: Leanpub.

SCELLATO, S., NOULAS, A., LAMBIOTTE, R. & MASCOLO, C. (2011): Socio-spatial Properties of Online Location-based Social Networks. In: ADAMIC, L.A., BAEZA-YATES, R.A. & COUNTS, S. (Hrsg.). Proceedings of the 5th International AAAI Conference on Weblogs and Social Media: ICWSM. Barcelona: AAAI Press, 329–336.

SCHEFFLER, T. (2014): A German Twitter Snapshot. In: CALZOLARI, N., CHOUKRI, K., DECLERCK, T., LOFTSSON, H., MAEGAARD, B., MARIANI, J., MORENO, A., ODIJK, J. & PIPERIDIS, S. (Hrsg.). LREC 2014, Ninth International Conference on Language Resources and Evaluation: May 26-31, 2014, Reykjavik, Iceland; proceedings. Reykjavik: European Language Resources Association (ELRA), 2284–2289.

SCHMIDT, J.-H. (2012): (Micro)blogs: Practices of Privacy Management. In: TREPTE, S. & REINECKE, L. (Hrsg.). Privacy Online: Perspectives on Privacy and Self-Disclosure in the Social Web. Berlin, Heidelberg: Springer, 159–174.

SCHMITZ, C. (2015): LimeSurvey, http://www.limesurvey.org (Stand: 2015) (Zugriff: 2016-07-25).

SCHOLZ, C. (2014): Generation Z: Wie sie tickt, was sie verändert und warum sie uns alle ansteckt. Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA.

SCHULZ, A., HADJAKOS, A., PAULHEIM, H., NACHTWEY, J. & MUHLHAUSER, M. (2013): A Multi-Indicator Approach for Geolocalization of Tweets. In: CHIARANDINI, L., GRABOWICZ, P.A., TREVISIOLM MICHELE & JAIMES, A. (Hrsg.). Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, 8 - 11 July 2013, Cambridge, Massachusetts, USA. Palo Alto, Calif.: AAAI Press, 573–582.

SCHWERINER VOLKSZEITUNG (2016): Rostock plant Wohngebiet bei Biestow, http://www.svz.de/lokales/rostock/rostock-plant-wohngebiet-bei-biestow-id15262146.html (Stand: 2016-11-05) (Zugriff: 2017-03-03).

SCIKIT-LEARN (2017): Support Vector Machines, http://scikit-learn.org/stable/modules/svm.html.

SEIP, C., KORDUAN, P. & ZEHNER, M.L. (2017): Web-GIS: Grundlagen, Anwendungen und Implementierungsbeispiele. Berlin, Offenbach: Wichmann.

SEKINE, S., GRISHMAN, R. & SHINNOU, H. (1998): A Decision Tree Method for Finding and Classifying Names in Japanese Texts. In: CHARNIAK, E. (Hrsg.). Proceedings of the Sixth Workshop on Very Large Corpora. Brunswick: ACL, 171–178.

SEVENONE MEDIA GMBH (2018): Media Activity Guide 2018: Trends in der Mediennutzung.

SHELTON, T., POORTHUIS, A. & ZOOK, M. (2015): Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. – Landscape and Urban Planning 142, 198–211.

SHIN, D. (2016): Urban Sensing by Crowdsourcing: Analysing Urban Trip behaviour in Zurich. – International Journal of Urban and Regional Research 40, 5, 1044–1060.

SHINYAMA, Y. & SEKINE, S. (2004): Named Entity Discovery Using Comparable News Articles. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (Hrsg.). Proceedings of the 20th international conference on Computational Linguistics. Genf: ACL, 848–853.

SIDARENKA, U. (2016): PotTS: The Potsdam Twitter Sentiment Corpus. In: CALZOLARI, N., CHOUKRI, K., DECLERCK, T., GOGGI, S., GROBELNIK, M., MACGAARD, B., MARIANI, J., MAZO, H., MORENO, A., ODIJK, J. & PIPERIDIS, S. (Hrsg.). Proceedings of the Tenth International Conference on Language Resources and Evaluation, 1133–1141.

SIDARENKA, U., SCHEFFLER, T. & STEDE, M. (2013): Rule-Based Normalization of German Twitter Messages. In: BEIßWENGER, M., LÜDELING, A. & STORRER, A. (Hrsg.). Proceedings of the GSCL Workshop Verarbeitung und Annotation von Sprachdaten aus Genres internetbasierter Kommunikation. Darmstadt: GCSL.

SIDARENKA, U. & STEDE, M. (2016): Generating Sentiment Lexicons for German Twitter. In: NISSIM, M., PATTI, V. & PLANK, B. (Hrsg.). Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media. Osaka, 80–90.

SILVERMAN, B.W. (1986[1]): Density Estimation for Statistics and Data Analysis. Chapman & Hall/CRC Monographs on Statistics and Applied Probability. Boca Raton, FL: Chapman and Hall/CRC.

SILVERTOWN, J. (2009): A new dawn for citizen science. – Trends in ecology & evolution 24, 9, 467–471.

SITES, D. (2014): Compact Language Detector 2, https://github.com/CLD2Owners/cld2 (Stand: 2014-02-05) (Zugriff: 2016-08-26).

SIVERT, C. & SHIRLEY, K.E. (2014): LDAvis: A method for visualizing and interpreting topics. In: CHUANG, J., GREEN, S., HEARST, M., HEER, J. & KOEHN, P. (Hrsg.). Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces. Baltimore, Maryland: Association for Computational Linguistics, 63–70.

SPIEGEL ONLINE (2011): Polizei wertete Tausende Handy-Daten aus: Demo in Dresden, http://www.spiegel.de/netzwelt/web/demo-in-dresden-polizei-wertete-tausende-handy-daten-aus-a-769275.html (Stand: 2011-06-19) (Zugriff: 2019-03-10).

SPLITTERBERGER, A. (2014): Praxishandbuch Rechtsfragen Social Media. Berlin, Boston: Walter de Gruyter.

STATISTA (2016): Nutzung von Videoportalen im Internet: Marktanteil von Video-Sharing-Plattformen in Deutschland im 1. Halbjahr 2016. Hamburg.

STATISTA (2018): Instagram: Statistics & Facts, https://www.statista.com/topics/1882/instagram/ (Stand: 2018) (Zugriff: 2019-02-08).

STEIGER, E., RESCH, B. & ZIPF, A. (2015): Exploration of spatiotemporal and semantic clusters of Twitter data using unsupervised neural networks. – International Journal of Geographical Information Science, 1–23.

STEINBRENNER, T. (2017): Die Geschichte der Social Media im Überblick, https://www.haufe.de/marketing-vertrieb/online-marketing/die-social-media-geschichte-im-ueberblick_132_298002.html (Stand: 2017-02-24) (Zugriff: 2017-03-03).

STOJANOVSKI, D., STREZOSKI, G., MADJAROV, G. & DIMITROVSKI, I. (2015): Twitter Sentiment Analysis Using Deep Convolutional Neural Network. In: ONIEVA, E., SANTOS, I., OSABA, E., QUINTIÁN, H. & CORCHADO, E. (Hrsg.). Hybrid artificial intelligent systems:

10th international conference, HAIS 2015, Bilbao, Spain, June 22-24, 2015 9121. Lecture notes in computer science Lecture notes in artificial intelligence 9121. Cham: Springer, 726–737.

STREICH, B. (2014): Subversive Stadtplanung. Wiesbaden: Springer Fachmedien Wiesbaden.

SUTTON, C. & MCCALLUM, A. (2012): An Introduction to Conditional Random Fields. – Foundations and Trends® in Machine Learning 4, 4, 267–373.

TERPSTRA, T., VRIES, A., STRONKMAN, R. & PRADIES, G.L. (2012): Towards a realtime Twitter analysis during crises for operational crisis management. In: ROTHKRANTZ, L., RISTVEJ, J. & FRANCO, Z. (Hrsg.). ISCRAM 2012 Conference Proceedings: 9th International Conference on Information Systems for Crisis Response and Management. Vancouver: Simon Fraser University.

THIEL, G. (2015): Geoinformationen als Impulsgeber für die digitale Informationsgesellschaft. In: KOLBE, T.H., BILL, R. & DONAUBAUER, A. (Hrsg.). Geoinformationssysteme 2015: Beiträge zur 2. Münchner GI-Runde. Berlin, Offenbach: Wichmann, 142–146.

THIELSCH, M.T. & WELTZIN, S. (2012): Online-Umfragen und Online-Mitarbeiterbefragungen. In: THIELSCH, M.T. & BRANDENBURG, T. (Hrsg.). Praxis der Wirtschaftspsychologie II: Themen und Fallbeispiele für Studium und Praxis. Münster: MV Wissenschaft, 109–127.

TRAYNOR, D. & CURRAN, K. (2013): Location-Based Social Networks. In: LEE, I. (Hrsg.). Mobile Services Industries, Technologies, and Applications in the Global Economy. Hershey: Information Science Reference, 243–253.

TSOTSIS, A. & PANZARINO, M. (2014): Google+ Is Walking Dead, http://techcrunch.com/2014/04/24/google-is-walking-dead/ (Stand: 2014-04-24) (Zugriff: 2016-04-29).

TWITTER INC. (2015): Annual Report 2015.

TWITTER INC. (2016a): About Us, https://about.twitter.com/company (Stand: 2016) (Zugriff: 2016-04-25).

TWITTER INC. (2016b): Twitter Developers, https://dev.twitter.com (Stand: 2016) (Zugriff: 2016-04-13).

TWITTER INC. (2018): Developer Agreement and Policy, https://developer.twitter.com/en/developer-terms/agreement-and-policy.html (Stand: 2018-05-25) (Zugriff: 2019-03-22).

TWITTER INC. (2019): Selected Company Financials and metrics.

