

Masterarbeit

zum Thema

Twitter während der COVID-19-Pandemie - Eine Analyse mit Techniken des Deep Learnings

Vorgelegt der Fakultät für Wirtschaftswissenschaften
der Universität Duisburg-Essen (Campus Essen)

von:

Paul Drecker

xxx

xxx

Erstgutachter: xxx

Zweitgutachter: xxx

Aktuelles Semester: Wintersemester 2020/2021, 7. Fachsemester

Studiengang: Master Volkswirtschaftslehre

Voraussichtlicher Studienabschluss: Wintersemester 2020/2021

Abgabetermin: 04.02.2021

Inhaltsverzeichnis

Inhaltsverzeichnis	I
Symbolverzeichnis	IV
Abkürzungsverzeichnis	IX
Abbildungsverzeichnis	XI
Tabellenverzeichnis	XIII
Abstract	1
Zusammenfassung	1
1 Einleitung	2
2 Twitter und Krisenkommunikation	5
2.1 Twitter	5
2.2 Kommunikation und staatliche Kommunikation auf Twitter während Krisen	7
2.2.1 Kommunikation auf Twitter in Krisenzeiten	7
2.2.2 Staatliche Krisenkommunikation auf Twitter während eines Krankheitsausbruches	9
3 Theoretische Grundlagen der Methodik	12
3.1 <i>Deep-Learning</i>	12
3.2 <i>Natural-Language-Processing</i> und semantischer Raum	14
3.3 <i>Topic-Modeling</i> von Twitter-Daten	16
3.4 Sentiment-Analysen	19
3.5 Soziale Netzwerkanalyse	23
4 Methodik und Daten	27
4.1 Künstliche neuronale Netze	27
4.1.1 Lernalgorithmus	29
4.1.2 Ausgabeschicht und Softmatrix	33
4.2 Methodik des <i>Topic-Modelings</i>	34
4.2.1 Bestimmung des semantischen Raums	35
4.2.2 <i>Uniform Manifold Approximation and Projection</i>	37

4.2.3	<i>Hierarchical Density-Based Spatial Clustering of Applications with Noise</i>	41
4.2.4	Bestimmung der Themen	44
4.2.5	Silhouettenkoeffizient	44
4.3	Methodik der Sentiment-Analyse	45
4.3.1	Convolutional Neural Networks	45
4.3.2	<i>Long Short-term Memory</i>	49
4.3.3	Regularisierungsarchitektur	53
4.4	Methodik der sozialen Netzwerkanalyse	55
4.5	Daten	57
4.5.1	Datenerhebung	58
4.5.2	Datenaufbereitung	59
5	Empirische Analyse	62
5.1	Untersuchung der Modelle	62
5.1.1	Untersuchung der <i>Topic Modeling</i> -Modelle und Bestimmung der Themen	63
5.1.2	Untersuchung verschiedener künstlicher neuronaler Netze zur Sentiment-Analyse	67
5.2	Regierungskommunikation	74
5.2.1	Kommunikation der Regierung	75
5.2.2	Integration der Regierung im Netzwerk	78
5.2.3	Zusammenfassung der Untersuchung der Regierungskommunikation	88
6	Diskussion	89
7	Fazit	93
A	Anhang	96
A.1	Visuelle Ergänzung zur Durchführung einer 1D Faltung (eigene Darstellung)	96
A.2	Visuelle Darstellung des Hyperparametertraining	96
A.3	Visuelle Darstellung der durch die Abstände im semantischen Raum gewonnenen Themen	98
A.4	Kommunikationsnetzwerk des BMGs, des BMFs und des Accounts von Olaf Scholz im Thema 38 (eigene Darstellung)	99
A.5	Untersuchte Regierungsaccounts	101

Literaturverzeichnis	102
Eidesstattliche Erklärung	113

Symbolverzeichnis

Symbole	Bedeutung
a_i^s	Neuron einer Schicht \mathbf{a}
\mathbf{a}	Ausgabevektor einer Schicht vor Verwendung der Aktivierungsfunktion
A	Cluster A
\mathbf{b}^c	Vektor der Verzerrung im <i>Input-Gate</i> zur Eingabe der Daten
\mathbf{b}^f	Vektor der Verzerrung im <i>Forget-Gate</i> -Ventil
\mathbf{b}^o	Vektor der Verzerrung im <i>Input-Gate</i> -Ventil
\mathbf{b}^q	Vektor der Verzerrung im <i>Output-Gate</i> -Ventil
\mathbf{b}^s	Vektor der Verzerrung in der Schicht s
$\mathbf{b}^{(1)}$	Vektor der Verzerrung in der ersten verdeckten Schicht
B	Cluster B
C_B	<i>Betweenness-Centrality</i>
C_D	<i>Degree-Centrality</i>
C_E	<i>Eigenvector-Centrality</i>
\mathbf{c}	Vektordarstellung des Kontextwortes
\mathbf{c}'	Vektordarstellung eines nicht beobachteten Kontextwortes
\mathbf{C}^t	Zellspeicher zum Zeitpunkt t einer Sequenz
$\tilde{\mathbf{C}}^t$	Eingabe in den Zellspeicher zum Zeitpunkt t einer Sequenz
$d(\mathbf{v}_i, \mathbf{v}_j)$	Distanz zwischen zwei Knoten
$dmreach$	<i>Mutual Reachability Distance</i>
D_{v_i}	Dichte am Knoten v_i
e	Ein Simplizialkomplex
\mathbf{E}	Menge der Simplizialkomplexe

Symbole	Bedeutung
f_i^t	Anteil der Weitergabe von Informationen von C_i^{t-1} nach C_i^t im <i>Forget-Gate</i> auf Basis von x_{ij}^t und h_j^{t-1}
FN	Anzahl an falsch-negativen Schätzungen
FP	Anzahl an falsch-positiven Schätzungen
\mathbf{F}	Matrix der vektorisierten Filter
g	Aktivierungsfunktion
$g^{(1)}$	Aktivierungsfunktion in der ersten verdeckten Schicht
h_i^s	Ausgabe eines Neurons i einer verdeckten Schicht s
\mathbf{h}^s	Ausgabevektor einer verdeckten Schicht s
$\mathbf{h}^{(1)}$	Ausgabevektor der ersten verdeckten Schicht
J	Anzahl der Klassen in den Daten
k	Anzahl der nächsten Punkte, die im Raum betrachtet werden
\mathbf{K}	Filter einer Schicht
l	Anzahl an Knoten eines Netzwerkes
$L(\hat{\mathbf{y}}, \mathbf{y})$	Verlustfunktion
M	Anzahl an Beobachtungen im Datensatz
n	Anzahl der Dimension eines Dokumentenvektors
n_{re}	Anzahl der reduzierten Dimension eines Dokumentenvektors
o_i^t	Anteil der Eingabe von Informationen in C_i^t im <i>Input-Gate</i>
PWI	<i>Probability-weighted amount of Information</i>
q_i^t	Anteil der Weitergabe von Informationen in C_i^t im <i>Output-Gate</i>
r	Anzahl an Neuronen einer Schicht
RN	Anzahl an richtig-negativen Schätzungen
RP	Anzahl an richtig-positiven Schätzungen

Symbole	Bedeutung
s	Schicht in einem künstlichen neuronalen Netz
$s(i)$	Silhouette einer Beobachtung i
s_C	Silhouettenkoeffizient
S_F^s	Ausgabe der <i>Convolution</i> -Schicht
\tanh	<i>Tangens hyperbolicus</i>
\tilde{T}	Menge aller Themenvektoren
\mathbf{T}	Wörter zu Kontextwörter-Kombinationen
u_{ij}^C	Gewicht zwischen Neuron i und j im <i>Input-Gate</i> zur Eingabe der Daten
u_{ij}^f	Gewicht zwischen Neuron i und j im <i>Forget-Gate</i> -Ventil
u_{ij}^o	Gewicht zwischen Neuron i und j im <i>Input-Gate</i> -Ventil
u_{ij}^g	Gewicht zwischen Neuron i und j im <i>Output-Gate</i> -Ventil
\mathbf{U}^t	Gewichtsmatrix zwischen den LSTM-Zellen zum Zeitpunkt t
\mathbf{U}^C	Gewichtsmatrix im <i>Input-Gate</i> zur Eingabe der Daten
\mathbf{U}^f	Gewichtsmatrix im <i>Forget-Gate</i> -Ventil
\mathbf{U}^o	Gewichtsmatrix im <i>Input-Gate</i> -Ventil
\mathbf{U}^q	Gewichtsmatrix im <i>Output-Gate</i> -Ventil
v_{i1}	x-Koordinate eines Knotens im euklidischen Raum
v_{i2}	y-Koordinate eines Knotens im euklidischen Raum
\mathbf{v}	Knoten im euklidischen Raum
\mathbf{V}	Menge der Knoten im Netzwerk
w_{ij}^C	Gewicht zwischen Neuron i und j im <i>Input-Gate</i> für Daten aus h^{t-1}
w_{ij}^f	Gewicht zwischen Neuron i und j im <i>Forget-Gate</i> -Ventil für Daten aus h^{t-1}

Symbole	Bedeutung
w_{ij}^o	Gewicht zwischen Neuron i und j im <i>Input-Gate</i> -Ventil für Daten aus h^{t-1}
w_{ij}^q	Gewicht zwischen Neuron i und j im <i>Output-Gate</i> -Ventil für Daten aus h^{t-1}
w_{ij}^s	Gewicht von h_i^s zu h_j^s
$w_h(e)$	Gewicht jedes Simplizialkomplexes der höheren Dimension
$w_l(e)$	Gewicht jedes Simplizialkomplexes der niedrigeren Dimension
\mathbf{W}^t	Gewichtsmatrix zwischen den LSTM-Zellen zum Zeitpunkt t
\mathbf{W}^c	Gewichtsmatrix für Daten aus h^{t-1} im <i>Input-Gate</i>
\mathbf{W}^f	Gewichtsmatrix für Daten aus h^{t-1} im <i>Forget-Gate</i> -Ventil
\mathbf{W}^o	Gewichtsmatrix für Daten aus h^{t-1} im <i>Input-Gate</i> -Ventil
\mathbf{W}^q	Gewichtsmatrix für Daten aus h^{t-1} im <i>Output-Gate</i> -Ventil
\mathbf{W}^s	Gewichtsmatrix in der Schicht s
$\mathbf{W}^{(1)}$	Gewichtsmatrix der ersten verdeckten Schicht
x_j^t	Neuron der Eingabe zum Zeitpunkt t der Sequenz
x_i	Neuron der Eingabe
\mathbf{x}	Eingabevektor einer Beobachtung
\mathbf{X}	Menge der Eingabedaten
\mathbf{X}_{re}	Menge der dimensionsreduzierten Eingabedaten
\mathbf{y}	Wahre Werte der Beobachtungen
$\hat{\mathbf{y}}$	Ausgabevektor eines künstlichen neuronalen Netzes und Schätzung des wahren Wertes
z_i	Neuron der letzten verdeckten Schicht

Symbole	Bedeutung
z	Letzte verdeckte Schicht
α_{ij}	Ausprägung innerhalb der Adjazenzmatrix ist eins wenn eine Verbindung besteht
$\alpha(\mathbf{v}_i, \mathbf{v}_j)$	Zeigt an, ob eine Verbindung zwischen zwei Knoten besteht
$\tilde{\alpha}_k(\mathbf{v}_i, \mathbf{v}_j)$	Zeigt an, ob eine Verbindung zwischen zwei Knoten die kürzest mögliche ist
α	Adjazenzmatrix
δ	Fehlerterm
ϵ	Epoche
η	Lernrate
$\kappa(\mathbf{x}_{re_i})$	Abstand von i zu dem k-sten Nachbarn
$\kappa(\mathbf{x}_{re_j})$	Abstand von j zu dem k-sten Nachbarn
λ	Lokale Dichte an einem Punkt
$\Phi(\mathbf{H}^s)$	Transponierte Matrix der durch Filter unterteilten Eingabe in die Faltung
σ	Logistische Sigmoidfunktion
$\tilde{\tau}$	Anzahl der Wörter in einem Satz
τ	Länge der Sequenz einer Eingabe
$\tilde{\omega}$	Vektordarstellung eines Worts im Textkorpus
ω	<i>One-hot</i> -Vektordarstellung eines Wortes
Ω	Menge aller Wortvektoren

Abkürzungsverzeichnis

Abkürzung	Bedeutung
BERT	<i>Bidirectional Encoder Representations from Transformers</i>
BMF	Bundesministerium der Finanzen
BMG	Bundesministerium für Gesundheit
BMBF	Bundesministerium für Bildung und Forschung
BMFSFJ	Bundesministerium für Familie, Senioren, Frauen und Jugend
BMI	Bundesministerium für Inneres
BMjV	Bundesministerium der Justiz und für Verbraucherschutz
BMVg	Bundesministerium der Verteidigung
BMWi	Bundesministerium für Wirtschaft und Energie
BMWL	Bundesministerium für Ernährung und Landwirtschaft
BIP	Bruttoinlandsprodukt
CDC	<i>Centers for Disease</i>
CNN	<i>Convolutional Neural Networks</i>
COVID-19	<i>Coronavirus Disease 2019</i>
HDBSCAN	<i>Hierarchical Density-Based Spatial Clustering of Applications with Noise</i>
IDF	<i>Inverse Document Frequency</i>
KfW	<i>Kreditanstalt für Wiederaufbau</i>
LDA	<i>Latent Dirichlet Allocation</i>
LSTM	<i>Long Short-term Memory</i>
NB	<i>Naive Bayes Classifier</i>
NLP	<i>Natural-Language-Processing</i>
UMAP	<i>Uniform Manifold Approximation and Projection</i>
PCA	<i>Principal Component Analysis</i>
PLSA	<i>Probabilistic latent semantic Analysis</i>

Abkürzung	Bedeutung
PWI	<i>Probability-weighted amount of Information</i>
PV-DBOW	<i>Bag of Words version of Paragraph Vector</i>
ReLU	<i>Rectified Linear Unit</i>
RNN	<i>Recurrent Neural Network</i>
SARS-CoV-2	<i>severe acute respiratory syndrome coronavirus type 2</i>
SVM	<i>Support Vector Machine</i>
TF	<i>Phasen Frequency of a Word</i>
t-SNE	<i>t-distributed stochastic neighbor embedding</i>
WHO	Weltgesundheitsorganisation

Abbildungsverzeichnis

1	Schaubild eines neuronalen Netzes (eigene Darstellung)	29
2	Anpassung der Distanz abhängig von der Dichte (eigene Darstellung)	40
3	Anpassung der Distanz zwischen Datenpunkten durch dmreach() (eigene Darstellung in Anlehnung an McInnes et al. 2016)	43
4	Schaubild einer LSTM Zelle (eigene Darstellung in Anlehnung an Olah 2015)	51
5	Zwei dimensionale Darstellung der auf fünf Dimensionen reduzierten Tweets (links) und zwei dimensionale Darstellung mit Clustern (rechts) (eigene Darstellung)	64
6	Die fünf Themen mit dem geringsten Abstand im semantischen Raum zum Wort "wirtschaft"	65
7	Darstellung der Themen zur Demonstrationen gegen Einschränkun- gen der Grundrechten	67
8	Hyperparametertraining CNN (eigene Darstellung)	70
9	Vergleich der Trainings- und Validierungs-Accuracy der künstlichen neuronalen Netze zum Schätzen des Sentiments (eigene Darstellung) .	73
10	Verteilung der Regierungstweets (eigene Darstellung)	78
11	Verteilung der Regierungstweets auf die extrahierten Themen (eigene Darstellung)	79
12	Kommunikationsnetzwerk des Themas 13 nach Eigenvector-Centrality und Degree-Centrality (eigene Darstellung)	80
13	Kommunikationsnetzwerk von BMF und BMWI innerhalb des Themas 13 (eigene Darstellung)	82
14	Kommunikationsnetzwerk des Themas 38 nach Eigenvector-Centrality (eigene Darstellung)	84
15	Kommunikationsnetzwerk des Themas 28 nach Eigenvector-Centrality (eigene Darstellung)	87
A1	Hyperparametertraining LSTM (eigene Darstellung)	96
A2	Hyperparametertraining CNN-LSTM (eigene Darstellung)	97
A3	Hyperparametertraining LSTM-CNN (eigene Darstellung)	97
A4	Die fünf Themen mit den geringsten Abstand im semantischen Raum zum Wort "Schliessungen" (eigene Darstellung)	98
A5	Die fünf Themen mit den geringsten Abstand im semantischen Raum zum Wort "hilfen" (eigene Darstellung)	98

Tabellenverzeichnis

1	Test-Accuracy der Modell zur Sentiment-Schätzung	73
2	Die Zehn Themen mit der höchsten Anzahl an Regierungstweets	75
3	Die Zehn Regierungsaccounts mit den meisten Tweets	76
4	Die Zehn Regierungsaccounts mit den meisten Tweets zu einem Thema	77
5	Anzahl an Regierungstweets in den ausgewählten Themen	78
6	Die Zehn Accounts im Kommunikationsnetzwerk des Thema 13 nach Metrik	81
7	Die Zehn Accounts im Kommunikationsnetzwerk des Thema 38 nach Metrik	85
8	Die Zehn Accounts im Kommunikationsnetzwerk des Thema 28 nach Metrik	88
A1	Auflistung der untersuchten Regierungsaccounts	101

Abstract

This thesis analyses German government communication on Twitter during the COVID-19 pandemic. Specifically, the influence of the government on economic policy topics is studied. By applying techniques of deep learning and concepts of natural language processing, the topics were extracted from the COVID-19 Twitter dataset collected from January 2020 to August 2020. The communication structures within the topics were investigated through social network analysis and the influence of the government in the network was identified. As a result, government communication was found to be widely spread across the themes. The examination of the economic policy communication networks showed a low influence on government communication. The analysis showed gaps in government communication especially in topics where restrictions on economic freedom of action were discussed.

Zusammenfassung

Die vorliegende Arbeit analysiert die deutsche Regierungskommunikation auf Twitter während der COVID-19-Pandemie. Im Speziellen wird der Einfluss der Regierung auf wirtschaftspolitische Themen untersucht. Durch die Anwendung von Techniken des *Deep-Learning* und Konzepten des *Natural-Language-Processing* werden die Themen aus dem von Januar 2020 bis August 2020 erhobenen COVID-19-Twitter-Datensatz extrahiert. Die Kommunikationsstrukturen innerhalb der Themen werden durch soziale Netzwerkanalysen untersucht und der Einfluss der Regierung im Netzwerk bestimmt. Im Ergebnis wird eine über die Themen weit gefächerte Kommunikation der Regierung festgestellt. Die Untersuchung der wirtschaftspolitischen Kommunikationsnetzwerke offenbart einen geringen Einfluss durch die Regierungskommunikation. Die Analyse zeigt insbesondere in Themen, in denen die Einschränkungen der wirtschaftlichen Handlungsfreiheit diskutiert wurden, Lücken in der Regierungskommunikation auf.

1 Einleitung

Am 31.12.2019 wurde durch die chinesische Regierung ein Ausbruch einer neuartigen Infektionskrankheit mit zu diesem Zeitpunkt unbekannter Ursache gemeldet (vgl. Taylor, 2021). Die Krankheit erhielt den Namen *Coronavirus Disease 2019 (COVID-19)* und der Virus selbst wird als *severe acute respiratory syndrome coronavirus type 2* oder auch als SARS-CoV-2 bezeichnet (vgl. Weltgesundheitsorganisation, 2020a). Von China aus verbreitete sich COVID-19 in der ganzen Welt und wurde von der Weltgesundheitsorganisation (WHO) am 12.03.2020 zu einer weltweiten Pandemie erklärt (vgl. Weltgesundheitsorganisation, 2020b). Innerhalb Deutschlands wurde ab dem 22.03.2020 bundesweit mit Maßnahmen zur Eindämmung der Ausbreitung von COVID-19 begonnen (vgl. Bundespressamt, 2020a). In der Folge der Pandemie und der damit verbundenen Maßnahmen kam es zu einem Nachfragerückgang und zu einer Unterbrechung von Lieferketten. Die Umsätze im Gastgewerbe und in der Luftfahrt sanken um 68 % und 76 % (vgl. Feld et al., 2020, S. 46). Im zweiten Quartal 2020 sank das Bruttoinlandsprodukt (BIP) um 9,7 % im Vergleich zum Vorquartal (vgl. Feld et al., 2020, S. 42). Um die Folgen der COVID-19-Pandemie und die damit verbundenen Kosten durch die Geschäftsschließungen abzumildern, beschloss die Bundesregierung eine Reihe von unterstützenden Maßnahmen. Zu Beginn der Pandemie wurde der Bezug von Kurzarbeitergeld erleichtert und eine Unterstützung für kleine Betriebe und Selbstständige in einer Höhe von 50 Milliarden Euro beschlossen. Ein Viertel davon wurde als Soforthilfe ausgezahlt, bei der eine Rückerstattung durch die Empfängern¹ erforderlich ist (vgl. Bundeswirtschaftsministerium, 2020). Zur Unterstützung von in Not geratenen Unternehmen wurden darüber hinaus sowohl Fremdkapital- als auch Eigenkapitalfinanzierungen durchgeführt.

Neben den akuten Auswirkungen der Pandemie auf die Wirtschaft entstehen zudem langfristige Folgen durch die Unsicherheit über den weiteren Verlauf der Pandemie und die Einschränkungen der Geschäftstätigkeit auf Basis politischer Maßnahmen. Die Maßnahmen zur Eindämmung der Pandemie sowie diejenigen zur Unterstützung der Wirtschaft werden durch die Regierung kommuniziert. Besteht durch ungenaue und/oder unzureichende Kommunikation Unsicherheit über das weitere Verhalten der Regierung, dann werden beispielsweise Investitionen zurückgehalten. Um in der Bevölkerung Verständnis für die politisch gesetzten Einschränkungen erreichen zu können und so die Unsicherheit zu reduzieren, äußert sich die Bun-

¹Aus Gründen der besseren Lesbarkeit wird im Folgenden auf die gleichzeitige Verwendung weiblicher und männlicher Sprachformen verzichtet und das generische Maskulinum verwendet. Sämtliche Personenbezeichnungen gelten gleichermaßen für beide Geschlechter.

desregierung über eine Reihe von Kommunikationskanälen. Einer dieser Kanäle ist die *Social-Media*-Plattform Twitter. Der Vorteil ihrer Nutzung in der Kommunikation der Bundesregierung liegt in der Kontrolle über die veröffentlichten Inhalte bei gleichzeitigem direktem Verfügbarmachen von Informationen für die Öffentlichkeit.

Durch die Untersuchung von Kommunikation während Krisen können Rückschlüsse auf die Bedürfnisse und Interessen der Bevölkerung gezogen werden. Wird in einer Krise darüber hinaus die Kommunikation der Regierung analysiert, so können Zusammenhänge extrahiert werden, die zu einer Optimierung der Interaktion mit der Bevölkerung während einer Krise beitragen. Aufgrund des stetigen Anstiegs der Bedeutung von *Social-Media*-Plattformen wurde die Kommunikation auf Twitter während verschiedener Krisen untersucht. Neben der Analyse von Krisen durch Naturkatastrophen wie Hurrikans (vgl. Pourebrahim et al., 2019) oder Tsunamis (vgl. Lu & Brelsford, 2014) beschäftigte sich die Forschung mit der staatlichen Kommunikation während Krankheitsausbrüchen, wie etwa von der Grippe (vgl. Yun et al., 2016a) oder Ebola (vgl. Crook et al., 2016). Bisherige wissenschaftliche Arbeiten, welche die Kommunikation der Regierung während der COVID-19-Pandemie untersuchten, verwendeten manuelle Klassifizierungen zur Analyse der von der Regierung über Twitter veröffentlichten Informationen und untersuchten nicht den Einfluss der Regierung in dem Kommunikationsnetzwerk (vgl. Rufai & Bunce, 2020).

In dieser Arbeit wird die Kommunikation der deutschen Bundesregierung auf Twitter während der COVID-19-Pandemie untersucht. Im Vergleich zu bisherigen Untersuchungen der Regierungskommunikation werden die Themen nicht manuell bestimmt, sondern mit Techniken des *Deep-Learnings* aus den vorhandenen Daten extrahiert (vgl. Rufai & Bunce, 2020; Wang et al., 2021). Durch die Wahl dieser Methodik werden Verzerrungen durch den Auswahlprozess der Daten verringert. Darüber hinaus ist bisher keine Arbeit bekannt, die sich mit der Krisenkommunikation im deutschsprachigen Twitter während der COVID-19-Pandemie befasst. Durch die verwendete Methodik und aufgrund der Menge an Datenpunkten ist die Analyse der Regierungskommunikation innerhalb bestimmter Themen möglich.

Der Schwerpunkt in der Untersuchung liegt auf der Kommunikation der Regierung zu wirtschaftlichen Themen. Hier soll zum einen analysiert werden, inwieweit die Bundesregierung diesen Kommunikationskanal nutzt, um ihre wirtschaftspolitischen Maßnahmen erklären zu können und Unsicherheiten zu senken. So führt beispielsweise ein geringer Einfluss der Bundesregierung im Kommunikationsnetzwerk zu keiner Verringerung der Unsicherheit, unabhängig davon, in welchem Maße zu wirtschaftlichen

Themen, wie der Soforthilfe, kommuniziert wird. Durch das Untersuchen der Kommunikationsnetzwerke wird des Weiteren festgestellt, welche Accounts innerhalb des Netzwerkes einen Einfluss ausüben. Dies soll Hinweise für eine gezielte Kooperation zu einer erhöhten Verbreitung der Informationen über Twitter liefert.

Hieraus leiten sich zwei konkrete Fragestellungen ab:

1. Zu welchem Themenbereich kommuniziert die Regierung während der COVID-19-Pandemie?
2. Welchen Einfluss hat die Regierung in Kommunikationsnetzwerken zu wirtschaftspolitischen Maßnahmen?

Da in dieser Arbeit nicht ausschließlich die Regierungskommunikation betrachtet wird, sondern ebenso die Regierungskommunikation innerhalb des Kommunikationsnetzwerkes, werden sämtliche deutschsprachigen Tweets vom 01.01.2020 bis zum 31.08.2020 erhoben und in die Analyse einbezogen. In diesem Zeitraum ist die erste Welle der Pandemie und das Abflachen dieser mit eingeschlossen. Der Beginn der zweiten Welle ist damit in der Datenbasis nicht enthalten.

Innerhalb dieses Zeitraumes wurden in der öffentlichen Diskussion verschiedene Themen im Zusammenhang mit der COVID-19-Pandemie besprochen. Im ersten Schritt der Analyse werden die Themen auf Twitter während der COVID-19-Pandemie erhoben und aus diesen die wirtschaftspolitischen Themen extrahiert. Im zweiten Schritt wurden verschiedene Modelle zur Bestimmung der Reaktionen innerhalb des Netzwerks untersucht. Die untersuchten Modelle wiesen eine geringe Genauigkeit der Schätzung auf und wurden entsprechend verworfen. Im dritten Schritt wird eine interpretierbare Darstellungsform für die Kommunikationsnetzwerke bestimmt.

Um diese drei Schritte durchführen zu können, wird zu Beginn der Arbeit die Plattform Twitter erläutert (vgl. Kapitel 2.1) und wissenschaftliche Untersuchungen zur Krisenkommunikation auf Twitter vorgestellt (vgl. Kapitel 2.2). Anschließend werden die theoretischen Grundlagen für die Datenbearbeitungsschritte und der Umgang mit Textdaten erläutert (vgl. Kapitel 3). Auf diesem theoretischen Wissen aufbauend werden die Methodiken der angewandten Modelle und die Datenerhebung und -aufbereitung beschrieben (vgl. Kapitel 4.1). Im nächsten Schritt werden die Themen während der COVID-19-Pandemie bestimmt und die wirtschaftspolitischen Themen extrahiert (vgl. Kapitel 5.1.1). In Bezug auf diese Themen kann dann die Regierungskommunikation untersucht werden (vgl. Kapitel 5.2).

2 Twitter und Krisenkommunikation

Im folgenden Kapitel werden das der Untersuchung zugrundeliegende Netzwerk Twitter und die möglichen Kommunikationsformen innerhalb des Netzwerkes beschrieben. Ebenfalls wird auf die Problematik der *Bots*² innerhalb des Netzwerkes eingegangen und die aktuellen Erkenntnisse zum Einfluss von *Bots* während der COVID-19-Pandemie werden präsentiert. Zudem wird vorgestellt wie in dieser Arbeit mit *Bots* umgegangen wird. Anschließend werden verschiedene wissenschaftliche Untersuchungen zur Kommunikation während Krisensituationen auf Twitter thematisiert. Insbesondere wird die wissenschaftliche Literatur zur Kommunikation auf Twitter während Krankheitsausbrüchen erläutert.

2.1 Twitter

Twitter ist eine im Jahr 2006 gegründete Microblogging Plattform. Im Vergleich zu anderen *Social Media*-Plattformen entsteht durch die *Blogging*-Struktur ein öffentlicher Raum. Aus der öffentlichen Orientierung folgt eine Fokussierung der Inhalte auf Schlagzeilen oder Nachrichten (vgl. Kwak et al., 2010, S. 600). Veröffentlichte Nachrichten auf Twitter werden als Tweet bezeichnet. Durch die Begrenzung der Anzahl der Zeichen auf 280 werden in Tweets Informationen kurz und knapp zusammengefasst (vgl. Twitter, 2017). Dies führt zum Weglassen von Abwägungen und Einschränkungen in der Kommunikation und dadurch werden die Tweets eindeutiger. Die Nutzer von Twitter haben die Möglichkeit, sich gegenseitig zu folgen. In der eigenen *Timeline* sind die Tweets der abonnierten Accounts zu sehen. Jeder öffentliche Tweet kann von einem anderen Nutzer wiederum unter den eigenen *Followern* geteilt werden. Dieser Vorgang wird retweeten genannt. Jeder Tweet oder Retweet kann durch eine Antwort kommentiert werden. Sowohl Retweets als auch Antworten der Nutzer sind auf der Profilseite des Nutzers einsehbar. Antwortet ein Nutzer auf einen Tweet, erscheinen die Antwort und der Tweet in den *Timelines* seiner *Follower* (vgl. Twitter, 2020a).

Twitter stellt eine Schnittstelle zum Abfragen von Daten bereit. Über diese Schnittstelle können umfangreiche Daten über Tweets und Nutzer erhalten werden. Neben dem Nutzernamen und dem Datum des Tweets können beispielsweise die Standortdaten der Nutzer gewonnen werden. Die Weitergabe der Standortdaten kann

²In dieser Arbeit wird der Begriff *Bots* für *Social Bots* eingesetzt. *Social Bots* im Twitter Kontext sind Twitter-Accounts die automatisiert beispielsweise Tweets erstellen oder Retweets durchführen (vgl. Ferrara, 2020, S. 4).

jedoch durch den Nutzer unterbunden werden. Die umfangreich verfügbaren Daten und der einfache Zugang zu den Daten haben dazu geführt, dass sich verschiedene Fachrichtungen mit Twitter beschäftigen. Durch den Schwerpunkt der Tweets auf Schlagzeilen und Nachrichten bietet sich Twitter für wirtschaftswissenschaftliche oder politische Analysen an, da ein Zusammenhang zwischen der Reaktion auf Twitter und aktuellen Nachrichten beziehungsweise der Reaktion auf die Nachrichten vermutet wird. Untersuchungen mit wirtschaftswissenschaftlichem Schwerpunkt zeigten einen Zusammenhang zwischen der Anzahl an Tweets oder des Sentiments³ der Tweets mit Aktienkursen (vgl. Bollen et al., 2011; Ruan et al., 2018; Zhang et al., 2011; Zheludev et al., 2014).

In Untersuchungen mit politischem Schwerpunkt wurde beispielsweise der Zusammenhang zwischen Meinungsumfragen und der Twitterkommunikation auf Basis des Sentiments untersucht (vgl. Bovet et al., 2018). Vor allem durch die US-Wahl im Jahr 2016 rückte die Problematik von *Fake News* und *Bots* in den Fokus der Öffentlichkeit und wurde deshalb einer wissenschaftlichen Untersuchung unterzogen. Bovet & Makse (2019) fanden deutliche Unterschiede in der Verteilung von *Fake News* zwischen Anhängern von Hillary Clinton und Unterstützern von Donald Trump. Bei der Verbreitung von *Fake News* wird *Bots* ein hoher Stellenwert zugeordnet. Caldarelli et al. (2020) stellten eine wichtige Rolle von *Bots* bei der Verbreitung von Informationen fest, und zwar unabhängig von *Fake News* oder normalen Tweets. Der Einfluss von durch *Bots* verbreiteten *Fake News* zur Wahlmanipulation bei der *Mid-Term-Wahl* in den USA im Jahr 2018 wurde von Deb et al. (2019) untersucht. Es wurde festgestellt, dass die Verzerrungen durch *Bots* zu einer Verzerrung von Analysen führen können, wenn einzelne Themen betrachtet und mit Daten außerhalb von Twitter verglichen werden.

Erste Untersuchungen zu *Bots* auf Twitter im Kontext der COVID-19-Pandemie liefern Hinweise darauf, dass vor allem politische Propaganda und Verschwörungstheorien durch *Bots* geteilt werden (vgl. Ferrara, 2020, S. 13). Accounts, die nicht als *Bots* identifiziert wurden, legten einen stärkeren Fokus auf die Diskussion von Themen der öffentlichen Gesundheit (vgl. Ferrara, 2020, S. 13). Die Qualität der geteilten Informationen bis Anfang April 2020 war mehrheitlich hoch oder mittel (vgl. Singh, Bode, et al., 2020; S. 362, Singh, Bansal, et al., 2020, S. 20). Quellen hoher Qualität sind Organisationen wie das *Centers for Disease* (CDC) oder die WHO. Als mittlere Qualität werden bekannte Nachrichtenquellen eingeordnet. Das Netzwerk separiert

³Das Sentiment oder die Stimmung eines Textes wird in Analysen in positiv und negativ oder auch in positiv, negativ und neutral unterteilt.

sich anhand der Qualität der Informationen (vgl. Singh, Bode, et al., 2020, S. 362). In einem Teil des Netzwerkes wurden Informationen mit niedriger Qualität untereinander geteilt, im zweiten solche mit hoher Qualität. Diese beiden Teile zeigten geringe Interaktionen untereinander. Der dritte Teil des Netzwerkes waren Accounts, die Informationen mittlerer Qualität teilten. Dieser Teil weist hohe Verbindungen zu den anderen zwei Teilen auf (vgl. Singh, Bode, et al., 2020, S. 362). In der ersten Phase der COVID-19-Pandemie wurden also auch in Netzwerkteilen, die bevorzugt Informationen mit niedriger Qualität teilen, Informationen mit mittlerer Qualität geteilt.

Die vorgestellten Arbeiten zeigen, dass beim Untersuchen von Twitter-Netzwerken die Möglichkeit von Verzerrungen durch *Bots* berücksichtigt werden muss. Gleichzeitig können die von *Bots* geteilten Informationen Auswirkungen auf andere Accounts im Netzwerk haben. Entsprechend führt das Löschen von *Bots* zum Löschen von Informationen über das Verhalten der anderen Accounts und zu Verzerrungen in der Analyse. *Bots* werden in dieser Arbeit entsprechend nicht aus dem Datensatz gelöscht. Eine mögliche Verzerrung durch *Bots* wird in dieser Arbeit durch zwei Schritte reduziert. Zum einen werden innerhalb der sozialen Netzwerkanalyse sowohl die einflussreichsten Accounts als auch die Accounts an Verbindungsstellen im Netzwerk analysiert. Zum anderen werden die als einflussreich erkannten Accounts manuell überprüft. Der Fokus dieser Arbeit liegt dabei nicht auf der Betrachtung des Einflusses von *Bots* im Netzwerk, sondern auf der Analyse der Regierungskommunikation. Offizielle Accounts werden von Twitter als solche gekennzeichnet.

2.2 Kommunikation und staatliche Kommunikation auf Twitter während Krisen

Die COVID-19-Pandemie stellt eine im *Social Media*-Zeitalter nie dagewesene Krise dar. Durch den globalen Charakter und den langen Zeitraum ist die COVID-19-Pandemie in ihrem Umfang nicht uneingeschränkt mit anderen Krisen vergleichbar. Anhand der Berücksichtigung von wissenschaftlichen Arbeiten zur Krisenkommunikation auf Twitter können dennoch im Verlauf der Arbeit Unterschiede und Gemeinsamkeiten herausgearbeitet und auf bisherige Erkenntnisse zurückgegriffen werden.

2.2.1 Kommunikation auf Twitter in Krisenzeiten

Die Kommunikation auf Twitter wurde bereits während verschiedener Krisen untersucht. Eine Betrachtung der Kommunikation während anderer Krisensituationen kann

für eine Einordnung des Netzwerks während der COVID-19-Pandemie behilflich sein. Pourebrahim et al. (2019) untersuchten die Kommunikation auf Twitter, während der Hurrikan Sandy im Jahr 2012 durch die USA zog. Für die Analyse wurden 13,7 Millionen Tweets von Twitter heruntergeladen und in drei Phasen unterteilt, jeweils ein Datensatz für den Zeitraum vor und nach dem Hurrikan und ein Datensatz für die Periode seines Auftretens (vgl. Pourebrahim et al., 2019, S. 6). Zur Verarbeitung der Tweets wurde TF-IDF⁴ genutzt und das Sentiment der Tweets wurde durch die Anwendung von *Support Vector Machine (SVM)* bestimmt und erwies sich in der Prä- und Post-Phase des Hurrikans hauptsächlich als positiv. In der Zeit des Hurrikans veränderte sich das Sentiment zum Negativen, wobei das höchste Verhältnis von negativen zu positiven Tweets beim Höhepunkt des Hurrikans erreicht wurde (vgl. Pourebrahim et al., 2019, S. 13). Die drei verschiedenen Zeiträume wurden ebenfalls mit Techniken der sozialen Netzwerkanalyse untersucht. Als Verbindungen zwischen den Knoten wurden die Antworten auf andere Nutzer verwendet. Zu allen drei Zeiträumen stellten Pourebrahim et al. (2019) die höchste *Degree-Centrality* für Accounts wie die *Federal Emergency Management Agency* fest.⁵ Die höchste *Eigenvector-Centrality* wurde für Personen des öffentlichen Lebens wie beispielsweise *Michael Bloomberg*⁶ ermittelt und nahm somit eine zentrale Rolle in der Verteilung der Informationen im Netzwerk ein (vgl. Pourebrahim et al., 2019, S. 14). Wichtige Knoten zur Verteilung der Informationen im Netzwerk stellten Wetteragenturen und katastrophenbezogene Behörden dar, was sich durch eine hohe *Betweenness-Centrality* äußerte (vgl. Pourebrahim et al., 2019, S. 14).

Pourebrahim et al. (2019) zeigten, dass mit Techniken der sozialen Netzwerkanalyse Informationen zur Krisenkommunikation auf Twitter gewonnen werden können. Darüber hinaus wurde die Bedeutung von Accounts einzelner Personen im Netzwerk deutlich. Ein möglicher Erklärungsansatz ist, dass diese Accounts ein weiteres Themenfeld als staatliche Accounts mit einem spezifischen Themenspektrum abdecken und deshalb in Krisenzeiten mehr Nutzer erreichen.

Lu & Brelsford (2014) untersuchten die Kommunikation auf Twitter zum Erdbeben und dem darauf folgenden Tsunami im Jahr 2011 in Japan. Wie bei Pourebrahim et al. (2019) wurde die Kommunikation in Zeiträume unterteilt und die Phasen vor und nach dem Erdbeben analysiert. Es wurden ein japanischer, ein englischer und ein globaler Tweetdatensatz erhoben. Lu & Brelsford (2014) argumentierten, dass 99%

⁴Erläuterung in Kapitel 3.4.

⁵*Eigenvector-Centrality*, *Degree-Centrality* und *Betweenness-Centrality* werden in Kapitel 3.5 erläutert.

⁶New Yorker Bürgermeister von 2002 bis 2013 (vgl. New York, 2020).

aller Menschen, die Japanisch als erste Sprache sprechen, in Japan leben und somit direkt oder indirekt vom Tsunami betroffen waren. Der englischsprachige Datensatz wurde als Vergleichsdatensatz genutzt. Im Vergleich zu den Zeiträumen vor und nach der Krise stieg die Anzahl der beteiligten Nutzer in allen Datensätzen deutlich an. Die Netzwerke wurden in verschiedene *Communities* unterteilt und in beiden Zeiträumen untersucht. Lu & Breisford (2014) stellten fest, dass die Nutzer nur in seltenen Fällen die *Community* verließen und sich dadurch in den beliebtesten *Communities* im englischsprachigen und im globalen Netzwerk durch den Tsunami keine Veränderung in den Themen ergab. Im japanischen Netzwerk änderte sich das Thema über alle *Communities* hinweg zu dem Thema Tsunami (vgl. Lu & Breisford, 2014, S. 10).

Lu & Breisford (2014) lieferten Hinweise darauf, dass regionale Krisen die Kommunikation auch nur in den entsprechenden regionalen Netzwerken stark verändern. Übertragen auf die Analyse der deutschsprachigen Tweets zur COVID-19-Pandemie lässt dies zwei Vermutungen zu: zum einen, dass der Fokus der Kommunikation des Netzwerkes auf den Themen im deutschsprachigen Raum liegt und zum anderen, dass lokale Dynamiken zwar einen Einfluss auf die Kommunikation im Netzwerk haben, sich die grundsätzliche Struktur jedoch nicht verändert. Ereignisse im Ausland werden zwar diskutiert, überlagern allerdings nicht die Diskussion inländischer Ereignisse. Gleichzeitig sind deutschsprachige Tweets aus Österreich und der Schweiz zwar in den Netzwerken enthalten, bilden jedoch eigene Cluster, wenn länderspezifische Ereignisse diskutiert werden.

2.2.2 Staatliche Krisenkommunikation auf Twitter während eines Krankheitssausbruches

Neben den Untersuchungen der Kommunikation in allgemeinen Krisenzeiten auf Twitter beschäftigten sich bereits verschiedene Arbeiten mit der staatlichen Kommunikation während eines Krankheitsausbruches. Die jährliche Grippewelle etwa wird von einer Informationskampagne zu Impfstoffen und Vorsichtsmaßnahmen begleitet. Yun et al. (2016b) untersuchten die Kommunikation über sieben Wochen in der Grippeaison 2013/2014. Die Accounts wurden über die Antworten zu einem Netzwerk verbunden. Jede Woche wurden die 100 einflussreichsten Accounts anhand der *Eigenvector-Centrality* bestimmt und manuell in eine Kategorie eingeteilt (vgl. Yun et al., 2016b, S. 70). Es wurde in individuelle Accounts, Accounts mit einem Gesundheitsbezug, Medienaccounts und Accounts von staatlichen oder halbstaatlichen Organisationen

unterschieden. Die betrachteten 700 Accounts enthielten zehn Accounts von Medien, acht Accounts mit einem Gesundheitsbezug und zehn Accounts von staatlichen oder halbstaatlichen Organisationen (vgl. Yun et al., 2016b, S. 70). Fünf der zehn Mediennaccounts zeigten über den Beobachtungszeitraum hinweg eine stabile *Eigenvector-Centrality*. Über den gesamten Beobachtungszeitraum hatten die Accounts mit einem Gesundheitsbezug den zweithöchsten Einfluss im Netzwerk. Accounts von staatlichen oder halbstaatlichen Organisationen zeigten in vereinzelten Wochen im Vergleich zu den anderen Kategorien einen hohen Wert im Eigenvektor Ranking. Der Einfluss dieser Accounts war jedoch immer nur temporär hoch (vgl. Yun et al., 2016b, S. 71). Aus den Ergebnissen leiteten Yun et al. (2016b) ab, dass bei der Verbreitung von Gesundheitsinformationen etablierte Medien einen großen Einfluss auf die Verteilung innerhalb des Netzwerkes haben.

Während des Ebola Ausbruches im Jahr 2014 wurden insgesamt 22.000 Fälle auf der Welt registriert. Dieser größte Ausbruch in der Geschichte führte nach den ersten Fällen in den USA teilweise zu einer Verängstigung in der Bevölkerung. Crook et al. (2016) untersuchten die Kommunikation des CDC in Livechat Events auf Twitter. Während dieses Livechats konnten die Twitternutzer Fragen an das CDC stellen und das CDC antwortete auf diese. Für die Auswertung wurden 512 Antworten manuell in drei Themen eingeordnet (vgl. Crook et al., 2016, S. 351). 43 % der Tweets beschäftigten sich mit den Maßnahmen zur Eingrenzung des Ebola Ausbruches. 36 % informierten über die Krankheit Ebola und 21 % der Tweets des CDC ordneten falsche oder unklare Informationen zu Ebola neu ein (vgl. Crook et al., 2016, S. 352).