UNIVERSITÄT ROSTOCK (2017): Sicherheitskonzept für die IT-Infrastruktur, https://www.itmz.uni-rostock.de/footer/ordnungen-und-regeln/sicherheitskonzept-fuer-die-it-infrastruktur/ (Stand: 2017) (Zugriff: 2017-03-06).

VAPNIK, V.N. (1998[2]): The nature of statistical learning theory. New York: Springer.

VETTERMANN, F., SEIP, C. & BILL, R. (2017a): Die Hanse Sail 2016 auf Twitter: Nutzung von Geolokalisation in Sozialen Netzwerken im kleinräumigen Maßstab. In: BILL, R., ZEHNER, M., GOLNIK, A., LERCHE, T., SCHRÖDER, J. & SEIP, S. (Hrsg.). Mit Geoinformation Planen. Berlin: GITO-Verlag, 123–131.

VETTERMANN, F., SEIP, C. & BILL, R. (2017b): Using Twitter for Geolocation Purposes during the Hanse Sail 2016 in Rostock. In: OTJACQUES, B., HITZELBERGER, P., NAUMANN, S. & WOHLGEMUTH, V. (Hrsg.). From Science to Society: New Trends in Environmental Informatics. Progress in IS. Cham: Springer, 171–180.

VETTERMANN, F., WEINZIERL, T. & BILL, R. (2018): Monitoring von Twitter-Nachrichten zur raumzeitlichen und thematischen Analyse in der Hansestadt Rostock: Der Twittermonitor Rostock. – GisScience, 1, 1–9.

VICENTE, R.C., FRENI, D., BETTINI, C. & JENSEN, C.S. (2011): Location-Related Privacy in Geo-Social Networks. – IEEE Internet Computing 15, 3, 20–27.

W3C (2009): SKOS Simple Knowledge Organization System Reference, https://www.w3.org/TR/skos-reference/ (Stand: 1.0.2009) (Zugriff: 2019-03-19).

WANG, Y., ZHOU, Z., JIN, S., LIU, D. & LU, M. (2017): Comparisons and Selections of Features and Classifiers for Short Text Classification. In: ZHANG, D. & RUXU, D. (Hrsg.). International Conference on Artificial Intelligence Applications and Technologies 261. Hawaii.

WARREN, S. & BRANDEIS, L. (1891): The Right to Privacy. – Harvard Law Review 4, 5, http://groups.csail.mit.edu/mac/classes/6.805/articles/privacy/Privacy_brand_warr2.html (Zugriff: 2019-03-19).

WESTERHOLT, R., STEIGER, E., RESCH, B. & ZIPF, A. (2016): Abundant Topological Outliers in Social Media Data and Their Effect on Spatial Analysis. – PloS one 11, 9.

WITTEN, I.H. & FRANK, E. (2005[2]): Data Mining: Pracitcal Machine Learning Tools and Techniques. San Francisco: Elsevier.

WOLD, H.M., VIKRE, L., GULLA, J.A., ÖZGÖBEK, Ö. & SU, X. (2016): Twitter Topic Modeling for Breaking News Detection. In: MAJCHRZAK, T.A., TRAVERSO, P., MONFORT, V. & KREMPELS, K.-H. (Hrsg.). Proceedings of the 12th International Conference on Web Information Systems and Technologies. Setúbal: SCITEPRESS - Science and Technology Publications Lda, 211 - 218.

WU, H.C., LUK, R.W.P., WONG, K.F. & KWOK, K.L. (2008): Interpreting TF-IDF term weights as making relevance decisions. – ACM Transactions on Information Systems 26, 3, 1–37.

YANG, Y. & WEBB, G.I. (2003): On Why Discretization Works for Naive-Bayes Classifiers. In: GEDEON, T.D. & FUNG, L.C.C. (Hrsg.). AI 2003.: Advances in Artificial Intelligence 2903. Berlin, Heidelberg: Springer, 440–452.

YU, H. & ROBINSON, D.G. (2012): The New Ambiguity of 'Open Government'. – UCLA Law Review 178, 59, 178–208.

ZEDNER, L. (2010): Pre-crime and pre-punishment: a health warning. – Criminal Justice Matters 81, 1, 24–25.

ZHANG, G.P. (2000): Neural Networks for Classification: A Survey. – IEEE Transactions on Systems, Man and Cybernetics, Part C 30, 4, 451–462.

ZHANG, J., TANG, J. & LI, J. (2007): Expert Finding in a Social Network. In: KOTAGIRI, R., KRISHNA, P.R., MOHANIA, M. & NANTAJEEWARAWAT, E. (Hrsg.). Advances in Databases: Concepts, Systems and Applications. Lecture Notes in Computer Science 4443. Bangkok: Springer, 1066–1069.

ZHANG, W. & GELERNTER, J. (2014): Geocoding location expressions in Twitter messages: A preference learning method. – Journal of Spatial Information Science, 9, 37–70.

ZHENG, X., HAN, J. & SUN, A. (2018): A Survey of Location Prediction on Twitter. – IEEE Transactions on Knowledge and Data Engineering 30, 9, 1652–1671.

ZHENG, Y. (2011): Location-Based Social Networks: Users. In: ZHENG Y., Z.X. (Hrsg.). Computing with Spatial Trajectories. New York: Springer, 243–276.

ZIEGELE, M. & QUIRING, O. (2012): Privacy in Social Network Sites. In: TREPTE, S. & REINECKE, L. (Hrsg.). Privacy Online: Perspectives on Privacy and Self-Disclosure in the Social Web. Berlin, Heidelberg: Springer, 175–189.

ZOOK, M. (2017): Crowd-sourcing the smart city: Using big geosocial media metrics in urban governance. – Big Data & Society 4, 1, 1–13.

Abbildungsverzeichnis

Abbildung 1-1: Compound Annual Growth Rate (CAGR) ausgewählter Medien in Deutschland im Verhältnis zur täglichen Gesamtmediennutzung 1
Abbildung 1-2: Überschwemmungen während des Sturmtiefs Axel am 04.01.2017 3
Abbildung 2-1: Offene und kommerzielle Angebote von Geodaten 13
Abbildung 2-2: Entwicklung Sozialer Netzwerke.....16
Abbildung 2-3: Komponenten und Rahmenbedingungen einer GDI 20
Abbildung 2-4: Hierarchie der GDIs in Europa.....21
Abbildung 2-5: Funktionsweise von LBSNs.....24
Abbildung 3-1: Die fünf V's von Big Data.....28
Abbildung 3-2: Grundlegende Struktur bei dem Verarbeitungszyklus von Big Data.....29
Abbildung 3-3: Die vier Schritte beim Text Mining 32
Abbildung 3-4: Überwachte (links) und unüberwachte (rechts) Klassifikation. 33
Abbildung 3-5: Überblick über die Beziehungen verschiedener Klassifikationsverfahren zueinander 34
Abbildung 3-6: Beispiel eines Entscheidungsbaums bezüglich der Wettereinschätzung. 35
Abbildung 3-7: Beispiel eines einfachen feedforward neuronalen Netzwerks 36
Abbildung 3-8: Zuweisung eines Themas zu einem Textelement.....37
Abbildung 3-9: Konzept des Vektormodells bei Wörtern anhand der Anfrage König - Mann + Frau = ? 44
Abbildung 3-10: Berechnung des Vektormodells mittels CBAG (links) und N-Grammen (rechts) 45
Abbildung 4-1: Funktionsweise der Twitter Streaming API 51
Abbildung 4-2: Funktionsweise der Twitter REST API.....51
Abbildung 4-3: Client-Server-Prinzip bei OL.....52
Abbildung 4-4: Architektur von GeoNetwork.....54
Abbildung 4-5: Grundfunktionen einer MapServer-Anwendung.....54
Abbildung 5-1: Regiopole Rostock 58
Abbildung 5-2: Überblick über die Stadtbereiche der Hansestadt Rostock mit ausgewählten Points of Interest (POI). 60
Abbildung 5-3: Bevölkerungssaldo der Hansestadt Rostock 61
Abbildung 5-4: Aktuelle und prognostizierte Bevölkerungsstruktur der Hansestadt Rostock.....62
Abbildung 5-5: Bevölkerungsentwicklung der Hansestadt Rostock bis 2025. 62
Abbildung 5-6: Web-Applikation von Klarschiff.HRO. 63
Abbildung 5-7: Prozessabläufe in Klarschiff.HRO. 64
Abbildung 5-8: Web-Interface von OpenData.HRO. 65
Abbildung 5-9: Umfragestruktur der KOGGE Online-Befragung. 66
Abbildung 6-1: Screenshot vom Frontend des Datenportals.....70
Abbildung 6-2: Grundlegende Datenstruktur des Datenportals auf Basis von Thesauri..70
Abbildung 6-3: Datenbankstruktur. 72
Abbildung 6-4: Struktur des vierschichtigen Gazetteers. 73
Abbildung 6-5: Verteilung der Ortsnamen in Rostock. 74
Abbildung 6-6: Grundlegender Aufbau des Streaming-Algorithmus.....75
Abbildung 6-7: Warteschlange bei Twitter (TWITTER INC. 2016b).....76
Abbildung 6-8: Grundlegende Struktur des Harvesters. 76
Abbildung 6-9: Vergleich der Trainingsgenauigkeit (Rot) und der Validierungsgenauigkeit (Grün) für mehrere Klassifikatoren zur Sentimentbestimmung. 83