Verschiedene Studien beschäftigten sich bereits mit der Kommunikation von offiziellen Accounts auf Twitter während der COVID-19-Pandemie. Rufai & Bunce (2020) analysierten Tweets der Regierungsvorsitzenden der G7 Staaten, inklusive Charles Michel, dem Präsidenten des Europäischen Rates, und Ursula von der Leyen, der Präsidentin der Europäischen Kommission. Bis auf Angela Merkel nutzen alle Vorsitzenden Twitter (vgl. Rufai & Bunce, 2020, S. 511). Twitteraccounts von Politikern werden teilweise auch durch Teams der Politiker bespielt. Es wurden Tweets bis Ende März 2020 erhoben, welche die Wörter *COVID-19* oder *Coronavirus* enthielten. Unterteilt wurde hierbei in informative, moralisch und politisch motivierte Tweets. Bei fast allen Regierungsvorsitzenden war der überwiegende Teil der Tweets informativ, wobei Emmanuel Macron hingegen in 60 % der Tweets an die moralischen Werte appellierte und ein deutlicher Ausreißer war (vgl. Rufai & Bunce, 2020, S. 514). Die anderen untersuchten Accounts posteten maximal 10 % (Trump) an Tweets

mit moralischen Appellen im Kontext der COVID-19-Pandemie (vgl. Rufai & Bunce, 2020, S. 514). Tweets, die eine politische Agenda mit COVID-19 verbanden, wurden ausschließlich von Donald Trump gepostet. Über alle Tweets zum Thema COVID-19 hinweg waren 27% der Tweets von Donald Trump politisch motiviert (vgl. Rufai & Bunce, 2020, S. 514).

Wang et al. (2021) untersuchten ebenfalls die von staatlichen und halbstaatlichen Accounts geposteten Tweets von Mitte Januar bis Ende April 2020. In die Analyse mit eingeschlossen wurden die Weltgesundheitsorganisation sowie staatliche Ämter in den USA, wie beispielsweise das *National Institute of Health* und das *Florida Department of Health*. Wang et al. (2021) stellten ab dem 23.02.2020 einen starken Anstieg der Tweets pro Tag von den ausgewählten Organisationen fest. Der hauptsächliche Themenschwerpunkt der Tweets waren das Informieren der Bevölkerung über die aktuelle Situation und das Bereitstellen von Informationen (vgl. Wang et al., 2021, S. 5). Ein Fokus der Arbeit war das Untersuchen der Koordination der öffentlichen Accounts untereinander. Analysiert wurde ein Retweet Netzwerk. Mit dem Fortschreiten der Pandemie stieg die Interaktion der staatlichen und halbstaatlichen Accounts miteinander an (vgl. Wang et al., 2021, S. 10). Im Mittelpunkt des Netzwerkes stand das CDC. Wang et al. (2021) stellten ein organisationsübergreifendes Teilen von Informationen zwischen staatlichen Stellen fest. In der Kommunikation wurde so sichergestellt, dass Informationen eine hohe Anzahl an Nutzern erreichte.

Insgesamt ist die Untersuchung der staatlichen Kommunikation auf Twitter bei Krankheitsausbrüchen auf die von dem Account kommunizierten Inhalte fokussiert. Yun et al. (2016b) bilden hier eine Ausnahme. Für eine erfolgreiche Verteilung der Informationen und die Wirkkraft der Maßnahmen während eines Ausbruches ist es jedoch notwendig, dass die genutzten Accounts auch einen Einfluss im Netzwerk haben. Im Vergleich zur aktuellen Forschung im Bereich der staatlichen Kommunikation auf Twitter während der COVID-19-Pandemie wird in dieser Arbeit eine größere Anzahl an Tweets analysiert. Anders als bei Rufai & Bunce (2020) und Wang et al. (2021) werden die Themen der Tweets nicht manuell festgelegt, sondern mit *Machine-Learning* beziehungsweise *Deep-Learning*-Methoden ermittelt. Die Bestimmung der Themencluster ist somit nicht von subjektiven Festlegungen abhängig. Wie von Yun et al. (2016b) wird in dieser Arbeit, anders als in bisherigen Arbeiten zur Kommunikation während der COVID-19-Pandemie, auch die Bedeutung der staatlichen Accounts innerhalb des Netzwerkes untersucht.

3 Theoretische Grundlagen der Methodik

In dieser Arbeit wird eine Reihe von verschiedenen Methodiken und Konzepten miteinander verbunden. Dieses Kapitel erläutert die Grundlagen der in Kapitel 4 beschriebenen Methodiken und soll die theoretischen Zusammenhänge zwischen diesen aufzeigen. Im ersten Teil dieses Kapitels wird das Konzept des *Deep-Learnings* vorgestellt und in Verhältnis zu anderen Methodiken gesetzt. Anschließend wird das *Natural-Language-Processing* (NLP) vorgestellt und die Grundlagen der Arbeit mit Textdaten werden erläutert. Die Analyse der Regierungskommunikation während der COVID-19-Pandemie kann in drei Abschnitte unterteilt werden. Es werden die Themen untersucht, zu denen die Regierung kommuniziert, die Stimmung der Antworten wird geschätzt und die Integration der Regierungskommunikation im Kommunikationsnetzwerk untersucht. Für jede dieser drei Analysen werden verschiedene Konzepte benötigt, deren Grundlagen in diesem Kapitel erläutert werden.

3.1 Deep-Learning

Der Bereich der künstlichen Intelligenz lässt sich in mehrere Segmente, die historisch betrachtet werden können unterteilen. Zu Beginn der Entwicklung der künstlichen Intelligenz wurden Modelle für spezifische Aufgaben manuell über mathematische Regeln definiert (vgl. Goodfellow et al., 2018, S. 2). Diese Expertensysteme genannten Modelle können in einer klar definierten Umgebung gute Ergebnisse erzielen (vgl. Goodfellow et al., 2018, S. 2). Die Regeln und Möglichkeiten bei Schachspielen sind beispielsweise klar beschreibbar. Entsprechend besiegte ein Schachcomputer im Jahr 1997 den Schachweltmeister. Vor allem abstrakte und formale Aufgaben, die für den Menschen eine besondere Herausforderung darstellen, sind für künstliche Intelligenzen lösbar (vgl. Goodfellow et al., 2018, S. 2). Im Gegensatz dazu sind Expertensysteme bei Aufgaben, die schwer zu formalisieren sind, dem Menschen deutlich unterlegen. Es ist zum Beispiel unmöglich, jede Stimmung (*Sentiment*) innerhalb eines Gespräches oder Textes durch Regeln zu definieren. Der grammatische Aufbau und die Bedeutung sind zwar einer Logik unterworfen, die Bedeutung jedes Wortes ist jedoch nicht in jedem Kontext konstant. Die Signifikanz eines Wortes für die Stimmung eines Textes kann je nach Text schwanken und in einem anderen Kontext kann ein Begriff auch das Gegenteil bedeuten. Auch der Mensch hat nicht für jede dieser Möglichkeiten eine Regelung erlernt, sondern handelt intuitiv auf der Grundlage seines bisher gesam-

melten Wissens, welches auf die neue Situation übertragen wird. Künstliche Intelligenzen, die diesen Ansatz verfolgen, werden dem *Machine-Learning* zugeordnet (vgl. Goodfellow et al., 2018, S. 5). Klassische Modelle des *Machine-Learnings* sind die logistische Regression, *Naive Bayes* oder baumbasierte Algorithmen wie *Random Forest* oder *Boosting* (vgl. Goodfellow et al., 2018, S. 3). In diesen Modellen oder Algorithmen werden Strukturen aus den gegebenen Daten extrahiert. Besonders entscheidend sind bei diesen Ansätzen die festgelegten Merkmale. Klassische Modelle des *Machine-Learnings* entscheiden nicht, welche Merkmale für die Problemstellung relevant sind, sondern sie erlernen, wie die Merkmale der Beobachtungen mit den Ergebnissen korreliert sind (vgl. Goodfellow et al., 2018, S. 3). Im *Machine-Learning* werden Modelle nach *supervised* und *unsupervised* unterschieden. *Supervised*-Ansätze extrahieren Informationen auf Basis des Wissens über den wahren Wert im Datensatz. Bei dem Nutzen von *Unsupervised*-Modellen werden keine Informationen über den wahren Wert verwendet.

Insofern bekannt oder vermutbar ist, welche Merkmale mit dem wahren Ergebnis korrelieren, bieten sich klassische *Machine-Learning*-Modelle an. Ist es jedoch unbekannt oder schwer zu beschreiben, welche Merkmale die entscheidenden für eine gute Schätzung sind, eignen sich Modelle des *Representation Learning* (vgl. Goodfellow et al., 2018, S. 4). *Representation Learning* umfasst Modelle wie Autoencoder und ist eine Unterform des *Machine-Learnings*. Bei der Analyse eines Textes auf das Sentiment wird beim *Representation Learning* beispielsweise die Bedeutung von Verben extrahiert. Steigt jedoch die Anzahl der möglichen Faktoren der Variation jedes Datenpunktes, führt dies zu einer erschwerten Extraktion von übergeordneten Merkmalen aus den Daten (vgl. Goodfellow et al., 2018, S. 5). Faktoren der Variation sind unbeobachtete Effekte, die sich auf die Beobachtung auswirken, wie beispielsweise leichte Verschiebungen in der Bedeutung der menschlichen Sprache, welche abhängig von einer Vielzahl an Einflussfaktoren ist. *Deep-Learning* löst dieses Problem, indem die komplexe Darstellung der Daten durch viele einfache Darstellungen abgebildet wird. Im Idealfall extrahiert jede dieser einfachen Darstellungen einen Faktor der Variation (vgl. Goodfellow et al., 2018, S. 6). Die Tiefe erreichen *Deep-Learning*-Modelle durch das Kombinieren und Aufeinander aufbauen von einfachen Darstellungen der Daten. Durch diesen Aufbau können unter anderem Informationen zur Verneinung in einem Satz und ein genutztes Verb zusammengeführt werden (vgl. Goodfellow et al., 2018, S. 6). *Deep-Learning*-Modelle sind aufgrund dieses Konzeptes in der Lage, kleinteilige Strukturen in Daten zu extrahieren, die für Menschen nicht erkennbar sind. Die Grundlagen der Algorithmen, die im *Deep-Learning* angewendet werden,

sind oft inspiriert vom menschlichen Ansatz der Datenverarbeitung. Deshalb und aufgrund der vernetzten und aufeinander aufbauenden Struktur werden die Modelle auch als künstliche neuronale Netze bezeichnet.

Künstliche neuronale Netze bestehen aus verschiedenen Schichten. Jede dieser Schichten besteht aus einzelnen Neuronen, die über Gewichte mit der nächsten Schicht verbunden sind. Jedes Neuron stellt eine Ausprägung der Beobachtung dar. Neuronale Netze werden durch einen Lernalgorithmus so angepasst, dass die Strukturen in den Daten extrahiert werden und die Informationen herausfiltern, die zur Beantwortung der Fragestellung am besten geeignet sind. Dieser Vorgang wird Training genannt. Zum Schätzen eines Wertes durchwandert eine Beobachtung das neuronale Netz. Die Schätzung wird mit dem wahren Wert verglichen und die Information über den Fehler wandert rückwärts durch das Netz. Die Gewichte werden daraufhin auf Grundlage des Fehlers angepasst. Der Aufbau und der Lernalgorithmus künstlicher neuronaler Netze werden in Kapitel 4.1 und 4.1.1 erläutert.

3.2 *Natural-Language-Processing* und semantischer Raum

In dieser Arbeit wird ein umfangreicher Datensatz von Tweets zur COVID-19-Pandemie analysiert. Wie oben beschrieben, handelt es sich bei Tweets um kurze Texte. Die Verarbeitung und die Analyse von Sprache und Texten stellen eine Besonderheit dar und werden in der Disziplin des *Natural-Language-Processings* untersucht. Die Besonderheit von Sprache ist, dass kleine Veränderungen in einem Satz die Bedeutung des Satzes stark ändern können. Gleichzeitig kann dieselbe Situation auf unterschiedliche Art und Weise beschrieben werden. Zum Beispiel weisen Sätze in passivem und aktivem Genus Verbi eine andere Satzstruktur auf, geben jedoch denselben Inhalt wieder (vgl. Goldberg, 2017, S. 1). Neben der Uneindeutigkeit von Sprache wird die Verarbeitung durch eine unendlich große Anzahl der Möglichkeiten an Bedeutungen erschwert (vgl. Goldberg, 2017, S. 2). Angenommen es kann für ein Wort eine eindeutige Definition gefunden werden, so steigt jedoch bereits in einer Phrase die Anzahl an möglichen Bedeutungen des Wortes. Diese möglichen Bedeutungen nehmen mit einer Einbettung des Wortes in Sätze und ganze Texte weiter zu. Für diese unendliche Anzahl an Bedeutungen ist das Definieren von regelbasierten Ansätzen wie Expertensystemen nur eingeschränkt möglich (vgl. Goldberg, 2017, S. 2).

Eine weitere Problematik beim Verarbeiten von Sprache ist es, dass zwischen Wörtern im selben Themenspektrum keine Korrelation in der Darstellung existieren muss. Die

Wörter „Pizza“ und „Hamburger“ bezeichnen zwei sich thematisch nahestehende Gerichte (vgl. Goldberg, 2017, S. 2). In der reinen Darstellungsform teilen sie jedoch, abgesehen vom „a“, keinen Buchstaben. Eine Verarbeitung von Texten anhand einer Korrelation von Buchstaben ist also nicht möglich. Um eine Verarbeitung von Texten zu ermöglichen, wird eine Repräsentation der Bedeutung von Wörtern, Sätzen oder Dokumenten berechnet. Lange Zeit wurde für die Unterscheidung von Wörtern das Verfahren der *one-hot*-Kodierung angewendet. Für jedes Wort im Textkorpus wird ein Vektor aus Nullen erstellt und an einer Stelle von anderen Wörtern im Textkorpus durch eine Eins unterschieden. In dieser Darstellung von Wörtern ist jedes Wort mathematisch unterschiedlich und die Wörter stehen in keinem logischen numerischen Verhältnis zueinander. Weitere mögliche Darstellungsformen von Texten sind die Kombinationen von *Bag of Words* mit *TF-IDF* oder *N-Gram*. Das *Bag of Words*-Verfahren betrachtet den gesamten Textkorpus und ordnet jedem Wort eine andere Zahl zu (vgl. Goodfellow et al., 2018, S. 525). Jeder Satz ist eine Vektordarstellung in der Länge der Anzahl der Wörter. *TF-IDF* wird in die Phasen *Frequency of a Word* (TF) und *Inverse Document Frequency* (IDF) aufgeteilt. In TF wird der normalisierte Anteil jedes Wortes an jedem Dokument berechnet (vgl. Luhn, 1957, S. 315). IDF berechnet die Häufigkeit des Wortes in sämtlichen Dokumenten (vgl. Sparck, 1972, S. 13). Durch Multiplikation von TD und IDF wird ein Score für jedes Wort im Textkorpus berechnet. *TF-IDF* ist eine Kombination von der Bedeutung des Wortes in jedem Dokument und der Häufigkeit des Wortes über sämtliche Dokumente. *TF-IDF* und *Bag of Words* werden häufig mit *N-Gram* kombiniert. Anstatt ein gesamtes Dokument zu betrachten, wird die Kombination der Wörter in einem vorgegebenen Ausschnitt erstellt, beispielsweise können *Bi-* oder *Trie-Gram* erstellt werden. Das *N-Gram*-Verfahren führt zu einer Berücksichtigung von Wortkombinationen in der Verarbeitung (vgl. Goodfellow et al., 2018, S. 515). In dem beschriebenen Verfahren ist eine Abstandsberechnung zwischen diesen Vektoren möglich. Der Abstand beschreibt jedoch keine semantische Relation der Wörter zueinander.

Ein entscheidender Fortschritt im *Natural-Language-Processing* wurde durch die Entwicklung der *Word-Embeddings* erreicht (vgl. Goldberg, 2017, S. 3). Durch diese Entwicklung ist es möglich, einen semantischen Raum zu erschaffen, in dem sämtliche Wörter des Raumes in einem logischen und messbaren Abstand zueinander stehen (vgl. Kapitel 4.2.1). Mathematisch gesehen ist ein semantischer Raum die Darstellung von Wörtern durch eine Wahrscheinlichkeitsschätzung in einem kontinuierlichen Raum (vgl. Bengio et al., 2003, p. S.1141; Schwenk, 2007, S. 514). Hierfür werden die Wörter in nummerische Werte umgewandelt. Jedes Wort wird durch einen Vektor

dargestellt (*Word-Embedding*), wobei die Anzahl der Dimensionen frei wählbar ist. In einem optimalen semantischen Raum sind dann Rechenoperationen mit Wörtern möglich. So ist beispielsweise das Wort „Rom“ aus der Übertragung des Verhältnisses von dem Wort „Frankreich“ zu „Paris“ auf „Italien“ zu ermitteln: - *Paris – France + Italy = Rome* (vgl. Mikolov et al., 2013, S. 9). Dies zeigt, dass in einem semantischen Raum die Wörter nicht nur nummerisch dargestellt werden, sondern die Abstände der Wörter zueinander einen nachvollziehbaren Zusammenhang haben. Anders ausgedrückt entsteht durch die mehrdimensionale Betrachtung der Wörter eine Art Verständnis der Bedeutung. Die Hauptstadt von Frankreich hat dasselbe Verhältnis zu Frankreich wie Rom zu Italien. Innerhalb dieses Raumes ist es ebenso möglich, Dokumentvektoren einzuordnen. Sämtliche Dokumentvektoren stehen in diesem Raum in einem logischen Verhältnis zueinander. Wie schon bei Wörtern steigt die semantische Ähnlichkeit mit einer geringeren Differenz zwischen den Vektoren. Zur Berechnung von *Word-Embeddings* werden künstliche neuronale Netze genutzt. Bengio et al. (2003) und Schwenk (2007) entwickelten neuronale Netze mit einer unterschiedlichen Architektur zur Berechnung von *Word-Embeddings*. Die von Bengio et al. (2003) berechneten *Word-Embeddings* zeigten eine bis zu 24 % bessere Schätzung der Wahrscheinlichkeitsverteilung im Vergleich zum besten *N-Gram/Bag of Words*-Modell (vgl. Bengio et al., 2003, S. 1148). Schwenk (2007) testete das auf Bengio et al. (2003) aufbauende neuronale Netz mit einem englischen, spanischen und französischen Textkorpus. Über alle Sprachen hinweg zeigte das entwickelte neuronale Netz eine präzisere Darstellung der Wörter als das genutzte *N-Gram/Bag of Words*-Modell (vgl. Schwenk, 2007, S. 506-510). 2013 veröffentlichten Mikolov et al. (2013) das neuronale Netz *Word2vec*, welches seitdem eines der am häufigsten genutzten Modelle zur Berechnung der *Word-Embeddings* ist und in dieser Arbeit verwendet wird (vgl. Kapitel 4.2.1) (vgl. Goldberg & Levy, 2014, S. 1).

Auf dem in dieser Arbeit verwendeten Textkorpus bedeutet die Berechnung eines semantischen Raumes, dass jeder Text eines Tweets durch einen Dokumentvektor im Raum dargestellt wird. Innerhalb dieses Raumes haben Tweets mit einer ähnlichen Bedeutung eine geringe Distanz zueinander und es bilden sich Bedeutungscluster.

3.3 **Topic-Modeling von Twitter-Daten**

Die Unterteilung der Tweets zur COVID-19-Pandemie in verschiedene Themen ist der erste analytische Schritt in dieser Arbeit. Durch die Unterscheidung wird zum einen untersucht, welche Themen im gesamten Netzwerk behandelt werden. Zum an-

deren können wie in den wissenschaftlichen Arbeiten, die in Kapitel 2.2.2 vorgestellt wurden, die von der Regierung behandelten Themen analysiert werden. Der Untersuchungsschwerpunkt dieser Arbeit ist die Integration der Regierungskommunikation innerhalb des Netzwerkes. Wird allein die Bedeutung der Regierungsaccounts im gesamten Netzwerk betrachtet, kann dies zu Verzerrungen in der Analyse führen. Beispielsweise können Themen mit einer hohen Anzahl an Tweets, zu denen sich eine Regierung nicht äußert, wie prominente COVID-19 Erkrankungen, dazu führen, dass der Einfluss der Regierungskommunikation im Netzwerk sinkt. Die Bestimmung der Themen (*Topic-Modeling*) wird mit *Deep-Learning* beziehungsweise *Machine-Learning*-Techniken durchgeführt.

Beim *Topic-Modeling* ist es das Ziel, einen Textdatensatz in verschiedene Themen zu unterteilen. Hierfür kann mit zwei unterschiedlichen Ansätzen gearbeitet werden. Der erste ist die Zuordnung anhand von Wahrscheinlichkeiten. Der zweite ist das Aufteilen in Cluster anhand einer Position im Raum. Ansätze wie *Latent Dirichlet Allocation* (LDA) und *Probabilistic latent semantic Analysis* (PLSA) ordnen jedem Dokument im Textdatensatz Wahrscheinlichkeiten zu, einem Thema anzugehören. Die Annahme hinter diesem Prinzip ist, dass Dokumente mit ähnlichen Wörtern ein ähnliches Thema behandeln. Dokumente mit einer ähnlichen Wahrscheinlichkeitsverteilung des Vorkommens der Wörter werden einem Thema zugeordnet. Zum Bestimmen von Abständen von Textdaten muss ein Raum existieren, in dem die Daten in einem numerischen Verhältnis zueinander stehen. Durch die Entwicklung von *Word-Embeddings* (vgl. Kapitel 4.2.1) konnte dies erreicht werden. Ist der Abstand zwischen zwei Dokumenten berechenbar, dann ist eine Vielzahl der Cluster-Algorithmen grundsätzlich dazu in der Lage, Cluster und somit Themen zu bilden.

Der am häufigsten für das *Topic-Modeling* genutzte Algorithmus ist LDA. Aus dem Aufbau der LDA entstehen zwei Nachteile. Einerseits ist das manuelle Setzen der Anzahl an Themen erforderlich und andererseits bestimmen sich die Themen durch die Wahrscheinlichkeitsverteilung der Wörter innerhalb der Themen. Häufige Wörter in den Themen sind entsprechend Wörter, die im normalen Sprachgebrauch häufig vorkommen und keine thematische Bedeutung aufweisen. Ein Beispiel hierfür sind Artikel oder Bindewörter. Um das Rauschen aus den Themen bereinigen zu können, werden Datenbearbeitungsschritte wie das Entfernen von Stoppwörtern genutzt. Dieses Bearbeiten der Daten kann zu einer Verzerrung der Informationen des Datensatzes führen. Auch vor dem in dieser Arbeit verwendeten Ansatz sind Datenbearbeitungss-

chritte notwendig, da Tweets aufgrund der Entstehung ein hohes Rauschen aufweisen (vgl. Kapitel 4.5.2). Der genutzte *Top2vec*-Algorithmus (vgl. Angelov, 2020) ist robuster gegen Einflüsse von Rauschen (vgl. Kapitel 4.2.3). *Top2vec* ist ein auf Abstand der Dokumente basierender Algorithmus. Der *Top2vec*-Algorithmus ordnet jeden Tweet einem Themenvektor zu. Das Thema wird durch die k nächsten Wortvektoren $\tilde{\omega} \subset \Omega$ zu diesem Thema beschrieben. Ein Tweet wird als Dokumentvektor $x \subset X$ dargestellt.

Um verschiedene Modelle zum *Topic-Modeling* miteinander vergleichen zu können, wird die *Mutual Information* innerhalb der Themen berechnet. Die *Mutual Information* gibt an, wie stark der statistische Zusammenhang innerhalb der Themen ist und wie viele Informationen die Tweets X im Thema miteinander teilen (vgl. Witten et al., 2011, S. 139). Wird ein LDA-Modell genutzt, dann wird jedem Tweet die Wahrscheinlichkeit, einem Thema anzugehören, zugeordnet. Die *Mutual Information* eines LDA-Modells errechnet sich aus den *Mutual Informations* zu jedem Thema, gewichtet mit den Wahrscheinlichkeiten der Zugehörigkeit (vgl. Angelov, 2020, S. 9). Im Fall vom *Top2vec*-Algorithmus ist die Zuordnung zu den Themen eindeutig. Die gesamte *Mutual Information* des Modells berechnet sich aus der Summe von *Mutual Informations* der Tweets zu den Wortvektoren. Die *Mutual Information* berechnet sich aus der multivariaten Verteilung $P(x|\tilde{\omega})$ multipliziert mit dem logarithmierten Verhältnis der multivariaten Verteilung zu den Randverteilungen (vgl. Manning et al., 2009, S. 272; Angelov, 2020, S. 10). Vereinfacht ausgedrückt beschreibt die *Mutual Information* die gemeinsamen Verteilungen von x und $\tilde{\omega}$ reduziert um die normalisierte Bedeutung der gemeinsamen Verteilung für x und $\tilde{\omega}$. Teilen x und $\tilde{\omega}$ eine geringe Informationsmenge, dann ist das Verhältnis der gemeinsamen Verteilung zu der Gesamtverteilung gering. Die Bedeutung der multivariaten Verteilung sinkt in der Berechnung der *Mutual Information*. Die gesamte *Mutual Information* ist eine wahrscheinlichkeitsgewichtete Menge an Informationen, die *Probability-weighted amount of Information (PWI)* des Modells und wird über alle Themen $\tilde{t} \in \tilde{T}$ über die k möglichen Wortvektoren in $\Omega_{\tilde{t}}$ und über alle Tweets X berechnet (vgl. Angelov, 2020, S. 10).

$$PWI(\tilde{T}) = \sum_{\tilde{t} \in \tilde{T}} \sum_{x \in X} \sum_{\tilde{\omega} \in \Omega} P(x|\tilde{\omega}) \log \frac{P(x|\tilde{\omega})}{P(x)P(\tilde{\omega})} \quad (1)$$

Curiskis et al. (2020) verglichen verschiedene Ansätze zum *Topic Modeling* miteinander. Hierzu wurde ein Twitter-Datensatz erhoben und Themencluster mit

dem wahrscheinlichkeitsbasierten Ansatz der LDA und den abstandsbasierten Algorithmen *Hierarchical*, *k-means*, *k-medoids* und *Non-negative Matrix Factorization* wurden bestimmt. Die Bestimmung des numerischen Abstands beziehungsweise des semantischen Raumes wurde mit *Doc2vec*, *Word2vec* und *TF-IDF* durchgeführt. Zur Messung der Qualität der Cluster wurden *Normalised Mutual Information* und *Adjusted Mutual Information* berechnet. Die größten gegenseitigen Informationen in den Clustern wurden bei der Kombination von *k-means*-Clustern mit *Doc2vec* erreicht (vgl. Curiskis et al., 2020, S. 13). Mit einem Wert von 0,19 bei *Normalised Mutual Information* wurde eine deutlich stärkere Informationskonzentration im Informationsgehalt der Cluster festgestellt, als es bei LDA mit 0,04 und bei *Word2vec* in Kombination mit *k-means* mit 0,1 der Fall war (vgl. Curiskis et al., 2020, S. 13).

Angelov (2020) verglich *Top2vec* mit LDA und PLSA. Hierzu wurde der Informationsgehalt in den einzelnen Themen auf Basis von *Mutual Information* berechnet und miteinander verglichen. Um eine höhere Aussagekraft erhalten zu können, wurden zwei verschiedene Datensätze analysiert: ein Zeitungsartikel-Datensatz mit entsprechend geringem Rauschen und einer hohen Länge der einzelnen Dokumente sowie ein Datensatz mit Kommentaren auf Yahoo und entsprechend mehr Rauschen und kurzen Texten. In beiden Datensätzen ist der Informationsgehalt in den durch *Top2vec* berechneten Themen deutlich höher als bei Themen, die durch LDA oder PLSA bestimmt wurden (vgl. Angelov, 2020, S. 15-16). Da die drei Ansätze die Anzahl der Themen unterschiedlich hoch bestimmen, wurde der Informationsgehalt für eine Spanne von Themen von eins bis hundert Themen berechnet. *Top2vec* zeigte auch außerhalb der optimal bestimmten Themenanzahl über sämtliche Themenzahlen hinweg den höchsten Informationsgehalt in den Themenclustern. Mit steigender Anzahl an Themen erhöht sich die Differenz im Informationsgehalt von *Top2vec* zu LDA und PLSA (vgl. Angelov, 2020, S. 16).

3.4 Sentiment-Analysen

Für die Analyse der Integration der Regierungskommunikation auf Twitter werden als Verbindungen im Netzwerk die Antworten auf die Tweets genutzt. Um ein besseres Verständnis der Reaktionen auf die Tweets zu erlangen, soll die Stimmung oder das Sentiment der Antworten durch ein künstliches neuronales Netz geschätzt werden. Hierfür werden verschiedene Netzwerkarchitekturen mit einem *Supervised*-Ansatz auf einem Tweet-Datensatz trainiert. Die Netzwerkarchitektur, welche die höchste Genauigkeit bei der Schätzung des Sentiments eines unbekannten Testdatensatzes

aufweist, wird zur Bestimmung der Sentiments der Antworten auf die Tweets genutzt.

Für das Durchführen von Sentiment-Analysen, also das Schätzen der Stimmung eines Textes, sind verschiedene Techniken bekannt. Lexikon-basierende Ansätze ermitteln das Verhältnis von positiver zu negativer Stimmung im Text auf der Grundlage der Anzahl der vorher als positiv oder negativ definierten Wörter (vgl. Goldberg 2017, S. 185). Durch die Beschränkung auf Lexika in der Schätzung werden grammatische Effekte wie die Negation von Wörtern allerdings nicht erfasst. Eine weitere Möglichkeit ist das händische Festlegen von Regelungen in Expertensystemen. In diesen müssen Zusammenhänge manuell erkannt und festgelegt werden. Aufgrund der geringeren Genauigkeit dieser Ansätze werden in der aktuellen Forschung beide selten genutzt und auch hier nicht weiter beschrieben. Die am häufigsten verwendeten Ansätze sind Algorithmen des *Machine-Learnings*. *Machine-Learning*-Algorithmen extrahieren die Strukturen in den Daten selbstständig und sind dadurch weniger verzerrungsanfällig als Expertensysteme, da auch für den Menschen nicht erkennbare Zusammenhänge erkannt werden können. Tweets zeichnen sich im Vergleich zu anderen Dokumenten durch ihre geringere Anzahl an Wörtern und das hohe Rauschen aus. Die Konzepte der genutzten Algorithmen unterscheiden sich nicht von anderen Dokumenten, jedoch hat diese abweichende Struktur einen Effekt auf die Genauigkeit der Modelle. Um dies in der folgenden Erläuterung berücksichtigen zu können, werden ausschließlich Modelle genannt, die bereits auf *Social Media*-Dokumente oder Dokumente vergleichbarer Strukturen angewendet wurden.

Der häufigste Ansatz zum Schätzen von Sentiments in Textdaten ist die Verwendung von *Supervised*-Modellen. Ein Datensatz mit Informationen über den wahren Wert wird in einen Trainings und in einen Testdatensatz unterteilt. Die Modelle werden auf dem Trainingsdatensatz trainiert und anschließend wird der Testdatensatz geschätzt. Das Training wird durch die Minimierung oder Maximierung einer Verlustfunktion durchgeführt. Innerhalb des Trainings wird der Trainingsdatensatz häufig in einen eigentlichen Trainings- und einen Validierungsdatensatz unterteilt. Besteht der Modellalgorithmus aus mehreren Iterationen, so wird in jeder Iteration der Validierungsdatensatz geschätzt. Die Genauigkeit einer Klassifizierung wird durch die Korrektklassifikationsrate (*Accuracy*) angegeben (vgl. Gleichung 2). Die *Accuracy* ist das Verhältnis aller richtig negativen (RN) und richtig positiven (RP) Schätzungen zu allen falsch positiven (FP) und falsch negativen (FN) Schätzungen (vgl. Burkov, 2019).

$$Accuracy = \frac{RN + RP}{RN + RP + FP + FN} \quad (2)$$

Eine der herausforderndsten Aufgaben beim Trainieren von *Supervised-Machine-Learning*-Modellen ist die Vermeidung von *Overfitting* (vgl. Goldberg, 2017, S. 253). Sind neuronale Netze zu stark an einen Trainingsdatensatz angepasst, kommt es beim Schätzen von bisher unbekannten Daten zu einer Verzerrung. Dies kann beispielsweise bei zu tiefen Netzen oder einer zu hohen Anzahl an Iterationen eintreten. Tritt eine Verzerrung ein, dann liegt die Schätzung systematisch neben dem wahren Wert (vgl. James et al., 2013, S. 34). Ist die Anpassung an die Trainingsdaten hingegen zu schwach, um allgemeine Strukturen erkennen zu können, so ist die Varianz um die Schätzung des wahren Wertes hoch und dies führt zu *Underfitting*. Während des Trainings soll das Modell daher möglichst so trainiert werden, dass es weder zu *Overfitting* noch zu *Underfitting* kommt (vgl. Goldberg, 2017, S. 253). Dies führt zu einer Abwägung zwischen einem genau geschätzten, aber verzerrten Wert und einer Schätzung mit hoher Varianz, jedoch mit geringer Verzerrung (vgl. James et al., 2013, S. 34-35). Die optimale Wahl der Modellparameter zum Trainieren der Modelle kann durch Hyperparametertraining untersucht werden. Ein neuronales Netz wird mit verschiedenen möglichen Parameterkombinationen trainiert. Anschließend wird das Modell mit der höchsten Genauigkeit bei der Schätzung eines Testdatensatzes ausgewählt. Eine weitere Möglichkeit zum Abwagen zwischen Varianz und Verzerrung sind verschiedene Anpassungen in der Architektur des Netzes. Diese Anpassungen sind Regularisierungstechniken und werden in Kapitel 4.3.3 beschrieben.

Von den klassischen *Machine-Learning*-Algorithmen finden zumeist die SVM und *Naive Bayes Classifier* (NB) Anwendung. Sluban et al. (2015) trainierten einen manuell klassifizierten Tweet-Datensatz mit SVM und analysierten die Sentiment Strukturen innerhalb von Netzwerken zu Umweltthemen wie beispielsweise dem Klimawandel. Das trainierte Modell erreichte eine Genauigkeit von bis zu 60 % bei der Schätzung von positiv und negativ. Zur Analyse eines Zusammenhangs zwischen Börsenkursen von Unternehmen und dem Sentiment in Tweets trainierten Smailović et al. (2014) ebenfalls einen binären SVM-Klassifikator. Durch umfangreiche Variation in den Datenbearbeitungsschritten wurde ein Genauigkeit von bis zu 78 % erreicht (vgl. Smailović et al., 2014, S. 7). Für die Erstellung eines Frühwarnsystems in Krisensituationen trainierten Brynielsson et al. (2014) eine SVM und einen NB zur Mehrfachklassifizierung von Stimmungen in Tweets. Hierfür wurde ein manuell

klassifizierter Trainingsdatensatz genutzt, welcher im Rahmen vom Hurrikan Sandy erhoben wurde. Bei der Schätzung von mehr als zwei Klassen erreichten die Modelle eine Genauigkeit von 60 % (vgl. Brynielsson et al., 2014, S. 5). Mit der Stimmung auf Twitter während Krisensituationen beschäftigten sich darüber hinaus auch Ruz et al. (2020). Durch die binäre Schätzung des Sentiments wurde mit SVM eine *Accuracy* von 81 % und mit NB eine Genauigkeit von 74 % erreicht.

Durch neue Ansätze im *Deep-Learning* in Kombination mit zunehmendem Zugang zu hoher Rechenleistung konnte die Forschung mit Sentiment-Analysen mithilfe der *Deep-Learning*-Technologie intensiviert werden (vgl. Yadav & Vishwakarma, 2020, S. 4339). Besonders *Convolutional Neural Networks* (CNN) und *Long Short-term Memory* (LSTM) Algorithmen weisen aufgrund der dahinterliegenden Konzepte ein hohes Potenzial auf.⁷ Die *Accuracy* von CNN und LSTM untersuchte Sosa (2017) in einem modellbasierten Experiment. Eingeschlossen in das Experiment wurden die Effekte von vortrainiertem *Word-Embeddings* (vgl. Kapitel 4.2.1), die des frühen Abbruches und die Höhe des genutzten *Dropout*-Wertes.⁸ Für CNN wurde eine *Accuracy* von 66 % und für das LSTM-Modell eine *Accuracy* von 72 % ermittelt. Des Weiteren wurde die *Accuracy* der Kombination von CNN und LSTM verglichen und untersucht inwiefern die Reihenfolge von CNN und LSTM in der Netzarchitektur die Genauigkeit beeinflusst. CNN-LSTM erreichte eine *Accuracy* von fast 70 % und LSTM-CNN eine Genauigkeit von 75 %. Die Untersuchungen des Einflusses von *Word-Embeddings* zeigen, dass im Modell trainierte *Word-Embeddings* über die untersuchten Modelle hinweg die höchste Genauigkeit aufweisen. Zudem wurde dargelegt, dass ein früher Abbruch eines Modells und die Höhe des *Dropout*-Wertes einen starken Effekt auf die Korrektklassifikationsrate der Modelle haben. Uysal & Murphey (2017) untersuchten die *Accuracy* von CNN und LSTM auf vier verschiedenen Datensätzen. Durch den Vergleich über mehrere Datensätze hinweg wurde gezeigt, dass die Vorhersage des Sentiments von Tweets eine besondere Herausforderung darstellt. Im Twitter-Datensatz erreichte das LSTM eine *Accuracy* von 71 % und das CNN-LSTM eine *Accuracy* von 67 %. Bei der Anwendung der Modelle auf Filmkommentare konnte hingegen mit einem LSTM Modell eine *Accuracy* von 89 % erzielt werden. Die Mehrzahl der Sentiment-Analysen nutzt englischsprachige Tweets, Cieliebak et al. (2017) hingegen erstellten einen deutschsprachigen Datensatz und erreichten mit SVM einen F1 Wert von bis zu 61 % und mit CNN 65 % (vgl. Cieliebak et al., 2017, S. 49).

⁷Die Methodik von CNN und LSTM wird in Kapitel 4.3 erläutert.

⁸Der theoretische Hintergrund eines *Dropout* in künstlichen neuronalen Netzen wird in Kapitel 4.3.3 erläutert.

In den vorgestellten Untersuchungen zu Sentiment-Analysen mit SVM oder NB wurde mit kleinen, spezifischen Datensätzen gearbeitet. In dieser Arbeit wird eine deutlich höhere Anzahl an Tweets verwendet. Dies bezieht sich sowohl auf den Datensatz, auf dem die Modelle trainiert werden, als auch auf die später zu schätzenden Antworten auf die Tweets zur COVID-19-Pandemie. Dieses bei der ersten Betrachtung sehr spezifisch wirkende Thema umfasst eine hohe Anzahl an Unterthemen. Durch den ungewöhnlich hohen Stellenwert in der Gesellschaft über einen langen Zeitraum hinweg wird auf Twitter im Zusammenhang mit COVID-19 eine Reihe an Themen behandelt. Ein zu spezifisches Trainieren von Modellen zum Schätzen des Sentiments würde zu Verzerrungen führen. Aufgrund dieser inneren Struktur der Daten bieten sich *Deep-Learning*-Techniken wie neuronale Netze an. Diese können mehrere spezifische Strukturen innerhalb der Daten gleichzeitig extrahieren und sind somit bei einer höheren Varianz der Daten flexibler.

3.5 Soziale Netzwerkanalyse

In dieser Arbeit werden zwei verschiedene Fragestellungen untersucht. Erst wird mithilfe von *Topic-Modeling* analysiert, zu welchen Themen die Regierung auf Twitter kommunizierte. Die zweite Fragestellung bezieht sich auf den Einfluss oder die Integration der Regierungskommunikation innerhalb dieser Themen im Netzwerk. Wenn angenommen wird, dass die Regierung verschiedene Tweets zur Aufklärung und Informationen über Maßnahmen für die Bevölkerung postet, dann stellt sich nicht nur die Frage, zu welchen Themen kommuniziert wird, sondern auch, ob die Tweets einen Einfluss im Netzwerk haben oder die Informationen nur einen kleinen Teil des Netzwerkes erreichen. Zum Untersuchen des Einflusses der Regierungskommunikation in einzelnen Themen wird die soziale Netzwerkanalyse genutzt, wobei darüber hinaus in Kombination mit der Sentiment-Schätzung auch die Reaktion auf die Regierungstweets betrachtet wird.

Ein Netzwerk besteht aus Knoten $\mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_l)$, die über Gewichte miteinander verbunden sind, wobei allerdings nicht zwischen allen Knoten eine Verbindung bestehen muss (vgl. Borgatti et al., 2009, S. 894). Ein Knoten wird in zwei Dimensionen mit $\mathbf{v}_i = (v_{i1}, v_{i2})$ beschrieben. Das Kernkonzept von Netzwerkanalysen ist das Prinzip der Struktur. Angenommen, eine Gruppe von Wissenschaftlern diskutiert über gleichwertige Ansätze zur Lösung einer Forschungsfrage, dann kann sich das Ergebnis der Diskussion abhängig von den Beziehungsmustern in der Gruppe verändern. Die Struktur des Netzwerkes verändert das Ergebnis (vgl. Borgatti et al.,

2009, S. 894). Ein weiterer Ansatz, ein soziales Netzwerk zu betrachten, ist eine von der Naturwissenschaft inspirierte Sicht. In dieser wird angenommen, dass sich der Zustand eines Knotens teilweise anhand seiner Position im Netzwerk bestimmt (vgl. Borgatti et al., 2009, S. 894). Der Charakter eines Menschen entwickelt sich beispielsweise in einem umfangreichen Netzwerk an Verknüpfungen mit Freunden und Bekannten. Auf die Fragestellung bezogen folgt daraus, dass die isolierte Betrachtung der Regierungskommunikation dazu führen kann, dass Informationen vernachlässigt werden.

Bei dem Erstellen von sozialen Netzwerken wirken drei Mechanismen: der theoretische Mechanismus, der Anpassungsmechanismus und der Bindungsmechanismus (vgl. Borgatti et al., 2009, S. 894-895). Alle drei Mechanismen sind schwierig zu messen, beeinflussen jedoch das Netzwerk und sind deshalb zu berücksichtigen. Der theoretische Mechanismus beschreibt eine Übertragung zwischen Knoten, woraufhin sich der Zustand eines Knotens ändert (vgl. Borgatti et al., 2009, S. 894). Bei der Analyse der Regierungskommunikation fließen Informationen von einem zum anderen Knoten. Die erste Information fließt vom Tweet zum Antwortgeber. Die Antwort auf einen Tweet steht somit immer im Kontext dieses Tweets. Gleichzeitig fließen jedoch auch Informationen von anderen Tweets zum Antwortenden. Führt dieser Einfluss zu einer weiteren Antwort, so werden die Einflüsse beobachtbar. Es ist jedoch davon auszugehen, dass ein hoher nicht beobachtbarer Informationsfluss zum Antwortenden existiert, wenn möglich sollte deshalb ebenfalls die Beziehungsstruktur der Antwortgeber untersucht werden. Der Anpassungsmechanismus besagt, dass Knoten mit ähnlichen Umwelteinflüssen homogener werden (vgl. Borgatti et al., 2009, S. 895). Auf die Kommunikation im Twitter-Netzwerk übertragen bedeutet dies, dass Accounts, die eine ähnliche Kommunikationsstruktur wie andere Accounts aufweisen, diesen thematisch nahestehen. Der dritte Mechanismus ist der Bindungsmechanismus. Dieser sagt aus, dass sich soziale Knoten aneinander binden können und dadurch eine neue Einheit entsteht, welche die Gemeinsamkeiten der einzelnen Knoten zusammenfasst (vgl. Borgatti et al., 2009, S. 895). Aus diesem und dem Anpassungsmechanismus folgt, dass sich in den sozialen Netzwerken Cluster entwickeln und diese als ein gemeinsamer Knoten aufgefasst werden können.

Bei der sozialen Netzwerkanalyse von Twitter-Daten wird eine hohe Anzahl an Personen betrachtet. Im einfachsten Fall sind die Gewichte innerhalb dieses Netzwerkes fest definiert. In der Darstellung ergibt sich eine unendlich hohe Anzahl an Möglichkeiten, das Netzwerk darzustellen. Um eine interpretierbare Darstellungsform erreichen zu

können, wurden verschiedene Algorithmen entwickelt. Aufgrund der hohen Anzahl an Datenpunkten wird in dieser Arbeit *OpenOrd* (Martin et al., 2011) verwendet.