Abbildung 6-10: Vergleich der unterschiedlichen Text-Vectorizer auf Basis des SB10k Korpus und Klassifikation in Positiv, Neutral und Negativ. 84
Abbildung 6-11: Genauigkeit des SVM-Klassifikators im Verhältnis zur Anzahl an Trainingsdaten. 85
Abbildung 6-12: word2vec-Modell vor (links) und nach der Verwendung der Word-Embeddings. 86
Abbildung 6-13: Struktur des CNN und des LSTM zur Sentimentanalyse. 87
Abbildung 6-14: Screenshot des Twittermonitors. 90
Abbildung 7-1: Absolute Anzahl der Tweets pro Tag im Untersuchungszeitraum (06.08.2019 – 30.09.2018). 93
Abbildung 7-2: Absolute Anzahl der Tweets pro Woche im Untersuchungszeitraum (06.08.2018 – 30.09.2018). 94
Abbildung 7-3: Mittlerer Tagesgang der Tweets im Untersuchungszeitraum (06.08.2018 - 30.09.2018). 94
Abbildung 7-4: Anamorphe Darstellung der Stadtbereiche auf Basis der aufsummierten Tweets der Ebene eins und tiefer im Untersuchungszeitraum (06.08.2018 - 30.09.2018). Anzahl der Tweets je Stadtbereich dargestellt. 95
Abbildung 7-5: Anamorphe Darstellung der Stadtbereiche auf Basis der aufsummierten Follower je Tweet der Ebene eins und tiefer im Untersuchungszeitraum (06.08.2018 - 30.09.2018). Anzahl der Follower je Stadtbereich in 10 tsd. dargestellt. 96
Abbildung 7-6: Signifikante Hot- (positives Sentiment) und Coldspots (negatives Sentiment) berechnet mittels Morans I aus der Gesamtwahrscheinlichkeit der Sentimentzuordnung. 99
Abbildung 7-7: Signifikante Hot- (positives Sentiment) und Coldspots (negatives Sentiment) berechnet mittels Morans I aus der Zuordnungswahrscheinlichkeit von 95 % zu positiven bzw. negativen Tweets. 99
Abbildung 7-8: Signifikante Hot- (positives Sentiment) und Coldspots (negatives Sentiment) berechnet mittels Morans I aus der Zuordnungswahrscheinlichkeit von 90 % zu positiven bzw. negativen Tweets. 100
Abbildung 7-9: Signifikante Hot- (positives Sentiment) und Coldspots (negatives Sentiment) berechnet mittels Morans I aus der Zuordnungswahrscheinlichkeit von 85 % zu positiven bzw. negativen Tweets. 100
Abbildung 7-10: Signifikante Hot- (positives Sentiment) und Coldspots (negatives Sentiment) berechnet mittels Morans I aus der Zuordnungswahrscheinlichkeit von 80 % zu positiven bzw. negativen Tweets. 101
Abbildung 7-11: Kernel Density anhand aller Tweets im Untersuchungszeitraum (06.08.2018 - 30.09.2018) von 0 - 6 Uhr. 102
Abbildung 7-12: Kernel Density anhand aller Tweets im Untersuchungszeitraum (06.08.2018 - 30.09.2018) von 6 - 12 Uhr. 103
Abbildung 7-13: Kernel Density anhand aller Tweets im Untersuchungszeitraum (06.08.2018 - 30.09.2018) von 12 - 18 Uhr. 103
Abbildung 7-14: Kernel Density anhand aller Tweets im Untersuchungszeitraum (06.08.2018 - 30.09.2018) von 18 - 0 Uhr. 104
Abbildung 7-15: Zeitlicher Verlauf der Tweets mit Bezug zur AfD-Demonstration am 22.09.2018. 104
Abbildung 7-16: Raum-zeitlicher Verlauf der Demonstrationen am 22.09.2018. 106
Abbildung 7-17: Anteil der zehn Accounts mit den meisten originären Tweets. 107
Abbildung 7-18: Rezeption und offizielle Meldung verschiedener Ereignisse der Demonstrationen am 22.09.2018. 108
Abbildung 7-19: Wordcloud der diskutierten Themen in der Hansestadt Rostock im Untersuchungszeitraum 110

Abbildung 7-20: Anzahl an Themen und deren Kohärenz mittels gensim LDA. 112
Abbildung 7-21: SVD-Interpolation der LDA mittels gensim auf Basis von sieben Klassen (links) und zwölf Klassen (rechts). 113
Abbildung 7-22: Top-Wörter nach Wichtung und Anzahl der einzelnen Themen auf Basis der gensim LDA für den gesamten Untersuchungszeitraum (06.08.2018 - 30.09.2018). 114
Abbildung 7-23: Twitternetzwerk aller Accounts mit Bezug zur Hansestadt Rostock im Untersuchungszeitraum (06.08.2018 - 30.09.2018). 116
Abbildung 7-24: Netzwerkdiagramm der Accounts mit Bezug zur AfD-Demonstration am 22.09.2018 in Rostock. Die Größe und Farbintensität der Punkte steht für die Anzahl an Followern innerhalb des Netzes. 117

Tabellenverzeichnis

Tabelle 2-1: Beispiele der Veränderung der Inhalte von Web 1.0 zu Web 2.0 8
Tabelle 2-2: Kategorisierung von Crowdsourcingsystemen 9
Tabelle 2-3: Klassifizierung Sozialer Netzwerke 17
Tabelle 2-4: Auswahl von OGC Web Services 22
Tabelle 2-5: Klassifizierung von LBSNs 24
Tabelle 3-1: Ausgewählte Text-Similarity-Funktionen 42
Tabelle 4-1: Beschreibung der einzelnen Elemente eines Tweets 49
Tabelle 5-1: Geäußerte Zahlungsbereitschaften (ZB) für naturnahe Fließgewässer und Feuchtgebiete anhand ausgewählter Quellen 67
Tabelle 6-1: Social Media spezifische Inhalte. 77
Tabelle 6-2: Genauigkeiten des SB10k Korpus 81
Tabelle 6-3: Konfusionsmatrix des SVM-Klassifikators. 84
Tabelle 6-4: Konfusionsmatrix des SVM-Klassifikators. 85
Tabelle 7-1: Lokationsgenauigkeit. 92
Tabelle 7-2: Zuordnung des Sentiments je Stadtbereich - > 0.5 eher positiv, <0.5 eher negativ. 98
Tabelle 7-3: Mittleres Sentiment je Thema. 101
Tabelle 7-4: Mittels TF-IDF ermittelte Trends im Untersuchungszeitraum, manuell nach Aussagekraft gefiltert. 111
Tabelle 7-5: Ergebnis der LDA für die KW 38. 113

Anhang

A - 1 Verwendete externe Bibliotheken

Tabelle A - 1: Liste der verwendeten, externen Bibliotheken und Erweiterungen für Python 3 und JavaScript.

Sprache	Bibliothek	Link	Verwendung
Python	fuzzywuzzy	https://github.com/seatgeek/fuzzywuzzy	Textvergleichsbibliothek für das Gazetteermatching
	gensim	https://radimrehurek.com/gensim/	Machine Learning Bibliothek für die LDA
	imblearn	https://imbalanced-learn.readthedocs.io/en/stable/api.html	Nutzung des SMOTE-Oversampling
	keras	https://keras.io/	Machine Learning Bibliothek für NN
	langid	https://pypi.org/project/langid/	Zuordnung der Sprache
	matplotlib	https://matplotlib.org/	Darstellung von Graphen
	nltk	https://www.nltk.org/	Sammlung von Algorithmen zur Textprozessierung
	numpy	http://www.numpy.org/	Handhabung mehrdimensionaler Arrays
	pandas	https://pandas.pydata.org/	Erstellen und Laden von Dataframes
	psycopg2	http://initd.org/psycopg/	Datenbankverbindung zu PostgreSQL über Python
	pyenchant	https://github.com/rfk/pyenchant	Textkorrekturbibliothek
	sklearn	https://scikit-learn.org/stable/	Umfassende Machine Learning Bibliothek, verwendet insbesondere für SMV
	tensorflow	https://www.tensorflow.org/	Machine Learning Bibliothek für NN
	twython	https://twython.readthedocs.io/en/latest/	Python-Bibliothek für den Zugriff auf die Twitter API
	word_cloud	https://github.com/amueller/word_cloud	Erstellen von Wordclouds zur Darstellung
JavaScript	Bootstrap	https://getbootstrap.com/	Responsives Webdesign
	Chart.js	https://www.chartjs.org/	Chartdarstellung in WebApp
	Font Awesome	https://fontawesome.com/	Textsymbole für Buttons etc.
	jQRangeSlider	https://ghusse.github.io/jQRangeSlider/	Datumsslider
	jQuery	https://jquery.com/	JavaScript-Erweiterung, benötigt für die meisten anderen Plugins
	OpenLayers3	https://openlayers.org/	Webmap-Darstellung
	Proj4js	http://proj4js.org/	Projektionsverarbeitung
	Twitter-Widgets	https://developer.twitter.com/en/docs/twitter-for-websites/embedded-tweets/overview	Einbindung der Tweets

A - 2 Sentimentanalyse

Tabelle A - 2: Liste der Stoppwörter, die für die Sentimentanalyse aus dem Nachrichtentext entfernt werden.

aber	demgemäß	en	jedem	seines	wem
acht	demgemaeß	entweder	jeden	seit	wen
achte	demselben	er	jeder	seitdem	wenn
achten	demzufolge	erst	jedermann	selbst	wer
achter	den	es	jedermanns	sich	werde
achtes	denen	etwa	jedes	sie	werden
allerdings	denn	etwas	jedoch	sieben	werdet
allgemeinen	denselben	euch	jemand	siebente	
als	der	euer	jemandem	siebenten	weshalb
also	deren	eure	jemanden	siebenter	wessen
am	derer	eurem	jene	siebentes	wie
an	derjenige	euren	jenem	sind	wieder
au	derjenigen	eurer	jenen	so	wieso
auch	derselbe	eures	jener	solche	wir
auf	derselben	folgende	jenes	solchem	wird
aus	des	früher	kam	solchen	wo
ausser	deshalb	frueher	kann	solcher	woher
ausserdem	desselben	fünf	kannst	solches	wohin
außer	dessen	fuenf	los	soweit	während
außerdem	deswegen	fünfte	manche	sowie	waehrend
bei	dich	fuenfte	manchem	startseite	währenddem
beide	die	fünften	manchen	steht	waehrenddem
beiden	diejenige	fuenften	mancher	suche	währenddessen
beim	diejenigen	fünfter	manches	tag	waehrenddessen
beispiel	dies	fuenfter	mann	tage	zehn
bereits	diese	fünftes	mein	tagen	zehnte
besonders	dieselbe	fuenftes	meine	teil	zehnten
bin	dieselben	für	meinem	tel	zehnter
bis	diesem	fuer	meinen	tun	zehntes
bisher	diesen	gegen	meiner	uhr	zeit
bist	dieser	gegenüber	meines	um	zugleich
da	dieses	gegenueber	mich	und	zum
dabei	dir	gehen	mir	uns	zunächst
dadurch	doch	gekannt	mit	unse	zunaechst
daher	dort	gesagt	mittel	unsem	zwanzig
dahin	drei	gewesen	nahm	unsen	zwar
dahinter	drin	geworden	neben	unser	zwei
damit	dritte	gibt	neun	unsere	zweite
danach	dritten	ging	neunte	unserer	zweiten
dann	dritter	gleich	neunten	unses	zweiter
daran	drittes	gott	neunter	unter	zweites
darauf	du	heisst	neuntes	vielleicht	zwischen
daraus	durch	her	nun	vier	zwölf
darum	durchaus	hin	ob	vierte	zwoelf
darunter	eben	ich	oder	vierten	übrigens
darüber	ebenso	ihm	rechten	vierter	uebrigens
darueber	ei	ihn	rechter	viertes	
das	eigen	ihnen	rechtes	vom	
dasein	eigene	ihr	rund	von	
daselbst	eigenen	ihre	sa	vor	
dass	eigener	ihrem	sache	wann	
davon	eigenes	ihren	sagt	war	
davor	ein	ihrer	sagte	waren	
dazu	einander	ihres	sah	warst	
dazwischen	eine	im	satt	wart	
daß	einem	immer	sechs	warum	
dein	einen	in	sechste	was	
deine	einer	indem	sechsten	weil	
deinem	eines	infolgedessen	sechster	weit	
deinen	einig	ins	sechstes	weiter	
deiner	einige	irgend	sei	weitere	
deines	einigem	ist	seid	weiteren	
dem	einigen	ja	seien	weiteres	
dementspreched	einiger	jahr	sein	welche	
demgegenüber	einiges	jahre	seine	welchem	
demgegenueber	einmal	jahren	seinem	welchen	
demgemäss	eins	je	seinen	welcher	
demgemaess	elf	jede	seiner	welches	

A - 3 Nachrichtenverteilung

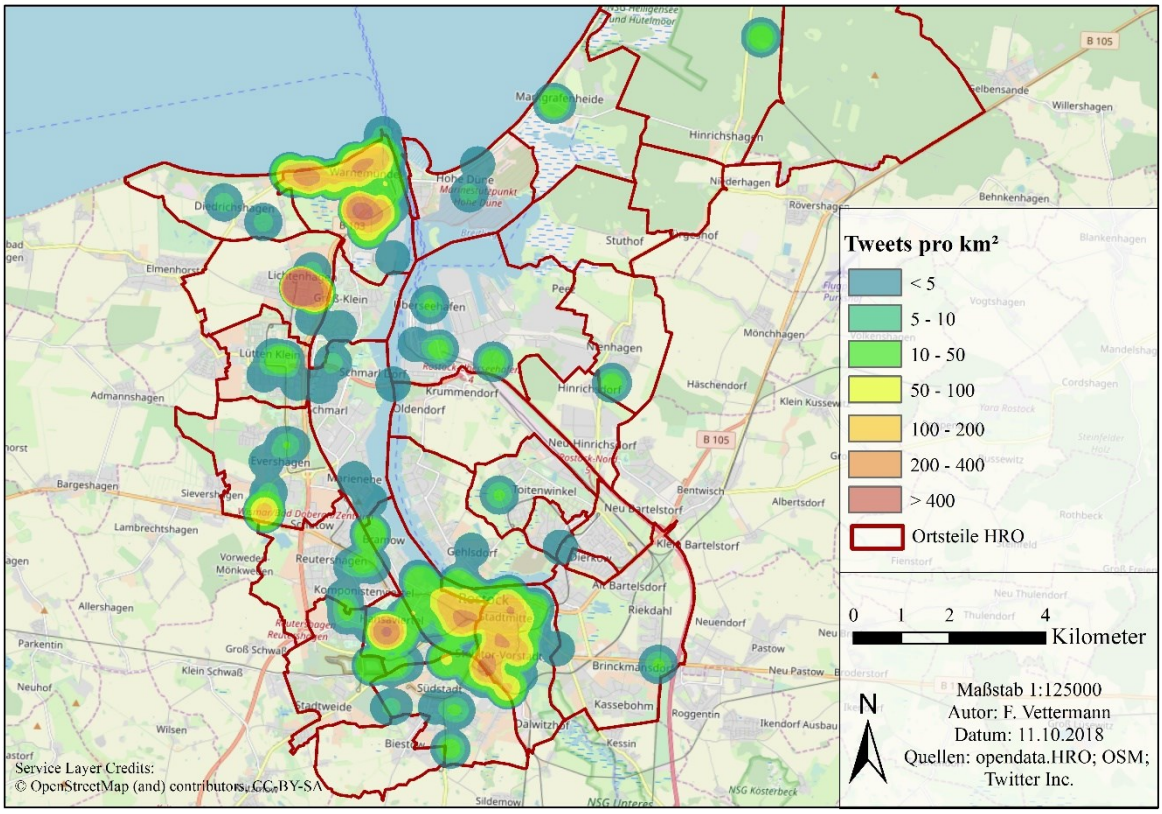


Abbildung A - 1: Kernel Density der Tweets der KW 32 2018.

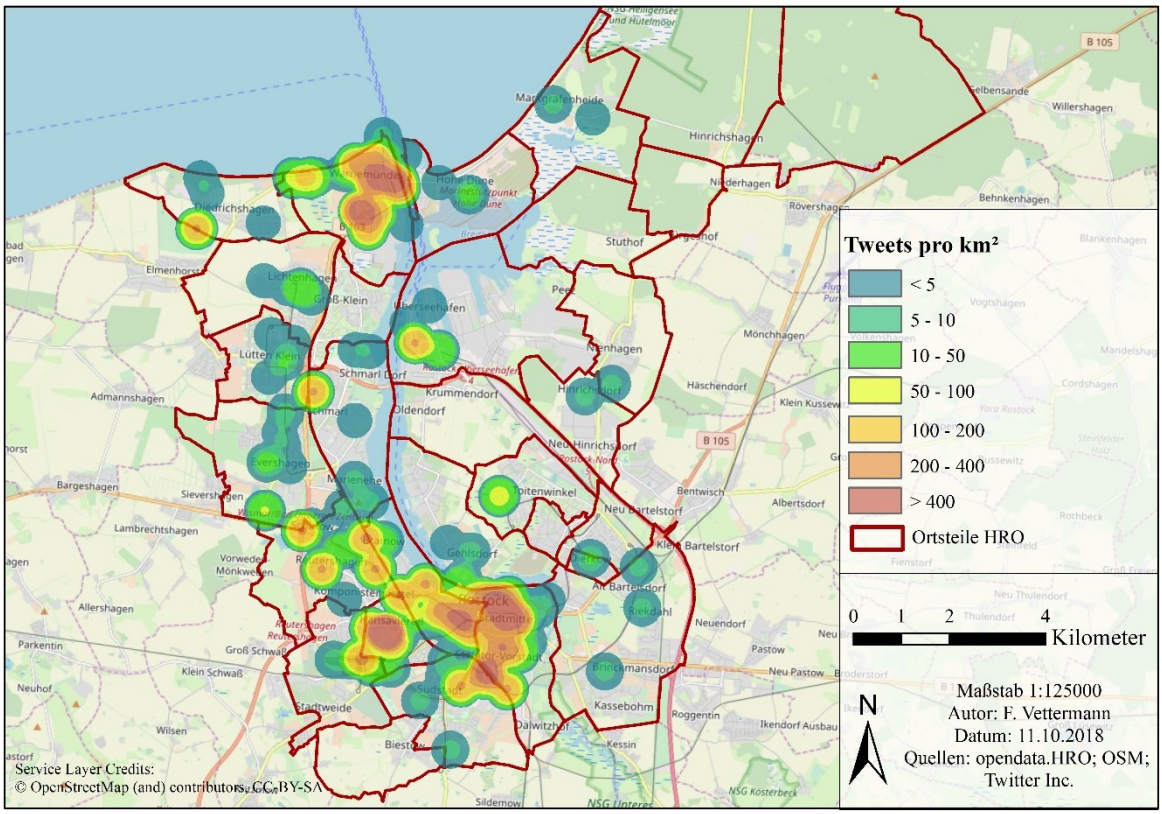


Abbildung A - 2: Kernel Density der Tweets der KW 33 2018.