Die öffentliche Information auf Twitter kann in drei unterschiedliche Arten unterteilt werden: dem Folgen, dem Retweeten und dem Antworten. Das Folgen einer Person auf Twitter signalisiert Interesse an den Informationen, die durch diese Person geteilt werden. Tweets der Person, der gefolgt wird, erscheinen in der eigenen *Timeline*. Hier ist jedoch unklar, ob die geteilte Information gelesen wird. Weiterhin ist das Messen eines Zustimmens oder Ablehnens der Information nicht möglich. Cha et al. (2010) fanden Indikationen dafür, dass die Twitter-Accounts mit der höchsten Anzahl an *Followern* nicht die Accounts mit dem höchsten Einfluss auf die Informationsverteilung im Netzwerk sind. Beim Retweeten wird der Tweet eines anderen Accounts über den eigenen Account mit den *Followern* geteilt. Diese Handlung symbolisiert, dass eine Auseinandersetzung mit dem Tweet stattgefunden hat. Die Information wurde aufgenommen und verarbeitet. Gleichzeitig zeigt das Retweeten eine gesteigerte persönliche Wichtigkeit des Tweets an, da der Tweet mit den eigenen *Followern* geteilt wird. Retweets können kommentiert oder auch unkommentiert sein. Ein Retweet kann also nicht in allen Fällen auf den Sentiment untersucht werden. Antworten Nutzer jedoch auf Tweets, dann kann hingegen das Sentiment der Antwort gemessen werden. Außerdem ist bei einer Antwort auf einen Tweet auch davon auszugehen, dass sich der Nutzer mit diesem Tweet auseinandergesetzt hat. Die Analyse von Antworten auf die Tweets führt also dazu, dass eine genauere Differenzierung innerhalb des Netzwerkes möglich ist. Es werden nicht nur die Position und die Verbindung zu anderen Accounts berücksichtigt, sondern auch das Sentiment der Verbindungen.

Für die Analyse der Regierungskommunikation innerhalb des Antwortnetzwerkes werden in dieser Arbeit drei verschiedene Messwerte betrachtet, die *Degree-Centrality* (vgl. Shaw, 1954), die *Eigenvector-Centrality* (vgl. Bonacich, 1972) und die *Betweenness-Centrality* (vgl. Freeman, 1977; Shaw, 1954). *Degree-Centrality* und *Eigenvector-Centrality* beziehen sich auf die Stellung von Knoten im Netzwerk. Es handelt sich hierbei um zwei verschiedene Ansätze zum Messen des Einflusses eines Knotens im Netzwerk (vgl. Cherven, 2015). Die *Degree-Centrality* (C_D) misst die Anzahl an Verbindungen zu anderen Knoten, dabei kann zwischen eingehenden und ausgehenden Verbindungen unterschieden werden (vgl. Gleichung 3) (vgl. Shaw, 1954, S. 139). Besteht zwischen zwei Knoten v_i und v_j eine Verbindung, dann ist $\alpha(v_i, v_j) = 1$ und anderenfalls ist $\alpha(v_i, v_j) = 0$ (vgl. Freeman, 1978, S. 220). Im Kontext des Vorgehens dieser Arbeit hat der Account im Netzwerk mit den meisten

Antworten die höchste *Degree-Centrality*.

$$C_D = \sum_{i=1}^I \alpha(\mathbf{v}_i, \mathbf{v}_j) \quad (3)$$

Die *Eigenvector-Centrality* C_E betrachtet die Verbindungen eines Knotens zu den weiteren Knoten nicht exklusiv, sondern darüber hinaus auch, mit welchen Knoten die Verbindung besteht (vgl. Gleichung 4) (vgl. Cherven, 2015). Ein Knoten kann eine relativ geringe *Degree-Centrality* aufweisen. Bestehen jedoch Verbindungen zu anderen Knoten mit vielen Verbindungen, dann ist die *Eigenvector-Centrality* hoch (vgl. Bonacich, 1972, S. 115). Ein Account wird im Netzwerk als wichtig betrachtet, wenn die direkten Nachbarn wichtig sind. Im Twitter-Netzwerk ist ein Account mit einer hohen *Eigenvector-Centrality* ein Account, auf dessen Tweets vor allem User antworten, die auf viele verschiedene Accounts antworten und eine hohe Aktivität aufweisen. In der Adjazenzmatrix $\alpha = (\alpha_{11}, \alpha_{12}, \dots, \alpha_{ij})$ ist $\alpha_{ij} = 1$, wenn \mathbf{v}_i mit \mathbf{v}_j verbunden ist, und $\alpha_{ij} = 0$, wenn nicht (vgl. Bonacich, 1972, S. 113). Die *Eigenvector-Centrality* eines Knoten ist (vgl. Das et al., 2018, S. 13):

$$C_E = \sum_{j=1}^I \alpha_{ij} \mathbf{v}_j \quad i = 1, 2, \dots, I \quad (4)$$

Mit der *Betweenness-Centrality* C_B werden Knoten im Netzwerk ermittelt, die Bereiche des Netzwerkes miteinander verbinden, welche eine große Distanz zueinander aufweisen (vgl. Gleichung 5). Knoten, die eine hohe *Betweenness-Centrality* haben, können als Brücke zwischen zwei Teilnetzwerken im Gesamtnetzwerk betrachtet werden (vgl. Cherven, 2015). Für die Berechnung der *Betweenness-Centrality* werden die kürzesten Verbindungen aller Knoten untereinander einbezogen (vgl. Freeman, 1977, S. 36). Nicht jeder Knoten ist direkt mit jedem Knoten verbunden. Die kürzeste Verbindung von einem Knoten zu einem anderen kann über einen dritten Knoten v_k gehen. Dieser Knoten kann theoretisch den Informationsfluss innerhalb der Verbindung unterbrechen. Die *Betweenness-Centrality* eines Knotens ist die Summe der kürzesten Verbindung zwischen zwei Knoten zu liegen. (vgl. Gleichung 5)(vgl. Freeman, 1977, S. 37). $\tilde{\alpha}_k$ ist eins wenn die Verbindung von \mathbf{v}_i zu \mathbf{v}_j die kürzeste Verbindung zwischen den Knoten darstellt.

$$C_B = \sum_{i \neq j \neq k} \tilde{\alpha}_k(\mathbf{v}_i, \mathbf{v}_j) \quad (5)$$

4 Methodik und Daten

Zum Untersuchen der Regierungskommunikation und der Integration der Regierung in den Kommunikationsnetzwerken während der COVID-19-Pandemie wird eine Reihe von verschiedenen Methoden genutzt. Sowohl zum Extrahieren der Themen aus dem COVID-19 Datensatz als auch zur Bestimmung des Sentiments innerhalb der Kommunikationsnetzwerke werden auf Prinzipien des *Deep-Learnings* aufbauende Modelle verwendet. Im ersten Teil des Kapitels wird in die Grundlagen der künstlichen neuronalen Netze eingeführt und beschrieben, wie diese Netze trainiert werden können. Anschließend wird das das *Top2vec*-Modell vorgestellt, welches auf einem durch ein neuronales Netz berechneten semantischen Raum aufbaut. Die darauf folgende Erläuterung beschreibt zwei spezifische Arten von künstlichen neuronalen Netzen die in anderen Arbeiten eine hohe Genauigkeit bei der Schätzung des Sentiments erzielten (vgl. Kapitel 3.4). Abschließend werden die Datenbeschaffung und die Datenaufbereitung dargestellt.

4.1 Künstliche neuronale Netze

In dieser Arbeit wird mit verschiedenen Arten von künstlichen neuronalen Netzen gearbeitet. Im ersten Schritt werden die allgemeinen Architekturen und zugrundeliegende Konzepte von künstlichen neuronalen Netzen zur Klassifizierung von Daten beschrieben (vgl. Abbildung 1). In dieser Beschreibung erfolgt eine Orientierung an dem einfachsten möglichen Aufbau eines Netzes, dem *Artificial Neural Network*. Der vorgestellte Aufbau und die Konzepte können auf die später beschriebenen Netze übertragen werden (vgl. Kapitel 4.3). Zur Vereinfachung der Darstellung wird in der Beschreibung der neuronalen Netze angenommen, dass nur eine Beobachtung das neuronale Netz durchwandert.⁹

Künstliche neuronale Netze werden in drei Schichten unterteilt, die je nach Modell weitere Schichten enthalten können. Diese drei Schichten sind die Eingabeschicht, die verdeckte Schicht und die Ausgabeschicht (vgl. Goodfellow et al., 2018, S. 7). Dabei

⁹Im späterem Verlauf ist die gesamte Eingabe in das Netz ein Tensor in fünf Dimensionen. Bei gleichbleibendem Konzept würde die Berücksichtigung aller Beobachtungen die Lesbarkeit unnötig erschweren.

stellen die Dateninformationen, die in das Netz gegeben werden, die Eingabeschicht \mathbf{x} dar und die Ausgabeschicht des neuronalen Netzes ist $\hat{\mathbf{y}}$, wobei $\mathbf{x} = (x_1, x_2, \dots, x_r)$ die Ausprägungen der Eingabe und im binären Fall $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2)$ die geschätzten Wahrscheinlichkeiten, einer Klasse anzugehören, sind. Entsprechend ist $\mathbf{y} = (y_1, y_2)$ und y_1 oder y_2 sind eins, wenn die Beobachtung der Klasse zugeordnet wird.

Die Schichten werden über Gewichte \mathbf{W}^s miteinander verbunden (vgl. Gleichung 6 und 7). $s = 1, \dots, r$ beschreibt die Anzahl der verdeckten Schichten. Für ein einfaches Modell ist jede Schicht \mathbf{h}^s innerhalb der verdeckten Schichten eine Abbildung der vorhergehenden Schicht \mathbf{h}^{s-1} (vgl. Goodfellow et al., 2018, S. 218). Dabei ergibt sich die erste verdeckte Schicht aus der Eingabeschicht. Darüber hinaus stellt g^s die Aktivierungsfunktion der jeweiligen Schicht und \mathbf{b}^s die Verzerrung dar (vgl. Goodfellow et al., 2018, S. 212). Die Ausgabe einer Schicht, wenn keine Aktivierungsfunktionen g^s wie *Rectified Linear Unit* (ReLU) genutzt wurden, ist $\mathbf{a}^s = (\mathbf{W}^s)^T \mathbf{h}^{s-1} + \mathbf{b}^s$ mit $\mathbf{a}^s = (a_1^s, a_2^s, \dots, a_r^s)$.¹⁰ Zum Berechnen der Ausgabeschicht wird in den neuronalen Netzen dieser Arbeit eine Softmax-Funktion genutzt. Die Eingabe in die Softmatrix ist die letzte verdeckte Schicht \mathbf{z} .¹¹

$$\mathbf{h}^{(1)} = g^{(1)}((\mathbf{W}^{(1)})^T \mathbf{x} + \mathbf{b}^{(1)}) \quad (6)$$

$$\mathbf{h}^{(2)} = g^{(2)}((\mathbf{W}^{(2)})^T \mathbf{h}^{(1)} + \mathbf{b}^{(2)}) \quad (7)$$

Die Schicht h^s enthält $i = 0, 1, \dots, r$ Neuronen, die über die Gewichte \mathbf{W}^s mit den Neuronen der vorherigen und der nachfolgenden Schicht verbunden sind. Die Anzahl der Neuronen muss nicht in allen Schichten identisch sein. Zur Vereinfachung wird jedoch angenommen, dass alle Schichten $\mathbf{h}^{(1)}$ bis $\mathbf{h}^{(r)}$ die gleiche Anzahl an Neuronen haben. Da jede Schicht aus mehreren Neuronen besteht, ist $\mathbf{W}^s \in \mathbb{R}^{r \times r}$.

Beim Vorwärts durchlaufen des Netzes wandern die Informationen durch w_{ji} von jedem Neuron j aus \mathbf{h}^{s-1} zu jedem Neuron i in \mathbf{h}^s . Daraus folgt, dass $\mathbf{h}^s = (h_0^s, h_1^s, \dots, h_r^s)$ und $\mathbf{h}^{s-1} = (h_0^{s-1}, h_1^{s-1}, \dots, h_r^{s-1})$ ist. Ein Ausgabeneuron a_i^s einer Schicht ergibt sich aus der gewichteten Summe der Ausgaben aus der Schicht \mathbf{h}^{s-1} zuzüglich der Verzerrung \mathbf{b}^{s-1} (vgl. Goodfellow et al., 2018, S. 234; LeCun et al., 2015, S. 437). Durch

¹⁰Eine ausführliche Erläuterung von ReLU findet sich in Kapitel 4.3.1.

¹¹Beschreibung der Softmax-Funktion und Ausgabeschicht finden sich in Kapitel 4.1.2.

die Betrachtung der Neuronen in der Berechnung wird aus Gleichung 7 Gleichung 8. Die Verzerrung fügt dem Modell eine trainierbare Konstante hinzu. Während eine Veränderung der Gewichte die Steigung der Funktion verschiebt, führt eine Anpassung der Verzerrung äquivalent zu der Konstanten in einer linearen Funktion zu einer Verschiebung der Funktion auf den Achsen. Dies wird durch die Annahme, dass $w_{0i}^s = b_i^s$ und $h_0^{l-1} = 1$ ist, deutlich (vgl. Rojas, 1996, S. 165). Aus diesem Grund wird die Verzerrung in dieser Arbeit nicht weiter notiert.

$$a_i^s = \sum_{j=0}^{r^{l-1}} w_{ji}^s h_j^{s-1} \quad (8)$$

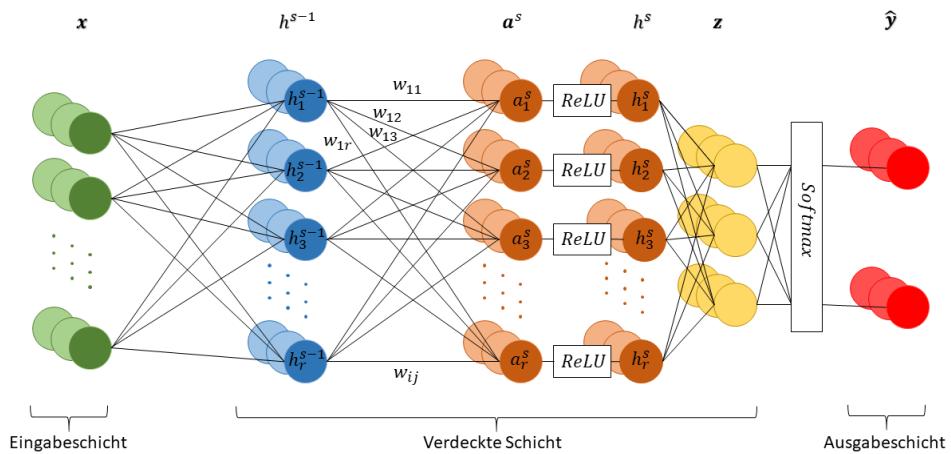


Abbildung 1: Schaubild eines neuronalen Netzes (eigene Darstellung)

4.1.1 Lernalgorithmus

Wie in Kapitel 3.1 beschrieben lernen künstliche neuronale Netze, die auf *Supervised Learning* basieren, durch den Fehler in den Schätzungen über mehrere Iterationen. In diesem Kapitel wird der *Supervised*-Lernalgorithmus dargestellt, der in den Modellen in den Kapiteln 4.2.1 und 4.3 Anwendung findet.

Der Lernalgorithmus beim *Supervised Learning* basiert auf zwei aufeinander aufbauenden Berechnungen, die während des Trainings iteriert werden. Beide Iterationen sind eine Epoche ϵ . Im ersten Schritt durchlaufen die Trainingsdaten das künstliche neuronale Netz und berechnen $\hat{\mathbf{y}}$. Anschließend wird die Güte der Schätzung anhand einer Verlustfunktion $L = L(\hat{\mathbf{y}}, \mathbf{y})$, die den geschätzten sowie den wahren Wert \mathbf{y} beinhaltet.

tet, bewertet. Die Gewichte werden in der ersten Iteration zufällig gewählt und auf Grundlage des Fehlers der Schätzung angepasst. Um eine Überreaktion auf den Fehler während des Trainings zu vermeiden, sind die zufällig bestimmten Gewichte kleiner eins. Das Ziel des gesamten Lernalgorithmus ist es, die Verlustfunktion zu minimieren (vgl. Goodfellow et al., 2018, S. 90). Die Anpassung der Gewichte erfolgt mittels eines Gradientenverfahrens (vgl. Goodfellow et al., 2018, S. 90-93). Die Richtung des Gradienten wird während der *Backpropagation* genutzt, um die Gewichte anzupassen und die Verlustfunktion zu optimieren, indem der zu verändernde Parameter in die Richtung angepasst wird, die der des Gradienten entgegengesetzt ist (vgl. Goodfellow et al., 2018, S. 62; Goldberg, 2017, S. 30). Bei der *Backpropagation* wandert die Information des Fehlers von den Neuronen i in \mathbf{h}^s zu den Neuronen j in \mathbf{h}^{s-1} . Die Anpassung der Gewichte führt zu unterschiedlichen Stärken der Verbindungen zwischen den einzelnen Neuronen aus verschiedenen Schichten. Die Kombination aus Gewicht und Neuron kann als Extraktion einer bestimmten Information aus den Eingabedaten \mathbf{x} betrachtet werden (vgl. Goodfellow et al., 2018, S. 7). Stellt sich diese Information im Laufe des Lernens als nicht zielführend zur richtigen Voraussage von y_i heraus, sinkt das dem Neuron zugeordnete Gewicht. Beim Vorwärts durchlaufen des Netzwerkes und Schätzen von \hat{y}_i reduziert sich der Einfluss dieser Informationen von x_i auf \hat{y}_i . Ein neuronales Netz in Kombination mit einem Lernalgorithmus extrahiert selbstständig die Ausprägungen aus den Daten, die einen hohen Informationsgehalt besitzen. Neben den Ausprägungen werden durch das Netz auch Beziehungen in den Daten nachvollzogen und in der Schätzung berücksichtigt (vgl. Goodfellow et al., 2018, S. 7). Die Gewichte eines final trainierten Netzes sind so angepasst, dass die Kombination der extrahierten Ausprägungen die Verlustfunktion minimiert.

Bei tiefen neuronalen Netzen kommt es zu einer großen Anzahl möglicher Kombinationen von Gewichten. Ein Durchlaufen des Lernalgorithmus über sämtliche möglichen Gewichtskombinationen würde eine hohe Rechenintensität erfordern und ist daher ab einer gewissen Netztiefe nicht möglich (vgl. Goodfellow et al., 2018, S. 311). Durch Anwendung des Gradientenverfahrens werden Informationen über die Richtung, in welche die Gewichte angepasst werden müssen, genutzt und die Effizienz wird erhöht (vgl. Goodfellow et al., 2018, S. 90). Damit ist es nicht erforderlich, sämtliche Kombinationen von Gewichten zu durchlaufen. Die Richtung und Stärke der Anpassung bestimmen sich durch die Steigung der Verlustfunktion, abhängig von einer Veränderung des Gewichts w_{ij} in Schicht \mathbf{h}^s zum Zeitpunkt ϵ . Diese Funktion $\frac{\partial L}{\partial w_{ij}^s}$ wird Gradient genannt (vgl. Rojas, 1996, S. 164).

Um die Wahrscheinlichkeit einer Überreaktion in der Anpassung der Gewichte und somit einen Anstieg der Verlustfunktion zu reduzieren, wird der Einfluss des Gradienten auf die Anpassung der Gewichte durch die Lernrate (η) vermindert (vgl. Gleichung 9). Die Lernrate wird während des Hyperparametertrainings bestimmt (vgl. Goodfellow et al., 2018, S. 93). Eine zu hohe Lernrate führt zu einem Überspringen des Minimums. Wird sie zu niedrig angesetzt, dann kann eine erhöhte Rechendauer dazu führen, dass das Minimum nicht erreicht wird. Die im Lernalgorithmus genutzten Gewichte w_{ij} in der Schicht \mathbf{h}^s in $\epsilon + 1$ bestimmen sich dann aus der Anpassung des Gewichts w_{ij}^s in ϵ durch den mittels der Lernrate reduzierten Gradienten.

$$(w_{ij}^s)^{\epsilon+1} = (w_{ij}^s)^\epsilon - \eta \cdot \frac{\partial L}{\partial (w_{ij}^s)^\epsilon} \quad (9)$$

Das Durchführen des Gradientenverfahrens mit einem vollständigen umfangreichen Trainingsdatensatz führt zu einem hohen Rechenaufwand (vgl. Goodfellow et al., 2018, S. 167). Dem entgegen steht das stochastische Gradientenverfahren. In diesem Verfahren wird durch Zufallsziehung in jeder Epoche ϵ des Lernalgorithmus eine Beobachtung ausgewählt und auf Basis des berechneten Fehlers der Gradient bestimmt und die Gewichte angepasst (vgl. Goodfellow et al., 2018, S. 167). Die zufällige Wahl einer Beobachtung führt wegen des Wegfalls der Summierung über alle Beobachtungen in der Verlustfunktion zu einer höheren Iteration des Lernalgorithmus. Dieser so approximativ bestimmte Gradient weist ein hohes Rauschen und somit auch eine hohe Varianz auf (vgl. Goodfellow et al., 2018, S. 311). Um das Rauschen zu reduzieren, ist ein häufig verwendetes Verfahren das *minibatch*-Gradientenverfahren (vgl. Goodfellow et al., 2018, S. 168). Anstatt der zufälligen Ziehung einer Beobachtung werden in jeder Epoche mehrere Beobachtungen gezogen. Eine Epoche wird damit beim Training in mehrere Schritte (*Batches*) unterteilt. Das Rauschen sinkt bei gleichzeitig gesteigerter Konvergenz an das Minimum der Verlustfunktion (vgl. Goodfellow et al., 2018, S. 309). Der Zufallsprozess reduziert darüber hinaus die Wahrscheinlichkeit einer Überanpassung an den Trainingsdatensatz. In den folgenden Erläuterungen zur Anwendung des Gradientenverfahrens über mehrere Iterationen wird zur Vereinfachung der Darstellung angenommen, dass keine *Batches* gebildet werden, sondern weiter eine Beobachtung das Netz durchwandert.