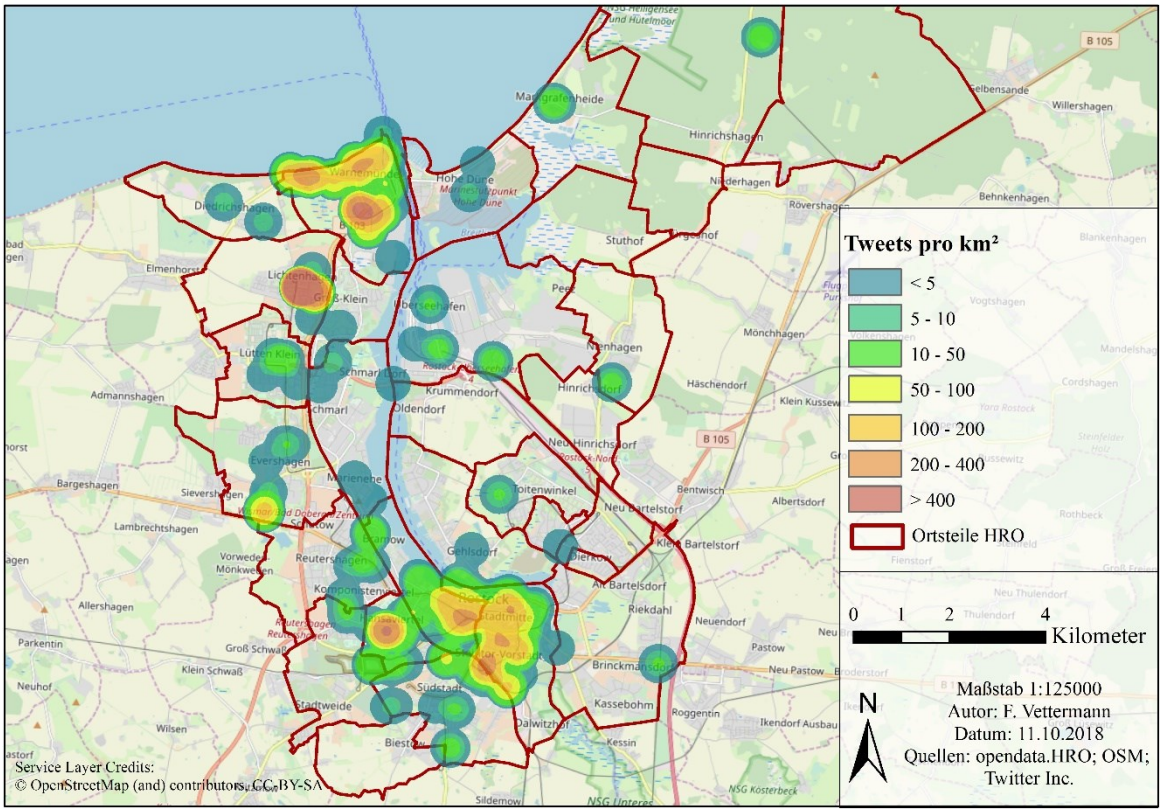


Abbildung A - 3: Kernel Density der Tweets der KW 34 2018.

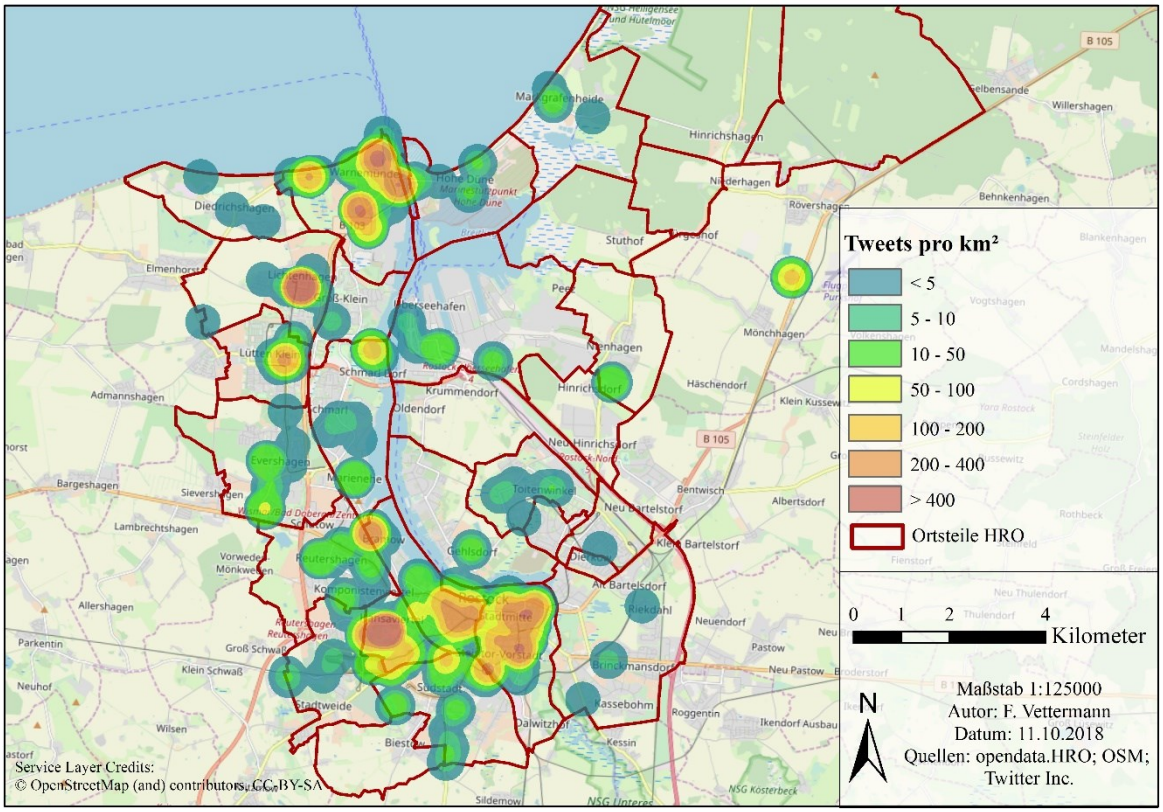


Abbildung A - 4: Kernel Density der Tweets der KW 35 2018.

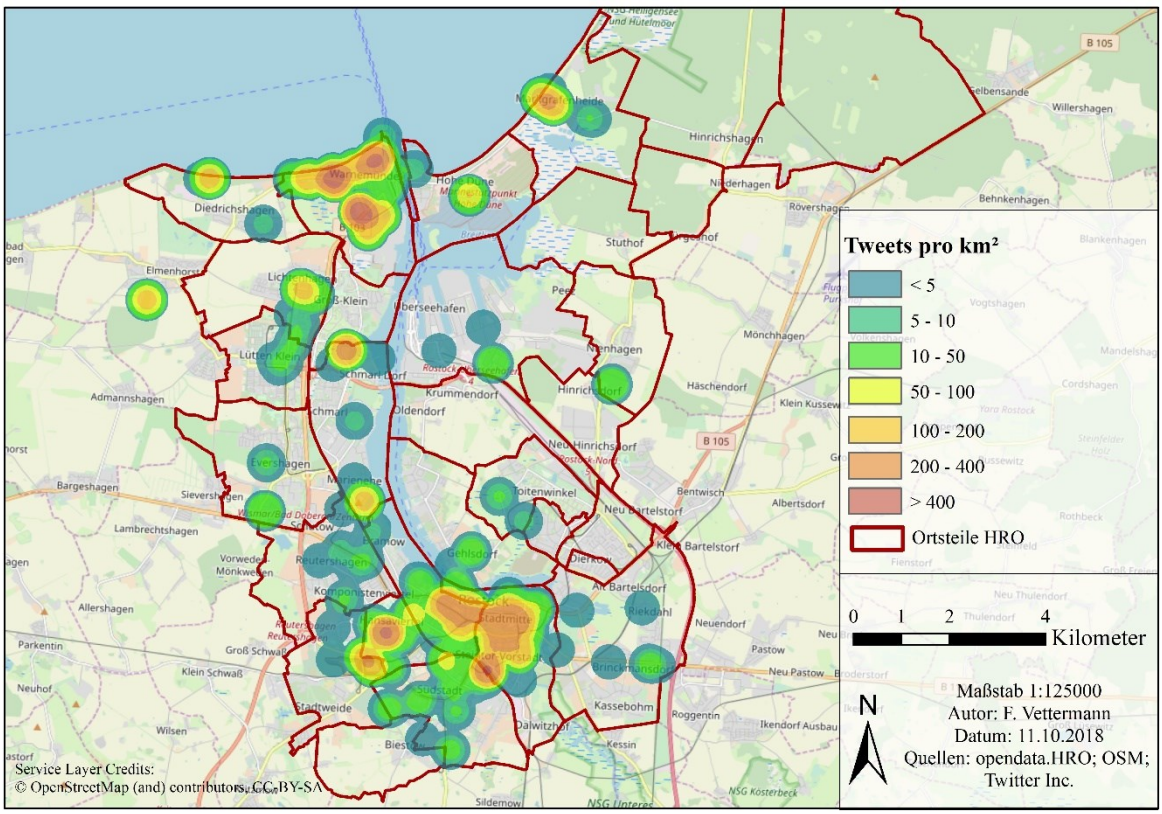


Abbildung A-5: Kernel Density der Tweets der KW 36 2018.