Aufgrund der Verzweigungen und der vielen möglichen Kombinationen innerhalb künstlicher neuronaler Netze ist es nicht möglich, die Anpassung der Gewichte in allen

Schichten auf Basis des Fehlerterms der letzten Schicht zu berechnen. Die Information des Fehlerterms wandert von der letzten Schicht rückwärts durch alle Schichten (*Backpropagation*) (vgl. Goodfellow et al., 2018, S. 225). Die Anpassung jeder Schicht \mathbf{h}^s hängt von der Anpassung der vorherigen Schicht \mathbf{h}^{s+1} ab. Die Information, die beim Vorwärtstlauf des Netzwerkes von Schicht zu Schicht weitergegeben wird, ist das Ergebnis einer Schicht \mathbf{h}^s . Bei der *Backpropagation* wird betrachtet, wie sich das Ergebnis einer Schicht \mathbf{h}^s verändern muss, um die Verlustfunktion zu reduzieren. Diese Veränderung wird durch ein Anpassen der Gewichte in \mathbf{W}^s erreicht. Aus der Gleichung 9 geht hervor, dass es für das Anpassen der Gewichte notwendig ist, den Gradienten $\frac{\partial L}{\partial w_{ij}^s}$ zu bestimmen (vgl. Goodfellow et al., 2018, S. 228). In der ersten verdeckten Schicht ist $\mathbf{h}^{s-1} = \mathbf{x}$. Der Gradient wird durch den Effekt einer Veränderung in der Ausgabe jedes Neurons auf die Verlustfunktion und den Effekt auf die Ausgabe der Neuronen durch eine Veränderung der Gewichte beeinflusst:

$$\frac{\partial L}{\partial w_{ij}^s} = \frac{\partial L}{\partial a_j^s} \frac{\partial a_j^s}{\partial w_{ij}^s} \quad (10)$$

Da in diesem Abschnitt zunächst keine Verlustfunktion definiert werden soll,¹² kann der erste Teil von Gleichung 10 als Fehlerterm δ definiert werden (vgl. Rojas, 1996, S. 164). Eine Veränderung der Ausgabe beziehungsweise eines Ergebnisses verändert die Richtung und Stärke der Reaktion im Gradientenverfahren um:

$$\delta_j^s = \frac{\partial L}{\partial a_j^s} \quad (11)$$

Der zweite Term in Gleichung 10 wird durch Einsetzen von a_j^s aus Gleichung 8 gelöst.

$$\frac{\partial a_j^s}{\partial w_{ij}^s} = \frac{\partial}{\partial w_{ij}^s} \left(\sum_{i=0}^{r^s} w_{ij}^s h_j^{s-1} \right) = h_j^{s-1} \quad (12)$$

Durch das Zusammenführen von Gleichung 10, 11 und 12 wird deutlich, dass die Anpassung der Gewichte während des Lernalgorithmus durch die Kenntnis des Fehlers aus der vorherigen Schätzung bestimmt wird (vgl. Rojas, 1996, S. 168). Die anzupassenden Gewichte der Ausgabe aus \mathbf{h}^{s-1} verbinden Schicht \mathbf{h}^s mit \mathbf{h}^{s-1} und der

¹²Die Definition der in dieser Arbeit verwendeten Verlustfunktion erfolgt in Kapitel 4.1.2.

Fehlerterm δ^s bezieht sich somit auf \mathbf{h}^{s+1} (vgl. Gleichung 13). Durch den bekannten Fehler in \mathbf{h}^{s+1} können die Gewichte in \mathbf{h}^s angepasst werden, sodass der Fehler der Schätzung $a_i^s = \sum_{j=0}^{r^s} w_{ij}^s h_j^{s-1}$ geringer wird.

$$\frac{\partial L}{\partial w_{ij}^s} = \delta_j^s h_i^{s-1} \quad (13)$$

Gleichung 9 kann als Veränderung in w_{ij} betrachtet werden und durch gleichzeitiges Einsetzen von $\frac{\partial L}{\partial w_{ij}^s}$ aus Gleichung 13 ergibt sich Gleichung 14. Es sei angenommen $\delta_j^s h_i^{s-1} > 0$ führt zu einer Erhöhung von w_{ij} und zu einer Erhöhung der eingehenden Werte von h_i^{s-1} (vgl. Gleichung 8). Daraus folgt ein Anstieg des Fehlerterms δ_j^s . Eine negative Anpassung wird dem Gewicht hinzugefügt und die Verlustfunktion sinkt in der nächsten Iteration, da die Veränderung $\Delta w_{ij}^s > 0$ ist.

$$\Delta w_{ij}^s = -\eta \delta_j^s h_i^{s-1} \quad (14)$$

4.1.2 Ausgabeschicht und Softmatrix

Im vorherigen Kapitel wurde erläutert wie künstliche neuronale Netze auf Grundlage des Fehlers der Schätzung über mehrere Iterationen lernen. In diesem Kapitel wird beschrieben, wie am Ende jeder Iteration $\hat{\mathbf{y}}$ geschätzt und aus der Schätzung der Verlust berechnet wird.

In dieser Arbeit sollen durch die künstlichen neuronalen Netze binäre Klassifizierungsprobleme gelöst werden. Das heißt, dass \mathbf{y} einen Vektor mit ganzen Zahlen als Ausprägungen darstellt (vgl. Goodfellow et al., 2018, S. 203). Im binären Fall wird für jede Beobachtung $\hat{y}_m = P(y = 1 | \mathbf{z})$ bestimmt (vgl. Goodfellow et al., 2018, S. 205). Aus der Perspektive des neuronalen Netzes ist jede Wahrscheinlichkeit ein Neuron. Die Anzahl der Beobachtungen ist $m = 1, \dots, M$. Das Ziel der Schätzung ist eine diskrete Wahrscheinlichkeitsverteilung, wodurch die Summe der geschätzten Wahrscheinlichkeiten eins ergeben soll. Um diese Darstellungsform zu erreichen, wird die Softmax-Funktion genutzt (Gleichung 15) (vgl. Goodfellow et al., 2018, S. 204). Die Eingabe in die Softmax-Funktion ist $\mathbf{z} = (\mathbf{W}^s)^T \mathbf{h}^s + \mathbf{b}^s$ (vgl. Goodfellow et al., 2018, S. 204). \mathbf{z} entspricht damit \mathbf{a}^s , ist jedoch explizit die Ausgabe aus der letzten verdeckten Schicht. \mathbf{W}^s ist die Gewichtsmatrix, \mathbf{h}^s die berechneten Werte in der vorherigen verdeckten Schicht und \mathbf{b}^s die Verzerrung. Jede geschätzte Wahrschein-

lichkeit z_i wird in das Verhältnis zu den Schätzungen aller Klassen J gesetzt. Durch das Anwenden von $\exp(z_i)$ werden alle Werte positiv und das Dividieren durch $\sum_j^J \exp(z_j)$ führt zu Normalisierung. Die Ausgabe aus der Softmax-Funktion ist für jede Beobachtung eine geschätzte Wahrscheinlichkeitsverteilung, den Kategorien anzugehören.

$$\text{softmax}(\mathbf{z})_i = \frac{\exp(z_i)}{\sum_{j=1}^J \exp(z_j)} \quad \text{fuer } i = 1, \dots r \quad (15)$$

Als Verlustfunktion wird in dieser Arbeit die *Cross Entropy*-Funktion $L(\hat{\mathbf{y}}, \mathbf{y}) = -\sum_m^M \log P(\mathbf{y}^m | \mathbf{z}^m)$ genutzt (vgl. Goodfellow et al., 2018, S. 146-147). Die *Cross Entropy*-Funktion ist die negative Log-Likelihood Funktion (vgl. Kapitel 4.2.2) (vgl. Goldberg, 2017, S. 28). Die zu berechnenden Wahrscheinlichkeiten $P(\mathbf{y}^m | \mathbf{z}^m)$ sind die Ausgabe der Softmax-Funktion. Wird die Softmax-Funktion in Kombination mit der *Cross Entropy*-Funktion oder Log-Likelihood Funktion genutzt, ist ausgeschlossen, dass $\hat{y}_i = 0$ wird und der Gradient nicht mehr berechnet werden kann (vgl. Goodfellow et al., 2018, S. 205). Entsprechend wird die Softmax-Funktion häufig in Kombination mit der Log-Likelihood Funktion verwendet und so die Softmax-Funktion logarithmiert (Gleichung 16) (vgl. Goodfellow et al., 2018, S. 205). Für eine Beobachtung ist $\log P(\mathbf{y}^1 | \mathbf{z}^1)$:

$$L(\hat{y}_i, y_i) = \log \text{softmax}(\mathbf{z})_i = z_i - \log \sum_{j=1}^J \exp(z_j) \quad (16)$$

$L(\hat{y}_i, y_i)$ wird maximal, wenn die Wahrscheinlichkeit für eine Ausprägung eins ist. Die Anwendung von $L(\hat{y}_i, y_i)$ führt dazu, dass die geringste Schätzung mit der höchsten Wahrscheinlichkeit am stärksten bestraft wird (vgl. Goodfellow et al., 2018, S. 204). Die Schätzungen von z_i , die geringer als der maximale Wert in \mathbf{z} sind, verlieren durch das Nutzen der Exponentialfunktion in der Minimierung an Bedeutung.

4.2 Methodik des *Topic-Modelings*

Während LDA über die Wahrscheinlichkeitsverteilung der Wörter in den Tweets die Themen bestimmt, ermittelt *Top2Vec* die Themen anhand des semantischen Abstands der Tweets im Raum. *Top2Vec* kann in vier Schritte unterteilt werden. Zuerst wird ein semantischer Raum berechnet, in dem Wörter und Tweets in ein num-

merisches Verhältnis zueinander gesetzt werden. Ein Raum mit hoher Dimensionalität führt zu einer steigenden Anforderung an die Rechnerkapazitäten bei dem Clustern der Tweets. Um die Berechnung der Cluster zu ermöglichen, wird die Dimension des Raumes mit *Uniform Manifold Approximation and Projection* (UMAP) (McInnes et al., 2020) auf fünf reduziert. Die so reduzierten Tweets werden mit *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN) geclustert (Campello et al., 2013). Der abschließende Schritt ist die Bestimmung des Wortvektors, welcher als Durchschnitt der sich im Cluster befindenden Tweets verstanden werden kann. Dieser Vektor ist somit eine Annäherung an die Beschreibung des Themas.

4.2.1 Bestimmung des semantischen Raums

Für die Berechnung des semantischen Raums wird der *Doc2vec*-Algorithmus genutzt (Le & Mikolov, 2014). *Doc2vec* stellt eine Erweiterung des Skip-Gram Modells aus *Word2vec* dar (vgl. Mikolov et al., 2013; Le & Mikolov, 2014, S. 1). Bei *Word2vec* und weiteren ähnlichen Algorithmen wie *GloVe* und *FastText* ist bisher unklar, weshalb eine gute Worddarstellung erreicht wird (vgl. Goldberg & Levy, 2014, S. 5). Ein möglicher Erklärungsansatz ist die distributionelle Hypothese, wonach Wörter, die im ähnlichen Kontext genutzt werden, zu einer ähnlichen Bedeutung tendieren (vgl. Harris, 1954, S. 152).

Das Skip-Gram Modell betrachtet jedes Wort ω im Datensatz und schätzt die Wahrscheinlichkeiten $p(\mathbf{c}|\omega)$, dass die Kontextwörter \mathbf{c} in der Umgebung liegen (vgl. Gleichung 17) (vgl. Goldberg & Levy, 2014, S. 1). Sowohl ω als auch \mathbf{c} sind Vektoren mit einer Länge gemäß der Anzahl aller möglichen Wörter im Textdatensatz. \mathbf{T} stellt sämtliche Wörter zu Kontextwörter Kombinationen dar. Die Anzahl der Kontextwörter, die betrachtet werden, ist ein Hyperparameter und abhängig vom Datensatz.

$$\arg \max_{(\omega, \mathbf{c}) \in \mathbf{T}} \prod p(\mathbf{c}|\omega) \quad (17)$$

Für die Berechnung der Wahrscheinlichkeiten wird ein einfaches neuronales Netz genutzt, welches eine verdeckte Schicht enthält. In der verdeckten Schicht werden die Eingabedaten zusammengefasst, um eine gewünschte *Embedding*-Dimension zu erhalten. Die Berechnung findet innerhalb einer Softmax-Funktion statt (vgl. Gold-

berg & Levy, 2014, S. 2). Zur Vereinfachung wird im Folgenden angenommen, dass keine verdeckte Schicht existiert. Als Verlustfunktion wird die Log-Likelihood Funktion genutzt (vgl. Goldberg & Levy, 2014, S. 2). Als Eingabedaten wird ein Textkorpus verwendet, bei dem die Wörter *one-hot*-kodiert werden (vgl. Goldberg, 2017, S. 89-90). Die Ausgabe des Algorithmus ist ein Vektor, der für jedes Wort im Datensatz die Wahrscheinlichkeit enthält, dass ein anderes Wort im Datensatz das nächste Wort ist. Das Ziel des Algorithmus ist es, diese Wahrscheinlichkeitsverteilung zu optimieren (vgl. Goldberg & Levy, 2014, S. 2). Die Einordnung des Algorithmus in *unsupervised* oder *supervised* ist nicht eindeutig. Zwar enthält der Datensatz keine Klassifizierung, jedoch wird ein *Supervised*-Ansatz verwendet. Um zu überprüfen, ob das geschätzte Wort richtig ist, wird der *one-hot* des Wortes genutzt. Für jedes Wort wird versucht, die Eins im Vektor zu schätzen.

Die Anwendung der Softmax-Funktion (Gleichung 15) auf die Problemstellung aus Gleichung 17 und anschließendes Logarithmieren führen zur Gleichung 18 (vgl. Goldberg & Levy, 2014, S. 2). \mathbf{c} und ω sind die Vektordarstellungen des betrachteten Kontextwortes und des betrachteten Wortes. \mathbf{c}' ist der Wortvektor für jedes mögliche Kontextwort. Das Skip-Gram Modell maximiert die Schätzung der Wahrscheinlichkeit, dass ein Wort im Kontext eines anderen erscheint.

$$\arg \max_{\Theta} \sum_{(\omega, \mathbf{c}) \in \mathbf{T}} \log p(\mathbf{c} | \omega) = \sum_{(\omega, \mathbf{c}) \in \mathbf{T}} (\log \exp(\mathbf{c} \cdot \omega) - \log \sum_{\mathbf{c}'} \exp(\mathbf{c}' \cdot \omega)) \quad (18)$$

Der auf dem Skip-Gram Modell aufbauende Algorithmus von *Doc2vec* ist *Distributed Bag of Words version of Paragraph Vector* (PV-DBOW) (vgl. Le & Mikolov, 2014, S. 4). Jedes Dokument besteht aus verschiedenen Wörtern. Die semantische Position im Raum jedes Dokuments kann als eine Kombination der Wörter innerhalb des Dokuments betrachtet werden (vgl. Le & Mikolov, 2014, S. 4). Darauf aufbauend wird bei der Bestimmung des Dokumentvektors mit PV-DBOW ähnlich vorgegangen wie im Skip-Gram Modell. Jedes Dokument wird in eine einzigartige *one-hot*-Kodierung umgewandelt. Mit diesen Vektoren wird mit demselben Aufbau wie beim Skip-Gram Modell in *Word2vec* ein zufällig ausgewähltes Wort aus dem Dokument geschätzt (vgl. Le & Mikolov, 2014, S. 4). Die Wörter sind *one-hot* kodierte Vektoren. Das Ziel des Lernalgorithmus ist es, dass die zufällig ausgewählten Wörter mit einer hohen Wahrscheinlichkeit richtig geschätzt werden. Die entstehende Wahrscheinlichkeitsverteilung ist die Vektordarstellung der Dokumente (vgl. Le &

Mikolov, 2014, S. 4). Zum Erreichen eines einheitlichen semantischen Raums ist es notwendig, Wortvektoren und Dokumentvektoren parallel zu trainieren (vgl. Angelov, 2020, S. 5). Hierfür werden die Dokumentvektoren als eine Art synthetisches Wort angesehen und innerhalb des Skip-Gram Modells trainiert.

4.2.2 *Uniform Manifold Approximation and Projection*

Wort- und Dokument- *Embeddings* werden häufig mit einer hohen Anzahl an Dimensionen trainiert. Die Anzahl der Dimension, welche genutzt wird, ist von den verwendeten Daten und der Fragestellung abhängig. Eine häufig genutzte Anzahl ist eine Dimension von 300. Innerhalb von 300 Dimensionen Cluster zu bilden ist rechenintensiv und ab einer gewissen Datenmenge nicht möglich. Um das Clustern mit HDBSCAN zu ermöglichen, wird beim *Topic-Modeling* der COVID-19 Tweets UMAP verwendet. Andere für Dimensionsreduzierung genutzte Ansätze sind *Principal Component Analysis (PCA)* und *t-distributed stochastic neighbor embedding (t-SNE)*. t-SNE und UMAP sind nicht lineare Verfahren, wohingegen die PCA auf der Zusammenfassung von Dimensionen als Linearkombinationen beruht. Mit steigender Komplexität der Daten sinkt die Genauigkeit der PCA. t-SNE ist auf die Reduzierung auf zwei bis drei Dimensionen begrenzt (vgl. McInnes et al., 2020, S. 2). Dimensionsreduzierung mit UMAP hingegen ist unabhängig von der Anzahl der Dimensionen, auf die zu reduzierenden ist. Somit können die Dimensionen der Daten vor dem Clustern auf die durch die Hardware händelbare Anzahl reduziert und so mehr Informationen für das Erstellen von Clustern genutzt werden. Ein weiterer Unterschied zwischen t-SNE und UMAP besteht in der Extraktion von globalen Strukturen in den Daten. Während t-SNE Zusammenhänge von benachbarten Daten betrachtet und nur lokale Zusammenhänge erkennt, ist UMAP durch das Nutzen von Mannigfaltigkeit in Kombination mit lokalen *Fuzzy simplizial* Mengen in der Lage, topologische Darstellungen in niedriger Dimension von hochdimensionen Daten zu erstellen (vgl. McInnes et al., 2020, S. 4). In manchen Räumen mit bestimmten topologischen Eigenschaften können lokale und globale Strukturen aufgefasst werden. Ein wichtiges Beispiel dafür ist eine Mannigfaltigkeit (vgl. Becht et al., 2019, S. 1-2). Ein topologischer Raum ist ein Raum von Menge, indem Teilmengen als offen ausgezeichnet werden, zum Beispiel sind in dem topologischen Raum der reellen Zahlen alle offenen Intervalle eine offene Teilmenge (vgl. Lee, 2011, S. 19-20). Die wissenschaftliche Diskussion, ob UMAP die globale Struktur aufgrund dieses Vorgehens, oder aufgrund der genutzten Startbedingungen klarer extrahiert ist nicht abgeschlossen (vgl. Kobak & Linderman, 2019,

S. 2).

Eine Mannigfaltigkeit ist ein topologischer Raum in dem sämtliche Punkte eine euklidische Umgebung haben und ein lokaler Abstandsbegriff definierbar ist (vgl. Goodfellow et al., 2018, S. 177). Eine Bewegung über diesen Bereich ist möglich. Die lokale Darstellung der Mannigfaltigkeit ist somit ein euklidischer Raum (vgl. Goodfellow et al., 2018, S. 177). Da bei Mannigfaltigkeiten Verformungen und Dehnungen keinen Einfluss auf die Eigenschaften haben, betrachten wir die Erde als Kugel. In Darstellungen auf Karten oder in der menschlichen Wahrnehmung ist die Erde jedoch im zweidimensionalen Raum (vgl. Goodfellow et al., 2018, S. 179). Eine Karte ist die Darstellung der Mannigfaltigkeit eines Globus im euklidischen Raum. Dies folgt der Mannigfaltigkeits Hypothese, dass hochdimensionale reale Daten eine niedrigdimensionale Darstellung haben (vgl. Goodfellow et al., 2018, S. 179). Der topologische Raum des Globus lässt sich verformen, die Eigenschaften des Raumes bleiben jedoch unverändert (vgl. Goodfellow et al., 2018, S. 177). Für eine zielführende Dimensionreduzierung ist es notwendig, dass die Daten nicht zufällig gleichförmig über die Mannigfaltigkeit verteilt sind. Wäre dies der Fall, könnte zwar durch Verformung eine Mannigfaltigkeit in niedriger Dimension gebildet werden, die zusammengefügten Dimensionen wären jedoch nicht zielführend. Daten wie Bilder oder Texte weisen eine stark zentrierte Verteilungen auf (vgl. Goodfellow et al., 2018, S. 179). Ein vollständig per Zufall aus Buchstaben zusammengesetzter Text ist gleichförmig in der Mannigfaltigkeit verteilt, praktisch allerdings nicht relevant. Die Sequenzen einer natürlichen Sprachfolge stellen einen kleinen Teil der möglichen Buchstabenkombinationen dar und sind somit mehrdimensional zentriert (vgl. Goodfellow et al., 2018, S. 179).

Auf die Anwendung innerhalb von *Top2vec* bezogen ist die Eingabe in UMAP ein Datensatz von Dokumenten in mehrdimensionaler Darstellung $\mathbb{R}^{M \times n}$. Hierbei ist M die Anzahl der Beobachtungen im Datensatz und n die Dimension der Dokumentvektoren. Mit UMAP wird die Anzahl der Dimensionen auf fünf reduziert. Neben der Information, die in jedem Dokumentvektor enthalten ist, betrachtet UMAP für die Reduzierung auch die Ausprägungen der Dokumentvektoren in der Umgebung und den Abstand zu diesen. Diese Datenpunkte werden über Simplizialkomplexe verbunden. Simplizialkomplexe sind die Verbindungen der Daten über mehrere Dimensionen. Ein Simplizialkomplex in einer Dimension verbindet zwei Punkte durch eine Linie. In zwei Dimensionen ist ein Simplizialkomplex eine Verbindung von drei Punkten (vgl. McInnes, 2018). Dies kann für höhere Dimensionen fortgesetzt werden. Ob sich aus zwei Punkten ein Simplizialkomplex bildet, ist abhängig von dem Abstand der Punkte

und im Algorithmus ein Hyperparameter (vgl. McInnes, 2018). Die Simplizialkomplexe können insgesamt die topologische Struktur approximieren. Der UMAP Algorithmus bestimmt diese Simplizialkomplexe im höheren dimensionalen topologischen Raum und berechnet aus diesen eine niedrigdimensionale Darstellung des höherdimensionalen topologischen Raumes (vgl. McInnes, 2018; McInnes et al., 2020, S. 11). Die Menge aller Simplizialkomplexe ist \mathbf{E} , wobei ein Simplizialkomplex e ist.