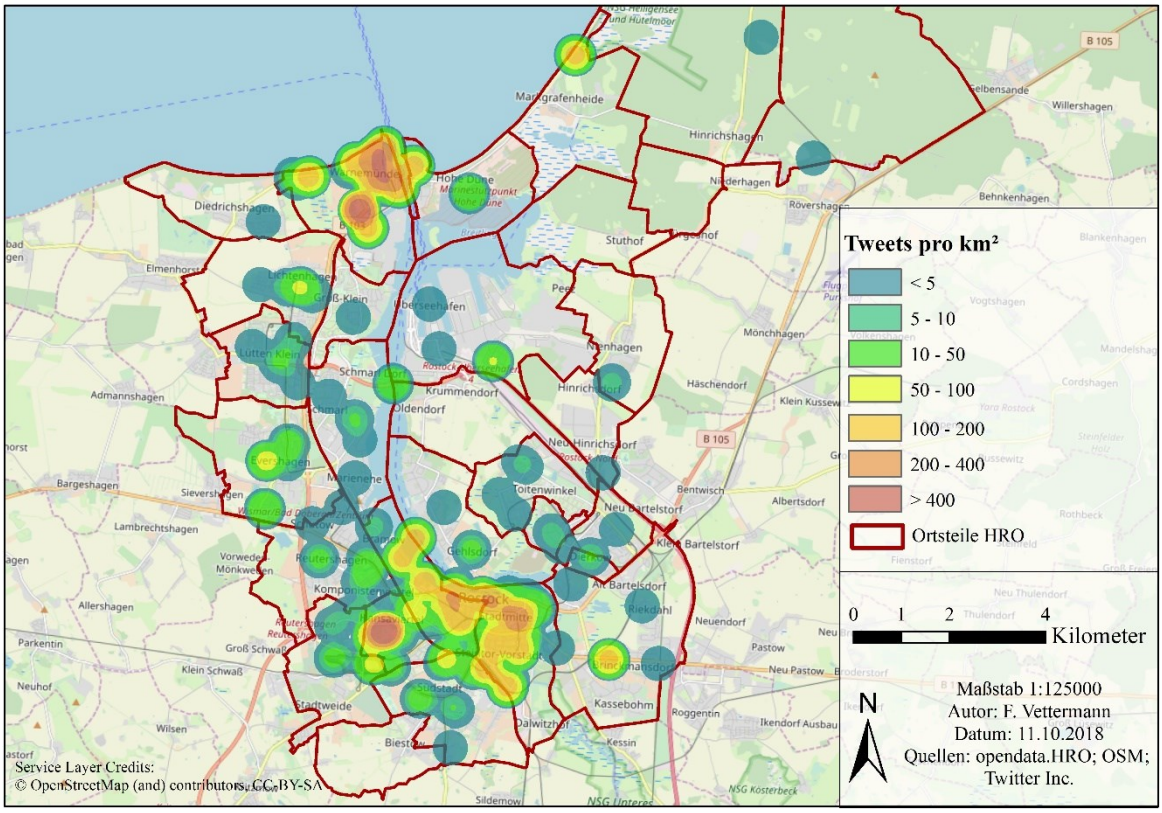


Abbildung A-6: Kernel Density der Tweets der KW 37 2018.

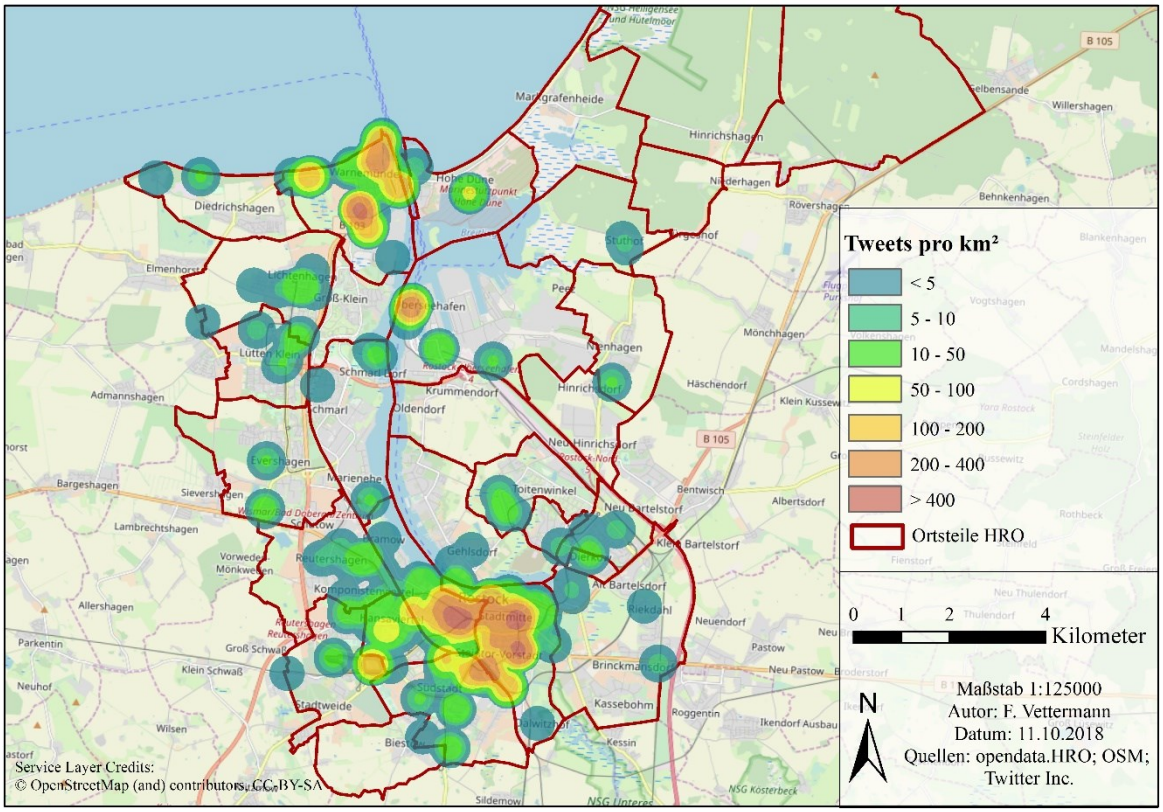


Abbildung A - 7: Kernel Density der Tweets der KW 38 2018.

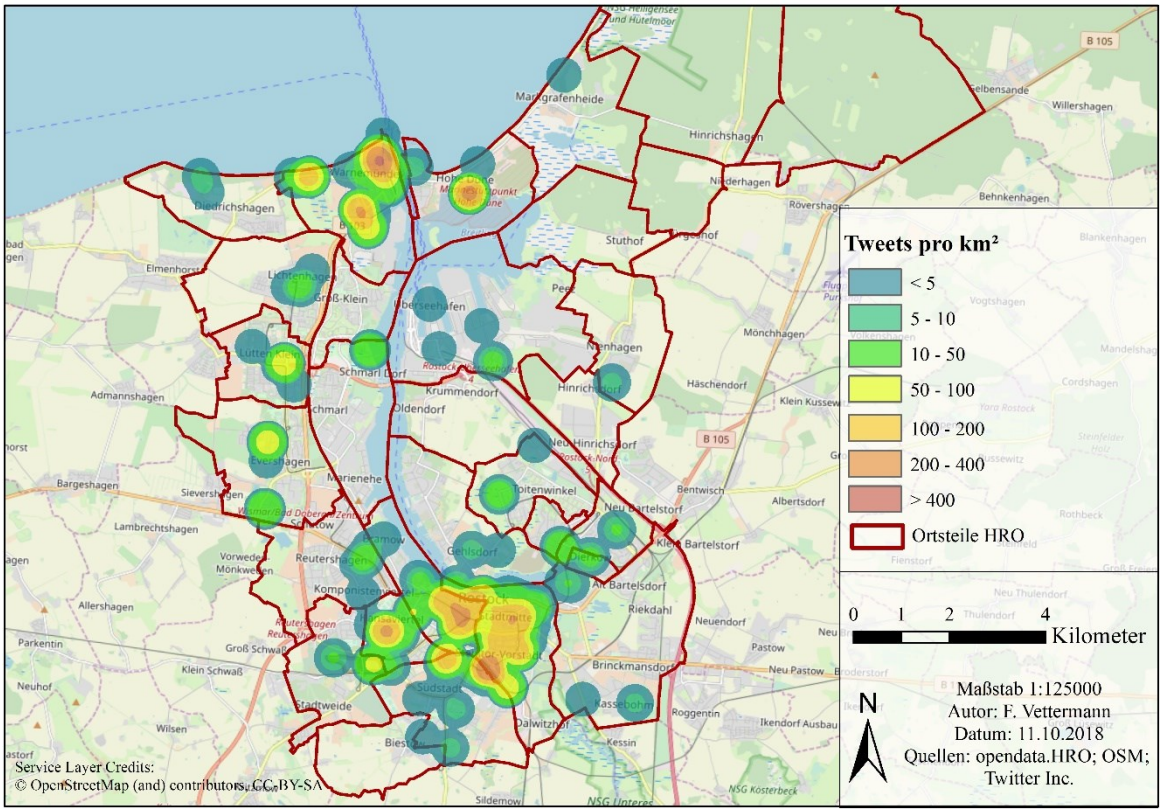


Abbildung A - 8: Kernel Density der Tweets der KW 39 2018.

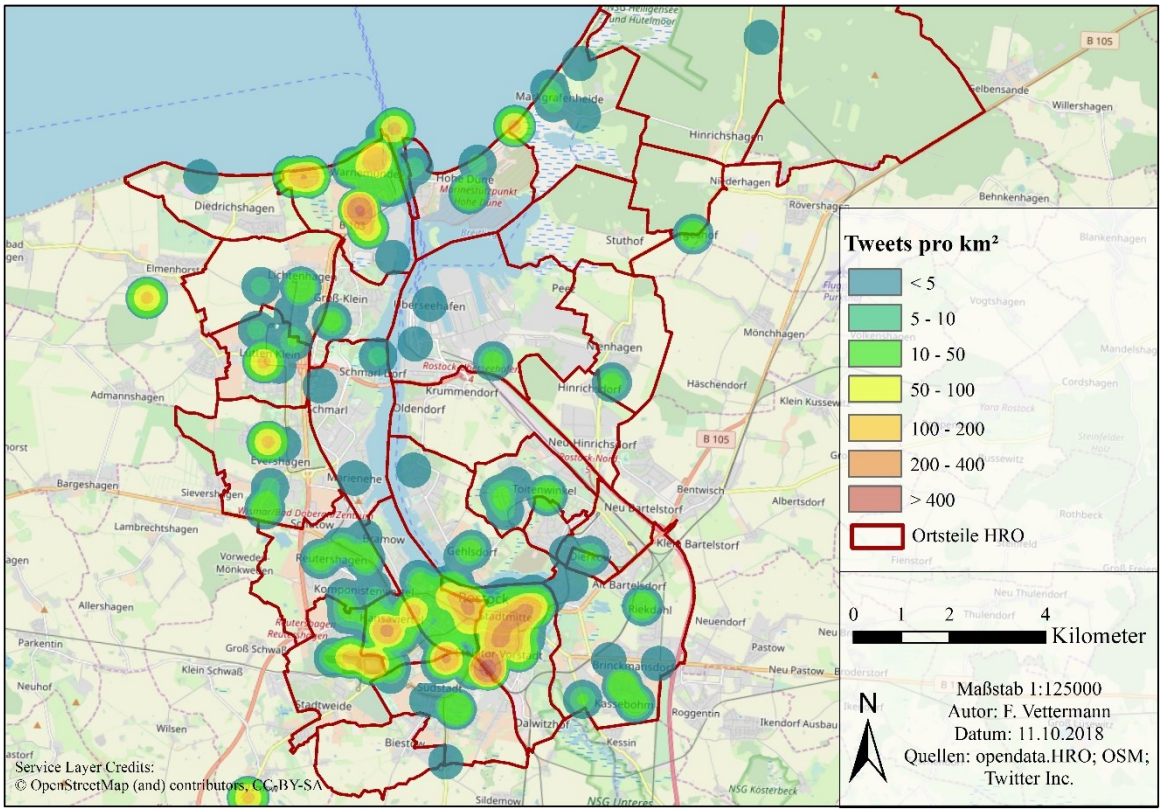


Abbildung A - 9: Kernel Density der Tweets der KW 40 2018.

Tabelle A - 3: Nachrichten je Thematik im Untersuchungszeitraum (06.08.2018 - 30.09.2018) nach Anzahl der Tweets je Stadtbereich.

	Tweets pro Einwohner und Woche
Biestow	0.00007
Brinckmansdorf	0.00056
Dierkow-Neu	0.00021
Dierkow-Ost	0.00182
Dierkow-West	0.00021
Evershagen	0.00119
Gartenstadt/Stadtweide	0.00945
Gehlsdorf	0.00098
Groß Klein	0.00007
Hansaviertel	0.04298
Kröpeliner-Tor-Vorstadt	0.00763
Lichtenhagen	0.00091
Lütten Klein	0.01246
Reutershagen	0.00077
Rostock-Heide	0.01764
Rostock-Ost	0.01841
Schmarl	0.00371
Stadtmitte	0.00126
Südstadt	0.01652
Toitenwinkel	0.00056
Warnemünde	0.03815

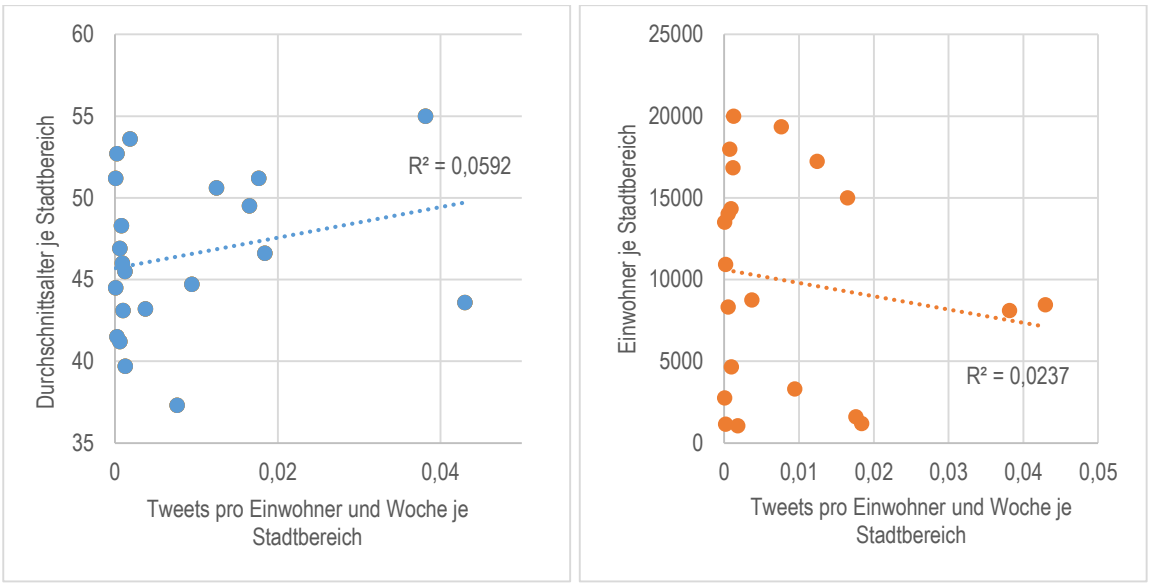


Abbildung A - 10: Korrelation zwischen Durchschnittsalter (links) und Einwohnerzahl rechts mit den Tweets pro Einwohner und Woche.

A - 4 Themenzuordnung

Tabelle A - 4: Nachrichten je Thematik im Untersuchungszeitraum (06.08.2018 - 30.09.2018) nach Anzahl der Tweets je Stadtbereich.

	Veranstal-tung	Sport	Sicherheit	Wasser	Urlaub	Arbeit	Sonstige	Gesamt	Anteil in %	Main-To-pic
Biestow	0	0	0	0	0	1	0	1	0.01	keine
Brinckmansdorf	0	0	1	1	1	1	32	36	0.28	keine
Dierkow-Neu	0	0	2	0	1	0	17	20	0.16	keine
Dierkow-Ost	0	0	2	0	1	0	12	15	0.12	keine
Dierkow-West	0	0	0	0	0	0	2	2	0.02	keine
Evershagen	0	6	24	3	9	10	107	159	1.25	Sport
Gartenstadt/Stadtweide	3	17	12	4	15	3	189	243	1.91	Sport
Gehlsdorf	1	0	1	0	3	1	29	35	0.28	keine
Groß Klein	0	0	0	0	2	0	2	4	0.03	keine
Hansaviertel	6	2 623	61	47	197	36	136	3 106	24.43	Sport
Kröpeliner-Tor-Vorstadt	7	53	594	43	61	34	424	1 216	9.56	Sicherheit
Lichtenhagen	1	6	598	2	20	2	752	1381	10.86	Sicherheit
Lütten Klein	1	5	2	13	9	11	89	130	1.02	Wasser
Reutershagen	4	3	9	3	10	5	75	109	0.86	Urlaub
Rostock-Heide	19	7	11	11	10	18	154	230	1.81	Veranstal-tung
Rostock-Ost	2	5	20	21	9	32	95	184	1.45	Arbeit
Schmarl	3	3	4	1	4	2	69	86	0.68	Sicherheit
Stadtmitte	74	162	439	157	152	78	1 639	2 701	21.24	Sicherheit
Südstadt	9	57	53	6	36	14	274	449	3.53	Sport
Toitenwinkel	0	0	0	0	2	3	52	57	0.45	keine
Warnemünde	183	469	107	53	330	36	1 373	2 551	20.06	Sport

Tabelle A - 5: Nachrichten je Thematik im Untersuchungszeitraum (06.08.2018 - 30.09.2018) nach Anzahl der Follower in 10 tsd. je Stadtbereich.

	Veranstalt-ung	Sport	Sicherheit	Wasser	Urlaub	Arbeit	Sonstige	Gesamt	Anteil in %	Main-Topic
Biestow	0.00	0.00	0.00	0.00	0.00	0.16	0.00	0.16	0.00	keine
Brinckmansdorf	0.00	0.00	0.00	0.01	0.11	0.06	9.01	7.62	0.05	keine
Dierkow-Neu	0.00	0.00	0.40	0.00	0.17	0.00	4.65	5.22	0.04	keine
Dierkow-Ost	0.00	0.00	0.49	0.00	0.07	0.00	2.90	2.11	0.01	keine
Dierkow-West	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.25	0.00	keine
Evershagen	0.00	0.51	4.78	0.96	0.55	2.24	644.93	653.87	4.43	Sicherheit
Gartenstadt/Stadtweide	2.64	151.75	0.59	0.23	0.51	128.30	324.15	608.08	4.12	Sport
Gehlsdorf	0.43	0.00	0.04	0.00	2.64	0.03	2.72	5.87	0.04	Urlaub
Groß Klein	0.00	0.00	0.00	0.00	0.40	0.00	0.06	0.46	0.00	keine
Hansaviertel	3.37	2 498.09	111.48	16.55	164.84	30.33	154.22	2 766.92	18.75	Sport
Kröpeliner-Tor-Vorstadt	11.99	17.09	1 505.92	296.88	168.73	136.67	544.50	2 346.50	15.90	Sicherheit
Lichtenhagen	0.10	0.24	215.28	0.25	2.27	1.57	102.55	321.86	2.18	Sicherheit
Lütten Klein	0.03	0.12	0.67	11.98	1.09	4.49	17.88	36.08	0.24	Wasser
Reutershagen	0.06	11.74	13.39	0.03	11.94	0.44	91.93	129.04	0.87	Sicherheit
Rostock-Heide	17.17	1.57	3.32	8.02	8.05	0.75	1 071.15	1 093.26	7.41	Veranstal-tung
Rostock-Ost	1.54	1.33	37.11	22.58	1.44	2.19	82.63	146.88	1.00	Sicherheit
Schmarl	0.58	1.36	1.59	0.12	11.26	1.41	36.50	52.60	0.36	Urlaub
Stadtmitte	15.38	233.40	1 572.06	134.47	194.52	308.21	1 599.88	3 820.22	25.88	Sicherheit
Südstadt	12.19	187.23	187.48	0.02	79.88	12.79	607.29	959.47	6.50	Sicherheit
Toitenwinkel	0.00	0.00	0.00	0.00	2.04	0.81	26.50	28.68	0.19	Urlaub
Warnemünde	58.18	741.98	73.26	7.97	175.14	6.93	826.62	1 775.16	12.03	Sport