In einem realen Datensatz sind die Daten nicht gleichmäßig über die Mannigfaltigkeit verteilt. Das führt dazu, dass beispielsweise in einer euklidischen Betrachtung nicht für sämtliche Daten benachbarte Datenpunkte ermittelt werden können (vgl. Abbildung 2 (1)) (vgl. McInnes, 2018). Dies gilt auch für Dimensionen größer zwei. Die Abstände in der Mannigfaltigkeit können in einem euklidischen Raum durch Kreise um die Punkte dargestellt werden (vgl. McInnes, 2018). In höheren Dimensionen ist der Kreis eine Kugel. Simplizialkomplexe bilden sich innerhalb dieser Kugel. Der Radius der Kugel um die Punkte und somit die Art wie sich die topologische Struktur zusammensetzt, ist schwierig zu bestimmen (vgl. McInnes, 2018). Ist der Radius zu klein, zerfällt der Simplizialkomplex. Bei einem zu großen Radius wird die Struktur dagegen unzureichend erkannt. Bei einer gleichmäßigen Verteilung der Daten über die Mannigfaltigkeit ist der zu bestimmende Radius eindeutig und die Struktur der Mannigfaltigkeit kann sehr genau dargestellt werden (vgl. McInnes, 2018). Um dieses Problem zu umgehen, wird eine Gleichverteilung der Daten angenommen (vgl. McInnes et al., 2020, S. 4). Aus dieser Annahme in Kombination mit den nicht gleichverteilten realen Daten folgt, dass der Begriff des Abstandes über die Mannigfaltigkeit variiert (vgl. McInnes, 2018). Es kommt zu einer Ausdehnung oder Verengung des Raumes, abhängig von der Dichte der Daten (vgl. McInnes, 2018). Durch die Anwendung von Riemannscher Geometrie kann unter der Annahme der Gleichverteilung für jeden Punkt ein Abstandsbegriff definiert werden (vgl. McInnes, 2018). Der Radius jeder Kugel um die Punkte reicht bis zu den nächsten k Nachbarn (vgl. Abbildung 2 (2)). Aus dieser theoretischen Erklärung wird deutlich, dass die Approximation der topologischen Struktur der Mannigfaltigkeit durch Simplizialkomplexe auch als *k-neighbour graph* verstanden werden kann, mit dem Unterschied, dass jeder Punkt einen anderen Begriff von Abstand aufweist (vgl. McInnes, 2018; McInnes et al., 2020, S. 14). Die Punkte besitzen einen lokalen metrischen Raum und die Entfernung zwischen Punkten sind damit sinnvoll bestimmbar (vgl. McInnes, 2018).

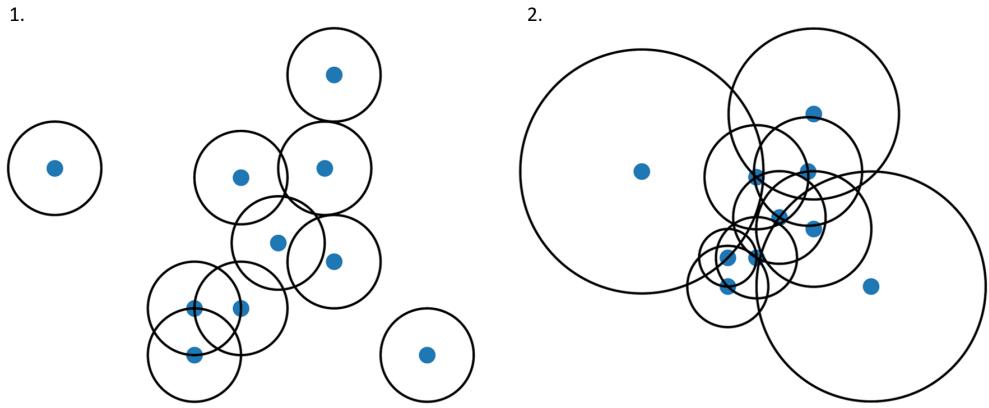


Abbildung 2: Anpassung der Distanz abhängig von der Dichte (eigene Darstellung)

Im abschließenden Schritt bei der Bestimmung des Raumes beziehungsweise Graphen wird davon ausgegangen, dass die betrachteten Räume eine Fuzzy Topologie aufweisen (vgl. McInnes, 2018). Eine Fuzzy Menge oder auch unscharfe Menge ist eine Menge, in der die Zugehörigkeit zu dieser unscharf ist. Auf die Abbildung der topologischen Struktur übertragen bedeutet dies, dass die Verbindungen der Punkte innerhalb der Kugeln eine Wahrscheinlichkeit erhalten innerhalb der Kugel zu liegen (vgl. McInnes, 2018). Die Wahrscheinlichkeit sinkt, wenn die Punkte am Rand der Kugel liegen. Der *k-neighbour graph* verändert sich zu einem gewichteten *k-neighbour graph*. Die topologische Beschreibung der Eingangsdaten ist eine Fuzzy simpliziale Menge. Sie enthält also sämtliche berechneten Simplizialkomplexe inklusive der Wahrscheinlichkeit das eine Verbindung der Knoten existiert.

Die niedrigdimensionale Darstellung der Eingangsdaten $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_M\}$ ist $\mathbf{X}_{re} = \{\mathbf{x}_{re_1}, \dots, \mathbf{x}_{re_m}\}$, wobei $\mathbf{x}_{re_m} \in \mathbb{R}^{n_{re}}$ und $n_{re} < n$ ist (vgl. McInnes et al., 2020, S. 11). Anders als bei der Darstellung als Fuzzy simpliziale Menge von \mathbf{X} , wird die Darstellung der Fuzzy simpliziale Menge von \mathbf{X}_{re} allein auf Basis der lokalen Information berechnet. Diese Darstellung wird ohne Information über die globale Struktur der Daten gewonnen (vgl. McInnes et al., 2020, S. 11). Das bedeutet, dass die einzelne berechneten Radien der Kreise nicht genutzt werden. Die Informationen über die globalen Strukturen, die aus den Eingangsdaten gewonnen werden und sich in der Fuzzy simplizialen Menge von \mathbf{X} befinden, werden durch einen Optimierungsprozess in die von \mathbf{X}_{re} integriert. Dies wird erreicht, indem die Darstellung als Fuzzy simpliziale Menge von \mathbf{X} mit der von \mathbf{X}_{re} verglichen wird (vgl. McInnes et al., 2020, S. 12). Für diesen Vergleich werden sämtliche Simplizialkomplexe in einer Dimension genutzt, da eine höhere Dimension zwar zu genaueren Ergebnissen führt, jedoch rechenaufwendig ist (vgl. McInnes, 2018). In einer Vorstellung eines *k-neighbour*

graph werden alle Gewichte zu den k Nachbarn eines jeden Punktes betrachtet. Die Nachbarn sind untereinander nicht verbunden. Die Punkte des Simplizialkomplexes sind entweder verbunden oder nicht und die Gewichte geben die Wahrscheinlichkeit an, ob die zwei Punkte verbunden sind. Somit sind diese Verbindungen Bernoulli Variablen (vgl. McInnes, 2018). Die Gewichte jedes Simplizialkomplexes werden in der höherdimensionalen Darstellung mit $w_h(e)$ und in der niedrigdimensionalen Darstellung mit $w_l(e)$ bestimmt. Der Abstand zwischen diesen Gewichten wird mit der *Cross Entropy*-Funktion verglichen und die Gewichte $w_l(e)$ werden über ϵ mit *Stochastic Gradient Descent* angepasst (vgl. McInnes, 2018; McInnes et al., 2020, S. 12). Das Zusammenführen der Bernoulli Verteilung und der *Cross Entropy*-Funktion führt zu Gleichung 19.

Wird die niedrigdimensionale Darstellung als Graph betrachtet, führt die Optimierung von Gleichung 19 zu einem Zusammenrücken oder Auseinanderdriften der Knoten (vgl. McInnes, 2018). Während der Optimierung nähern sich die Gewichte der niedrigdimensionalen Darstellung $w_l(e)$ an $w_h(e)$ an. Ist $w_l(e) > w_h(e)$, führt das zu einem negativen ersten Term und gleichzeitig ist der zweite Term positiv. Einem Absinken des Gewichtes folgt eine Annäherung beider Terme nach null. Im Graphen führt das zu einer steigenden Entfernung der Knoten voneinander (vgl. McInnes, 2018). Mit $w_l(e) < w_h(e)$ ist der zweite Term negativ und im Umkehrschluss der erste Term positiv. Ein Anheben des Gewichtes führt somit zu einem Zusammenrücken der Knoten. Durch die Anpassung der Gewichte während der Optimierung werden die globalen Informationen, die in den Gewichten der höherdimensionalen topologischen Beschreibung in der Fuzzy simplizialen Menge gespeichert sind auf die Gewichte in der niedrigdimensionalen Darstellung übertragen (vgl. McInnes, 2018; McInnes et al., 2020, S. 12).

$$\sum_{e \in E} w_h(e) \log\left(\frac{w_h(e)}{w_l(e)} + (1 - w_h(e)) \log\left(\frac{1 - w_h(e)}{1 - w_l(e)}\right)\right) \quad (19)$$

4.2.3 Hierarchical Density-Based Spatial Clustering of Applications with Noise

Der anschließende Schritt im *Top2vec*-Algorithmus besteht darin, Cluster innerhalb der auf fünf Dimensionen reduzierten Daten zu finden. Übertragen auf die Aufgabe von *Top2vec* sollen die Tweets gefunden werden, die einem gemeinsamen Thema angehören. Die Anzahl der Themen innerhalb sämtlicher Tweets ist nicht bekannt.

nt. Auch ein manuelles Festlegen der Anzahl der Cluster wie bei *k-nearest neighbours* wird durch die fünf dimensionalen Eingangsdaten erschwert. Durch die Art der genutzten Daten ergeben sich Strukturen innerhalb der Themen. Ein Thema kann ein Unterthema eines anderen sein. Beispielsweise können Themen wie „Soforthilfe“ und „Schließung von Restaurants“ auch Unterthemen des Clusters „Wirtschaft“ sein. Ebenso ist unklar, ob jeder Tweet einem Thema zuordenbar ist. Ein Tweet kann als Rauschen aufgefasst werden, wenn er keinem Thema zuzuordnen ist. Besonders Tweets, die zwar in der Datenbeschaffung den Kriterien entsprechen, aber davon abgesehen keine Inhalte enthalten, werden keinem Thema zugeordnet. Nicht erkanntes Rauschen kann zwei Cluster fälschlicherweise miteinander verbinden. HDBSCAN ist in der Lage, die Anzahl der Cluster selbstständig zu bestimmen und neben einer genauen Zuordnung von Tweets zu Themen auch Rauschen zu erkennen (vgl. McInnes & Healy, 2017, S. 2). Die Verzerrung der Themen durch mögliches Rauschen sinkt. HDBSCAN stellt die Kombination aus einem verteilungsbasierten und einem hierarchischen Clusteralgorithmus dar (vgl. McInnes & Healy, 2017, S. 7-8).

Das Unterteilen in Cluster basiert auf *Single-Linkage-Clustering*. *Single-Linkage-Clustering* beruht auf aufeinander aufbauenden Prozessen (vgl. Johnson, 1967, S. 248). Im ersten Schritt wird eine Distanzmatrix berechnet, in der die Distanz jedes Punktes zu sämtlichen anderen Punkten definiert ist (vgl. Johnson, 1967, S. 248). Die zwei Punkte, die den niedrigsten Distanzwert aufweisen, werden als ein Cluster zusammengefasst. Anschließend wird erneut die Distanzmatrix berechnet, wobei die Distanz zu dem Cluster der minimale Abstand ist. In diesem Schritt werden weitere Punkte zusammengefasst oder Punkte dem Cluster hinzugefügt. Der Prozess wird fortgeführt, bis sämtliche Punkte in einem Cluster zusammengefasst sind (vgl. Johnson, 1967, S. 248). Die dimensionsreduzierte Ausgabe aus Kapitel 4.2.2 \mathbf{X}_{re} ist die Eingabe im HDBSCAN.

Durch das Nutzen von einzelnen Punkten zum Bestimmen von Clustern ist das *Single-Linkage-Clustering* anfällig für Rauschen. Um die Robustheit des Clusters gegenüber dem Rauschen zu erhöhen, wird im HDBSCAN die Dichteverteilung der Daten betrachtet. Cluster können als Bereich mit hoher Dichte aufgefasst werden, Bereiche mit geringer Dichte hingegen als Rauschen (vgl. McInnes & Healy, 2017, S. 7). Durch Definieren des Abstandes der Daten abhängig von der Dichte kann die Wahrscheinlichkeit reduziert werden, dass das Rauschen einem Cluster zugeordnet wird (vgl. McInnes & Healy, 2017, S. 7). Die Dichte wird durch die Distanz zu den k -ten nächsten Nachbarn bestimmt. Um die Dichte in die Distanzmessung zwischen den

Daten zu integrieren, wird die *Mutual Reachability Distance* ($dmreach$) genutzt (vgl. Gleichung 20) (vgl. McInnes & Healy, 2017, S. 7). $d(\mathbf{x}_{re_i}, \mathbf{x}_{re_j})$ ist der Abstand der Punkte i und j zueinander. $\kappa(\mathbf{x}_{re_i})$ ist die Distanz von i zu dem k -sten Nachbarn und $\kappa(\mathbf{x}_{re_j})$ der Abstand von j zu dem k -sten Nachbarn. Die neue Distanz zwischen den Punkten ist die höchste der drei Distanzen.

$$dmreach(\mathbf{x}_{re_i}, \mathbf{x}_{re_j}) = \max\{\kappa(\mathbf{x}_{re_i}), \kappa(\mathbf{x}_{re_j}), d(\mathbf{x}_{re_i}, \mathbf{x}_{re_j})\} \quad (20)$$

$\kappa(\mathbf{x}_{re_i})$ und $\kappa(\mathbf{x}_{re_j})$ sind der Radius von Kreisen um die Punkte i und j (vgl. Abbildung 3). In höheren Dimensionen ist der Kreis eine Kugel. Ist $\kappa(\mathbf{x}_{re_i}) > \kappa(\mathbf{x}_{re_j})$, sind die k nächsten Nachbarn um j in einem weiteren Umkreis um j gestreut als die k nächsten Nachbarn von i um i . Die Dichte der Punkte um j ist geringer als um i . Ist nun $d(\mathbf{x}_{re_i}, \mathbf{x}_{re_j}) < \kappa(\mathbf{x}_{re_j})$, wird die Distanz zwischen i und j auf $\kappa(\mathbf{x}_{re_j})$ gesteigert. Die Bereiche mit hoher und niedriger Dichte driften auseinander (vgl. McInnes et al., 2016).

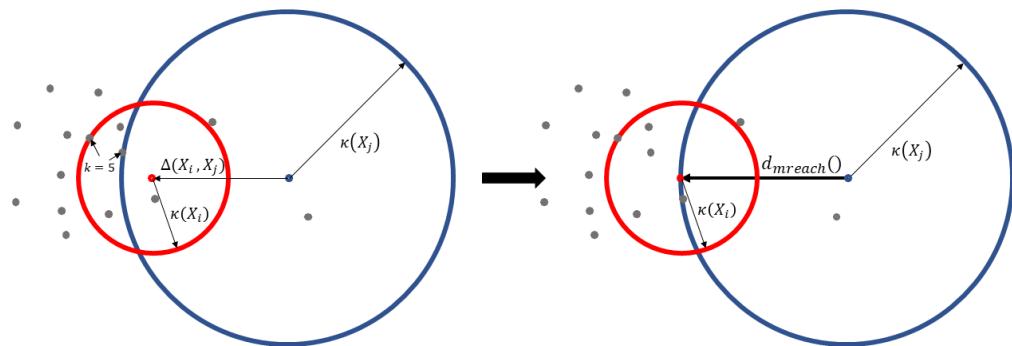


Abbildung 3: Anpassung der Distanz zwischen Datenpunkten durch $dmreach()$ (eigene Darstellung in Anlehnung an McInnes et al. 2016)

Die lokale Dichte eines Punktes ist die Inverse des Abstandes zwischen zwei Punkten $\lambda = \frac{1}{dmreach(\mathbf{x}_{re_i}, \mathbf{x}_{re_j})}$ (vgl. McInnes & Healy, 2017, S. 8). Steigt die Distanz zwischen zwei Punkten, sinkt λ . Zum Bestimmen des Clusters wird der durch *Single-Linkage-Clustering* berechnete Baum umgekehrt durchwandert (vgl. McInnes & Healy, 2017, S. 8). Das Aufteilen in zwei Cluster wird nicht als Teilung, sondern als Verlieren von Beobachtungen betrachtet. Mit steigender Distanz λ verlieren die Cluster an

Beobachtungen (vgl. McInnes & Healy, 2017, S. 8). Bei jedem Aufteilen der Daten wird kontrolliert, ob die entstehenden Cluster eine Mindestanzahl an Datenpunkten enthalten. Die Mindestanzahl ist ein zu definierender Hyperparameter. Entstehen beim Teilen des Clusters zwei neue Cluster, die die Mindestanzahl an Beobachtungen enthalten, wird das Cluster geteilt (vgl. McInnes & Healy, 2017, S. 8). Enthält eines der entstehenden Cluster weniger Beobachtungen als die Mindestanzahl, werden diese Punkte als Rauschen betrachtet und das andere entstehende Cluster wird als Fortsetzung des alten gedeutet (vgl. McInnes & Healy, 2017, S. 8).

4.2.4 Bestimmung der Themen

Nach der Bearbeitung mit UMAP und HDBSCAN befinden sich die Tweets in einem semantischen Raum in fünf Dimensionen. Jeder Tweet ist einem Cluster zugeordnet oder wurde als Rauschen identifiziert. Jedes durch HDBSCAN bestimmte Cluster stellt ein Thema dar (vgl. Angelov, 2020, S. 8). Durch die logischen und berechenbaren Verhältnisse der Dokument- und Wortvektoren zueinander (vgl. Kapitel 4.2.1) entspricht der Durchschnitt der Dokumentvektoren eines Clusters der durchschnittlichen Bedeutung des Clusters (vgl. Angelov, 2020, S. 8). Der abschließende Schritt von *Top2vec* ist es, die so berechneten Themenvektoren von der numerischen Form in ein Wort zu übertragen. Im semantischen Raum ist die Beschreibung des Themenvektors das Wort, welches die geringste Distanz zu dem Themenvektor aufweist (vgl. Angelov, 2020, S. 8). Für ein besseres Verständnis der Themen werden in der Analyse die k nächsten Wörter betrachtet. Die Wortvektoren, die ein Thema beschreiben, liegen in der Mitte eines Clusters. Dies bedeutet im Umkehrschluss, dass Wörter, die von mehreren Dokumenten in verschiedenen Clustern geteilt werden, zwischen den Clustern liegen (vgl. Angelov, 2020, S. 8). Stopwörter werden so nicht in die Bestimmung des Themas mit einbezogen.

4.2.5 Silhouettenkoeffizient

Anders als von Angelov (2020) vorgeschlagen, wird in dieser Arbeit zum Vergleich verschiedener *Topic-Modeling*-Modelle nicht die *Mutual Information* genutzt. Dieser Ansatz stellte sich für die Größe des verwendeten Datensatzes als nicht praktikabel heraus. Um verschiedene Modelle miteinander vergleichen zu können, wird der Silhouettenkoeffizient (vgl. Rousseeuw, 1987) verwendet. Für jede Beobachtung im Datensatz wird bestimmt, wie präzise die Zuordnung zum Cluster ist (vgl. Gleichung 21). Dieser Wert wird Silhouette s_C genannt. Es sei angenommen, dass eine Beobachtung

dem Cluster A angehört, dann ist die Silhouette die Differenz zwischen der Distanz der Beobachtung $d(i, B)$ zum Mittelpunkt des Clusters B und der Distanz $d(i, A)$ zum Mittelpunkt des Clusters A , geteilt durch die höhere Distanz der beiden Distanzen zu den Clustern (vgl. Rousseeuw, 1987, S. 56).

$$s(i) = \frac{d(i, B) - d(i, A)}{\max\{d(i, A), d(i, B)\}} \quad (21)$$

Diese Berechnung wird für alle Beobachtungen zu Cluster-Kombinationen durchgeführt und der Durchschnitt aus diesen wird ermittelt. Der Silhouettenkoeffizient liegt in dem Bereich $-1 \leq s_C \leq 1$ (vgl. Rousseeuw, 1987, S. 56). Je näher s_C an eins liegt, desto eindeutiger sind die Cluster bestimmt worden. Bei einem Wert um null liegt eine Vielzahl der Dokumente zwischen den Clustern und je kleiner s_C wird, desto unbestimmter werden die Cluster.

4.3 Methodik der Sentiment-Analyse

Der Bereich des *Deep-Learnings* umfasst viele verschiedene Arten von künstlichen neuronalen Netzen. Für Klassifizierungsprobleme mit Textdaten, wie eine Sentiment-Analyse, zeigen auf CNN (LeCun, 1989) und LSTM (vgl. Hochreiter & Schmidhuber, 1997) basierende Netze eine hohe Genauigkeit. Anders als in Kapitel 4.1 vorgestellt bestehen in diesen Modellen die verdeckten Schichten nicht aus einer Multiplikation von Eingabe und Gewichten, sondern aus verschiedenen gewichteten Datenbearbeitungsschritten. Im Folgenden werden die theoretischen Grundlagen von CNN und LSTM erläutert.

4.3.1 Convolutional Neural Networks

CNN sind an den menschlichen Umgang mit visuellen Informationen angelehnt (vgl. Goodfellow et al., 2018, S. 405). Um zu erkennen, welches Objekt betrachtet wird, konzentrieren Menschen sich auf die Merkmale, die nach bisherigen Erfahrungen ausschlaggebend für die Einordnung in eine Kategorie sind. Beispielsweise liegt bei dem Erkennen eines Hundes der Informationsgehalt in der Gesichts- und/oder Schnauzenform. Andere Merkmale wie die Farbe treten dabei in den Hintergrund und werden als Rauschen bei der Entscheidung herausgefiltert. An diesem Vorgehen bei der Objekterkennung setzt CNN an. Das Filtern der entscheidenden Merkmale während

des Trainingsprozesses und das gleichzeitige Reduzieren von Rauschen führt zu einer steigenden Genauigkeit der Schätzungen und senkt die Trainingszeit (vgl. Goodfellow et al., 2018., S. 374).