A - 5 Themenfindung

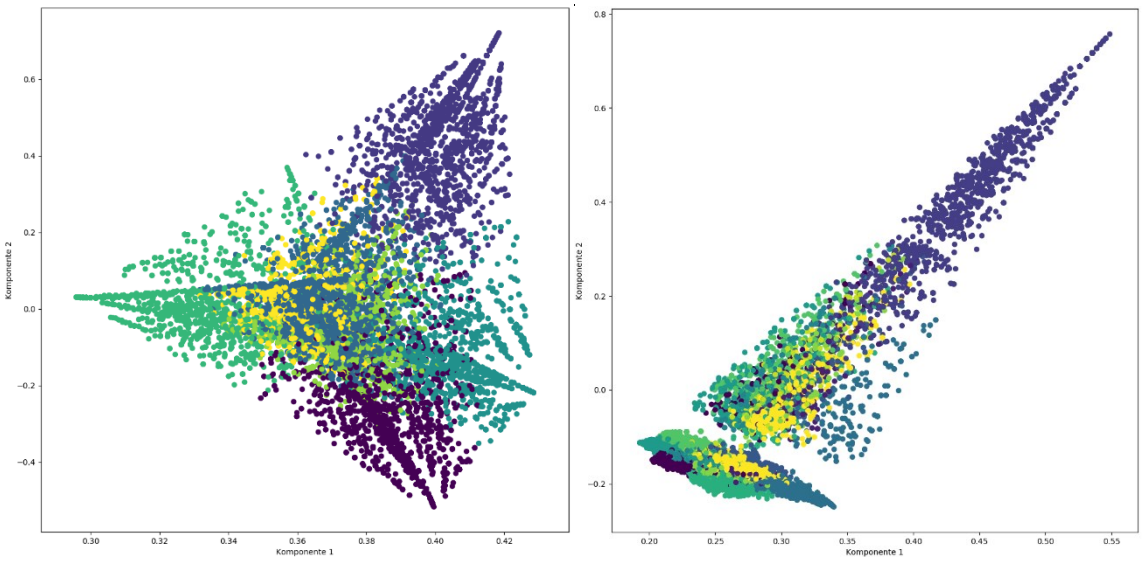


Abbildung A - 11: SVD-Interpolation der LDA mittels Scikit-Learn auf Basis von sieben Klassen (links) und zwölf Klassen (rechts).

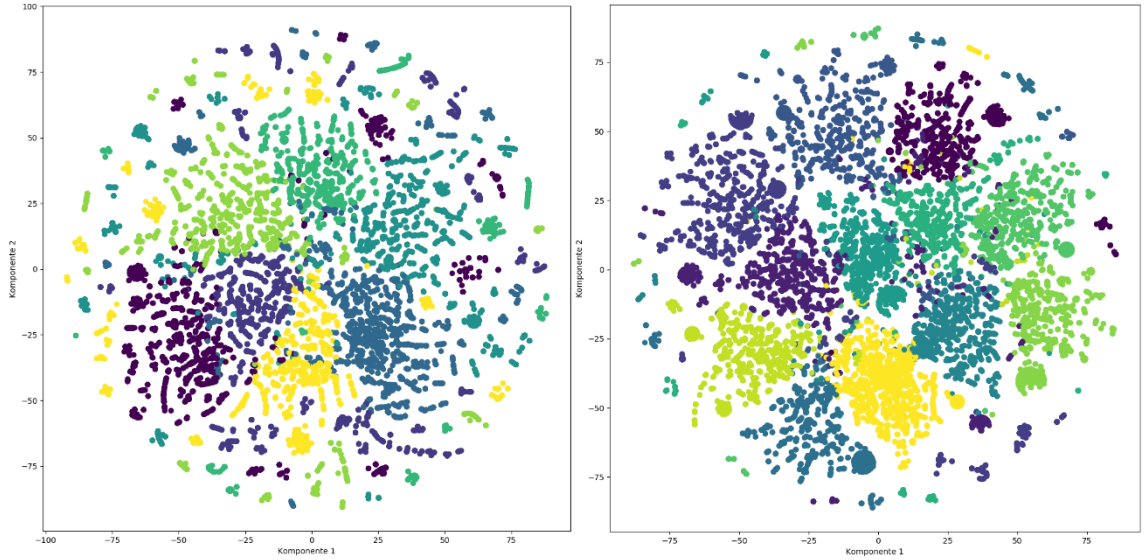


Abbildung A - 12: t-SNE-Interpolation der LDA mittels Scikit-Learn auf Basis von sieben Klassen (links) und zwölf Klassen (rechts).

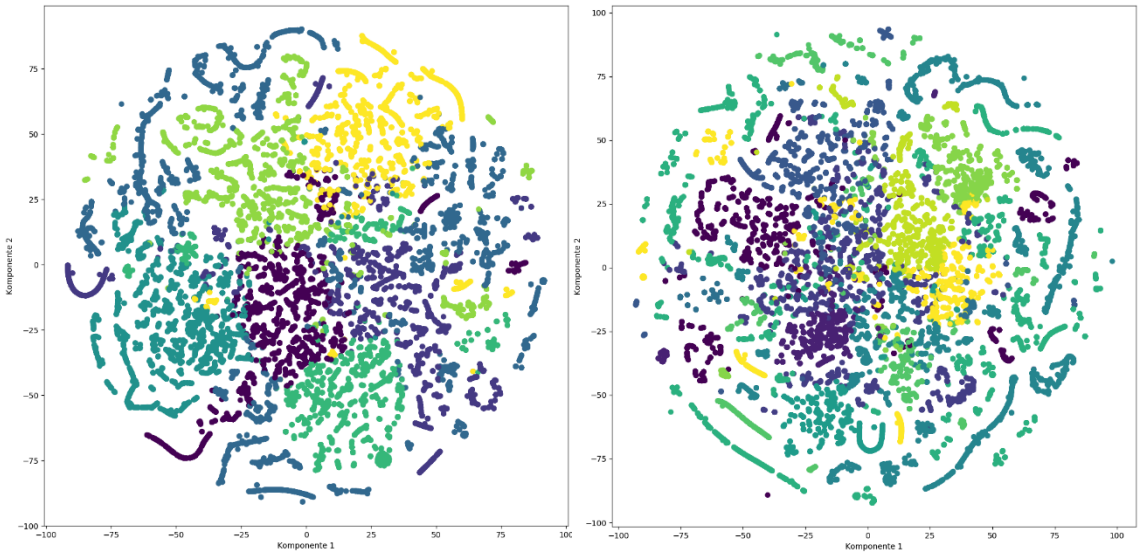


Abbildung A - 13: t-SNE-Interpolation der LDA mittels gensim auf Basis von sieben Klassen (links) und zwölf Klassen (rechts).

Tabelle A - 6: Wöchentliche Themen und deren Schlagwörter abgeleitet aus der gensim LDA.

KW	Thema	Schlagwörter	Max. Gewicht
32	1	rostock, richtung, stau, ostseestadion	0.077
	2	rostock, bekommt, reicht, rest, monat	0.095
	3	rostock, nsu, ort, mordes, enttarnung	0.083
	4	rostock, hbf, fotos, warnemünde, stadthafen	0.048
	5	rostock, sail, hanse, aktuell, nächsten	0.132
	6	hansesail, warnemünde, rostock, freitag, seebrücke	0.213
	7	Rostock, hansa, unterhaching, unfall, spvgg	0.129
33	1	arzt, ostseestadion, schlägt, kogge, jaaa	0.105
	2	rostock, stau, flixbus, liga, verletzte	0.087
	3	offenburg, rostock, via, warnemünde, schwaben	0.102
	4	rostock, stuttgart, vfb, hansa, dfbpokal	0.135
	5	rostock, berlin, fc, richtung, polizei	0.119
	6	rostock, syrerin, baby, geschubst, übergriff	0.156
	7	teich, rostock, schubst, nazis, feige	0.226
34	1	rostock, polizei, hansa, polizisten, würzburger	0.119
	2	rostock, lichtenhagen, deutsche, begannen	0.131
	3	chemnitz, rostock, mob, august, bilder	0.092
	4	rostock, fehler, regierung, bringt, wiederholen	0.072
	5	rostock, lichtenhagen, jahrestag, jagen	0.089
	6	rostock, fc, baby, jagdszenen, eskaliert	0.077
	7	rostock, liga, massage, syrerin, geschubst	0.094
35	1	rostock, chemnitz, rostocker, mölln, solingen	0.145
	2	abend, warnemünde, shop, mädchen, bilder	0.069
	3	rostock, hansa, liga, studieren, hansarostock	0.103
	4	rostock, sex, massage, porno, escort	0.073
	5	rostock, ficken, gruppe, klick, geile	0.084
	6	rostock, deutsche, verstärkung, phil, premierleague	0.087
	7	ostseestadion, rostock, leben, team schönster	0.091
36	1	rostock, hamburg, frauen, mädchen, treffen	0.086
	2	rostock, altem, syrische, monate, großmutter	0.064
	3	warnemünde, rostock, richtung, hafen, mv	0.059
	4	rostock, hansarostock, hansa, liga, fch	0.097
	5	rostock, mittelmeer, flüchtlingen, aufnahme, bereit	0.113
	6	rostock, dresden, teen, wasser, oh	0.060
	7	rostock, gefickt, halt, dunjahayali, verlassen	0.102
37	1	rostock, warnemünde, gezeigt, blog, checkt	0.083
	2	rostock, tsv, september, via, berlin	0.123
	3	rostock, hambibleibt, stadion, zug, ddr	0.059
	4	rostock, hansa, höcke, verhindern, mädchen	0.117
	5	rostock, treffen, lübeck, klick, faschistenbande	0.058
	6	rostock, hambacherforst, warnemünde, liebe, chemnitz	0.039
	7	rostock, münchen, afd, ostseestadion, cruise	0.105
39	1	rostock, warnemünde, Samstag, schauspieler, frank	0.099
	2	rostock, marokkanische, münchen, nordosten, arkona	0.048
	3	rostock, hansa, halle, abgesagt, liga	0.086
	4	rostock, teilnehmern, veranstaltung, zitat, hbf	0.110
	5	rostock, vermisst, domscheit, ots, kessler	0.092
	6	rostock, archivtag, dat, noafd, fcknzs	0.135
	7	rostock, münster, rostocker, afd, deutschland	0.062