Die Eigenschaft, wichtige Informationen zu lokalisieren, führt dazu, dass CNN auch im NLP zum Einsatz kommen (vgl. Goldberg, 2017, S. 152). Diese Lokalisierung bezieht sich hier jedoch nicht auf die Position von Wörtern in Sätzen, sondern auf die Lokalisierung der Positionen im Satz mit dem höchsten Informationsgehalt unter Berücksichtigung des Wortes in einem Bedeutungscluster. Durch das Nutzen von Wortvektoren wird jedes Wort als Vektor mit n Dimensionen aufgefasst. Die Wortvektoren können vor Anwendung des Netzes trainiert werden oder eine Schicht innerhalb des Lernalgorithmus sein. Der Abstand von Wörtern im Raum sinkt mit zunehmender Ähnlichkeit der Bedeutung (vgl. Kapitel 4.2.1) (vgl. Goodfellow et al., 2018, S. 374). Durch das Anwenden des CNN werden die Bedeutungscluster im Raum entdeckt, die wichtige Informationen zur Klassifizierung liefern (vgl. Goldberg, 2017. S152). Dieses führt dazu, dass die konkret genutzten Wörter im Satz unerheblich werden und sich somit die Komplexität der Daten reduziert. Es ist nicht mehr relevant, welche Wort verwendet wurde, sondern die Kombination an Bedeutungsclustern, denen das Wort angehört, ist wichtig. Beim Anwenden des trainierten Netzwerks können auch unbekannte Wörter im Raum lokalisiert und verarbeitet werden. Die Genauigkeit der Schätzung steigt. Im CNN werden vier Schichtarten genutzt, um die Daten zu filtern und die wichtigen Merkmale zu erkennen: die *Convolution*-Schicht, die ReLU-Schicht, die *Pooling*-Schicht und das *Flattening*. Diese Schichten bauen aufeinander auf und werden im Folgenden erläutert.

Convolution-Schicht

Wie in Kapitel 4.2.1 beschrieben, wird jedes Wort als Vektor im Raum dargestellt. Nach Verlassen der *Embedding*-Schicht ist ein Tweet eine Matrix beziehungsweise ein Tensor (\mathbf{H}^{s-1}) in der Größe $\tilde{\tau} \times n$, wobei $\tilde{\tau}$ die Anzahl der Worte und n die Dimension des Wortvektors darstellt. Jedes Wort $t \in \tilde{\tau}$ kann auch als Neuron der Eingabeschicht aufgefasst werden. Der gesamte Textkorpus ist ein Tensor mit $\mathbb{R}^{\tilde{\tau} \times n \times m}$. Das Grundprinzip der Faltungsoperation¹³ ist das Nutzen von Filtern zum Extrahieren von Merkmalen, auch Kernel K^s genannt (vgl. Goodfellow et al., 2018, S. 370-373). In jeder Faltungsoperation werden mehrere Filter (\mathbf{K}_i^s) genutzt, die unterschiedliche Parameter aufweisen (vgl. Goodfellow et al., 2018, S. 371). Bei Textdaten wird eine so

¹³Faltung ist im mathematischen Kontext das Berechnen einer Funktion $s(t)$, die beschreibt, wie eine Funktion $x(t)$ durch eine Funktion $w(a)$ verändert wird (vgl. Goodfellow et al., 2018, S. 370). Im diskreten Fall ist das $s(t) = (x * w)(t) = \sum_{a=-\infty}^{\infty} x(a)w(t-a)$ (vgl. Goodfellow et al., 2018, S. 371).

genannte 1D-Convolution Schicht verwendetet (vgl. Goldberg, 2017, S153). Jeder Filter ist eine Matrix in $\mathbb{R}^{\tilde{r} \times u}$ und kann als ein Neuron der *Convolution*-Schicht aufgefasst werden. n_f wird innerhalb des Modells gewählt und ist die Dimension des Filters, wobei $n_f < n$. Die Gewichte w_{ij} verbinden jedes Eingaberoutinen i der Matrix \mathbf{H}^{s-1} mit einem Filter \mathbf{K}_i^s . Alle Filter einer Faltungsoperation zusammengefasst bilden einen Tensor mit drei Dimensionen. Die Filter fassen keine Informationen über Worte hinweg zusammen. Durch Anpassung der Gewichte der Filter während des Lernalgorithmus kommt es zur höheren Gewichtung der Filter, welche die Merkmale der Textdaten am eindeutigsten extrahieren. Die Filter werden an der Position $(0, 0)$ der Matrix angelegt und die Werte der Filter mit den Werten der Matrix multipliziert und aufsummiert (vgl. Gleichung 22). Anschließend wandert der Filter von links nach rechts über die Matrix (vgl. Goodfellow et al., 2018, S. 373). Mit welcher Weite der Filter wandert, ist von den Daten abhängig und kann während der Warengruppenoptimierung bestimmt werden. Die Anzahl der möglichen Faltungen ist $\tilde{i} = 0, \dots, \tilde{l}$. Das Ergebnis einer Faltungsoperation wird als Merkmals karte (*Feature MAP*) dargestellt und mit \mathbf{S}_F bezeichnet (vgl. Goodfellow et al., 2018, S. 372). Die Faltungsoperation mit einer Eingabe \mathbf{H}^{s-1} und einem Filter K^s stellt sich dann wie folgt dar:¹⁴

$$S_F^s(\tilde{i}) = (\mathbf{H}^{s-1} * \mathbf{K}^s)(\tilde{i}) = \sum_i \sum_{n_f} \mathbf{H}^{s-1}(i, \tilde{i} + n_f) \mathbf{K}^s(i, n_f) \quad (22)$$

Um die Effizienz der Faltungsoperation zu erhöhen, werden die einzelnen Berechnungen der Filter mit den Daten für jede Verschiebung und jeden Filter nicht aufeinanderfolgend, sondern parallel durchgeführt. Hierfür wird jedes \mathbf{H}^s entsprechend der genutzten Filtergröße unterteilt und die Werte werden nach der Unterteilung als Spalten in einer transponierten Matrix $\phi(\mathbf{H}^s)$ gespeichert (vgl. Wu, 2017, S. 18). Um die Faltungsoperation durchzuführen, werden sämtliche Filter vektorisiert und zu einer Matrix \mathbf{F} zusammengefasst. Die Faltungsoperation ist damit die Matrixmultiplikation $\phi(\mathbf{H}^s)\mathbf{F}$ (vgl. Wu, 2017, S. 18). Wie in Kapitel 4.1.1 beschrieben, werden die Anpassungen der Gewichte in der *Backpropagation* durch die Anwendung der Kettenregel bestimmt. Durch Einsetzen von $\phi(\mathbf{H}^s)\mathbf{F}$ in Gleichung 13 zeigt sich, dass die Anpassung des Gewichts w_{ij} jedes Filters \mathbf{K} durch die Multiplikation der gefilterten Eingabedaten mit dem Signal des Fehlerterms bestimmt wird (Gleichung 23) (vgl. Wu, 2017, S. 19).

¹⁴vgl. visuelle Ergänzung im Anhang A.1.

$$\frac{\partial L}{\partial w_{ij}^s} = \delta_j^s \phi(\mathbf{H}^{s-1})^T \quad (23)$$

ReLU-Schicht

In der ReLU-Schicht kommt es während des Lernalgorithmus nicht zu einer Anpassung von Parametern. Durch die Anwendung von ReLU wird eine mögliche Linearität in den Daten reduziert (vgl. Gleichung 24) (vgl. Goodfellow et al., 2018, S. 213). Darüber hinaus kommt es durch das Setzen von negativen Werten auf null zu einer Reduzierung der Berechnungszeit (vgl. Goodfellow et al., 2018, S. 213). In der Merkmalskarte auftretende negative Werte stellen Bereiche in den Daten dar, in denen keine Strukturen mit einem hohen Informationsgehalt erkannt wurden. Das Nullsetzen der negativen Werte führt zu keinem Informationsverlust und reduziert die Komplexität der Berechnung.

$$ReLU(a)_i = \max(0, a_i) = \begin{cases} 0 & a_i < 0 \\ a_i & \text{sonst} \end{cases} \quad (24)$$

Pooling

Um die Robustheit des Modells gegenüber Verschiebungen der Informationsmerkmale in den Daten zu erhöhen, wird *Pooling* genutzt. *Pooling* fasst die durch die *Convolution*-Schicht extrahierten Bedeutungscluster zusammen und führt so zu einer höheren Robustheit des Modells gegenüber leichten Verschiebungen in den Daten (vgl. Goodfellow et al., 2018, S. 379-380). Wörter mit ähnlicher Bedeutung werden dadurch als ähnlich erkannt. Durch das Zusammenfassen der Informationen der Merkmalskarte wird darüber hinaus die Größe der Matrix reduziert und dies führt zu einer Steigerung der Effizienz des Modells (vgl. Goodfellow et al., 2018, S. 379). Wie bereits in der *Convolution*-Schicht ist die Eingabe der Schicht der Tensor \mathbf{H}^{s-1} . Vergleichbar zur *Convolution*-Schicht bewegt sich ein Rechteck über die Matrizen. Innerhalb der Rechtecke werden je nach *Pooling*-Verfahren verschiedene mathematische Operationen vorgenommen. Im Bereich der Textklassifizierung ist das Max-*Pooling* etabliert (vgl. Goodfellow et al., 2018, S. 379). Innerhalb der Rechtecke wird das Maximum der Ausprägungen ermittelt und dieser Wert in H_i^s übertragen. Ein alternatives Verfahren wäre das Bilden des Durchschnittes innerhalb sämtlicher Quadrate (vgl. Goldberg, 2017, S156). Die Anwendung von Max-*Pooling* legt den Fokus auf

das Vorhandensein eines wichtigen Merkmals, unabhängig von der Merkmalsdichte in der Umgebung. Durch die *Pooling*-Schicht reduziert sich die Matrix auf $\tilde{\tau}^s = \frac{\tilde{\tau}^{s-1}}{\tilde{\tau}}$ und $n^s = \frac{n^{s-1}}{n}$ (vgl. Goodfellow et al., 2018, S. 380; Wu, 2017, S. 23).

Flattening

Der abschließende Schritt des CNN ist die Vektorisierung jeder Ausgabe aus der *Pooling*-Schicht. Dies ist notwendig, um die Daten in ein klassisches neuronales Netz zu geben und mit der Softmax-Funktion arbeiten zu können (vgl. Kapitel 4.1.2). Die Dimension jedes Vektors ist $\tilde{\tau} \cdot n$.

4.3.2 Long Short-term Memory

LSTMs sind auf *Recurrent Neural Network* (RNN) aufbauende Netzwerke (vgl. Hochreiter & Schmidhuber, 1997, S. 1736). Um Sprache zu verstehen, arbeitet der Mensch mit Informationen in Sätzen über die Zeit. Das in einem Moment ausgesprochene Wort eines Gesprächspartners alleine liefert Informationen, der Kontext wird jedoch erst durch die Einbettung im Satz deutlich. RNNs sind das künstliche Äquivalent eines Kurzzeitgedächtnisses. Wörter, die sich am Satzanfang befinden, sind zur Einordnung späterer Wörter notwendig. Informationen am Anfang einer Sequenz haben einen Einfluss auf die Bedeutung von Ausprägungen zum späteren Zeitpunkt in der Sequenz. Daraus folgt, dass RNNs Informationen über eine Sequenz ($1 \leq t \leq \tau$) hinweg nutzen. Der verwendete Textkorpus besteht aus einer Vielzahl an Tweets. Im Folgenden wird jedoch lediglich ein Tweet innerhalb des RNNs betrachtet. Der gesamte Datensatz durchläuft die Schicht wie ein einzelner Tweet. Die Eingabe einer Beobachtung in ein RNN ist eine Matrix \mathbf{X} . Ist das RNN nicht die erste verdeckte Schicht im Netz, dann ist die Eingabe ein Vektor \mathbf{h}^{s-1} .¹⁵ Im Folgenden wird jedoch von einer Eingabe \mathbf{X} ausgegangen. Jeder Teil der Sequenz der Eingabe wird einem anderen Berechnungsschritt t im RNN zugeordnet (vgl. Goodfellow et al., 2018, S. 418). Ein Teil der Sequenz entspricht einem Wortvektor \mathbf{x} des Satzes. Die Ausgabe in \mathbf{h}^t ergibt sich aus der Eingabe in \mathbf{x}^t und der Ausgabe zum vorherigen Zeitpunkt \mathbf{h}^{t-1} (vgl. Gleichung 25). Dies kann auch als eine Funktion g^t über sämtliche Eingaben aus der Sequenz \mathbf{x} betrachtet werden, wobei die Anzahl der Schritte in der Sequenz t durch die Länge von \mathbf{x} bestimmt wird (vgl. Goodfellow et al., 2018, S. 219). Die \mathbf{h}^1 bis \mathbf{h}^τ sind durch Gewichte miteinander verbunden (vgl. Goodfellow et al., 2018, S. 421).

¹⁵Im Fall eines vorgelagerten CNN

$$\mathbf{h}^t = g^t(\mathbf{x}^t, \mathbf{x}^{t-1}, \mathbf{x}^{t-2}, \dots, \mathbf{x}^2, \mathbf{x}^1) \quad (25)$$

Die aufeinander aufbauende Struktur des RNNs führt zu dem *Vanishing-Gradient* beziehungsweise dem *Exploding-Gradient*-Problem (vgl. Pascanu et al., 2013, S. 2). Im Zuge der *Backpropagation* wandert der Fehlerterm rückwärts von der Softmatrix durch das Netz. Die anzupassenden Gewichte sind zu Beginn des Lernalgorithmus zufällig und kleiner eins (vgl. Kapitel 4.1.1). Eine RNN-Schicht mit $t = 3$ ist dann für eine Sequenz $\mathbf{h}^3 = (\mathbf{W})^T \mathbf{x} \cdot (\mathbf{W})^T \mathbf{h}^{t+1} \cdot (\mathbf{W})^T \mathbf{h}^{t+2}$. Ist der zufällige Werte für \mathbf{W} nahe null, dann sinkt die Ausgabe der Schicht gegen null. Umgekehrt bedeutet dies in der *Backpropagation*, dass der Gradient $\frac{\partial L}{\partial \mathbf{W}^\tau} = \delta^\tau \mathbf{h}^{\tau-1}$ geringer ist und mit sinkendem τ weiter sinkt. Der sinkende Gradient führt zu einer langsamen Anpassung der Gewichte zu Beginn der RNN-Schicht. Da jedoch die Anpassung der Gewichte in t von den Gewichten in $t - 1$ abhängig ist, führt dies zu einer Kettenreaktion im Netzwerk. Es sei angenommen, die Gewichte in τ sind vollständig angepasst, basieren jedoch auf Berechnungen aus $\tau - t$, die aufgrund des geringen Gradienten fehlerhaft sind. Mit einer höheren Zahl an Iterationen des Lernalgorithmus passen sich die Gewichte in $\tau - t$ an. Daraus folgt eine Veränderung von \mathbf{h}^τ , welche zu einer Veränderung des Fehlerterms führt. Ein vollständig trainiertes Netz ist aufgrund dieser Abhängigkeit theoretisch mit einer hohen Anzahl an Iterationen erreichbar, jedoch nicht praktikabel. Das *Exploding-Gradient*-Problem beschreibt den gegenüberstehenden Effekt. Bei sehr hohen Gewichten multiplizieren sich deren Effekte und ein Optimieren des Netzes ist nicht mehr möglich. (vgl. Pascanu et al., 2013, S. 2).

Die Problematik des *Vanishing-Gradient* und des *Exploding-Gradient* wird im LSTM mittels der Trennung des Transports der Informationen durch die Zellen und des Extrahierens der Informationen gelöst (vgl. Hochreiter & Schmidhuber, 1997, S. 1736). Die Neuronen innerhalb eines RNNs werden durch LSTM-Zellen ersetzt, welche mehrere unterschiedliche Datenbearbeitungsschritte ausführen (vgl. Abbildung 4) (vgl. Goodfellow et al., 2018, S. 455; Olah, 2015). Informationen werden von LSTM-Zelle zu LSTM-Zelle weitergegeben, wobei jede Zelle neue Informationen durch die Eingabe \mathbf{x}^t erhält. Jede LSTM-Zelle hat die Ausgabe \mathbf{h}^t . \mathbf{h}^t wird an die nächste Zelle und die nächste Schicht weitergegeben. In der nächsten Zelle wird \mathbf{h}^{t-1} jedoch nicht als Übertragung von Informationen verstanden, sondern als Übermittler vom Kriterium zur Berechnung der zu übertragenden Informationen. Die Informationen, die durch die Zellen wandern, werden durch den Zellspeicher \mathbf{C}^t dargestellt (vgl. Olah,

2015). Sämtliche vorgestellten Modelle, die auf einem Lernalgorithmus beruhen, werden durch Anpassung der Gewichte optimiert. Für die Berechnung von \mathbf{C}^t aus \mathbf{C}^{t-1} werden jedoch keine Gewichte genutzt oder anders ausgedrückt: $w_{ij} = 1$. So können Informationen die gesamte LSTM-Schicht durchwandern ohne das Entstehen des *Vanishing-Gradient-* beziehungsweise des *Exploding-Gradient-*Problems (vgl. Hochreiter & Schmidhuber, 1997, S. 1745). Die Anpassung der zu extrahierenden Strukturen aus den Daten wird durch das Hinzufügen von drei verschiedenen Ventilen erreicht: das *Forget-Gate* (1), das *Input-Gate* (2) und das *Output-Gate* (3) (vgl. Abbildung 4)(vgl. Hochreiter & Schmidhuber, 1997, S. 1743). Durch diese Ventile werden die zu extrahierenden Strukturen zwischen \mathbf{C}^{t-1} und \mathbf{C}^t bearbeitet. σ ist die logistische Sigmoidfunktion $\sigma = \frac{1}{1+\exp(-x)}$ (vgl. Hochreiter & Schmidhuber, 1997, S. 1768). Die Ausgabe der Sigmoidfunktion liegt zwischen null und eins. Durch Nutzung von σ wird bestimmt, mit welchem Anteil die Informationen die Ventile durchqueren.

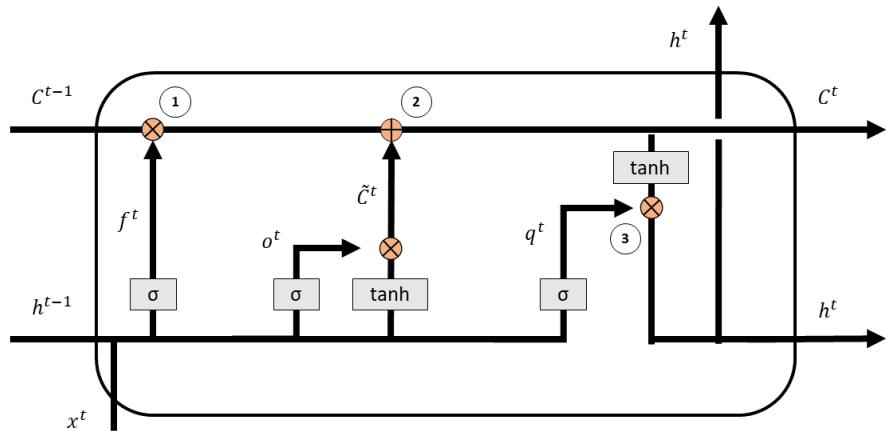


Abbildung 4: Schaubild einer LSTM Zelle (eigene Darstellung in Anlehnung an Olah 2015)

Forget-Gate

Die zu Beginn der Sequenz extrahierten Informationen wandern mit \mathbf{C} durch die Zellen des LSTMs. Der erste Bearbeitungsschritt in einer LSTM-Zelle bestimmt, welche Daten aus dem Zellspeicher weiter genutzt werden (vgl. Gleichung 26). Hierzu wird die Ausgabe \mathbf{h}^{t-1} mit \mathbf{W} gewichtet und die Eingabe \mathbf{x}^t mit den Gewichten \mathbf{U} , wobei $\mathbf{U} \in \mathbb{R}^{i \times j}$ und $\mathbf{U} = (u_{11}, u_{12} \dots, u_{ij})$ (vgl. Goldberg, 2017, S. 180; Goodfellow et al., 2018, S. 456). \mathbf{U} wird wie \mathbf{W} im Lernalgorithmus angepasst. Die Gewichte der Ventile sind im Lernalgorithmus unabhängig voneinander. So sind die Gewichtsmatrizen des *Forget-Gate* \mathbf{U}^f und \mathbf{W}^f , des *Input-Gate* \mathbf{U}^o und \mathbf{W}^o und des *Output-Gate* \mathbf{U}^q und

W^q. Die Verzerrung ist \mathbf{b}^f . Durch das Nutzen der Sigmoidfunktion wird abhängig von den Informationen festgelegt, welche Daten aus der vorherigen Schicht übernommen werden (vgl. Goodfellow et al., 2018, S. 456). Es wird anhand der neuen Informationen aus \mathbf{x}^t geprüft, welche Informationen nicht weiter verwendet werden (vgl. Olah, 2015). Wenn beispielsweise die bisherige gespeicherte Information war, dass keine Verneinung genutzt wurde, in \mathbf{x}^t nun aber eine Verneinung vorhanden ist, wird die Information aus \mathbf{C}^{t-1} gelöscht beziehungsweise der Einfluss reduziert. f_i^t beschreibt, welcher Anteil der Informationen aus C_i^{t-1} in C_i^t übertragen werden soll.

$$f_i^t = \sigma(b_i^f + \sum_j u_{ij}^f x_j^t + \sum_j w_{ij}^f h_j^{t-1}) \quad (26)$$

Input-Gate

Während im *Forget-Gate* ausschließlich Informationen gefiltert werden, werden im *Input-Gate* Informationen zum Zellspeicher hinzugefügt. Informationen aus \mathbf{h}^{t-1} und \mathbf{x}^t werden in $\tilde{\mathbf{C}}^t$ (vgl. Gleichung 28) zusammengefasst und mit o^t gewichtet (vgl. Gleichung 27) (vgl. Hochreiter & Schmidhuber, 1997, S. 1744). Die Information aus \mathbf{x}^t , dass eine Verneinung vorliegt, kann so dem Zellspeicher hinzugefügt werden. Die Eingabe in \mathbf{C}^t $\tilde{\mathbf{C}}^t$ wird durch *Tangens hyperbolicus* ($\tanh(x) = \frac{\exp(2x)-1}{\exp(2x)+1}$) auf zwischen eins und minus eins skaliert (vgl. Goldberg, 2017, S. 45,180). Das Erlauben von negativen Werten ist notwendig, um eine negative Anpassung während des Lernalgorithmus zu ermöglichen (vgl. Olah, 2015). Die Gewichtsmatrizen \mathbf{U}^C und \mathbf{W}^C sind im Lernalgorithmus unabhängig von \mathbf{U}^o und \mathbf{W}^o .

$$o_i^t = \sigma(b_i^o + \sum_j u_{ij}^o x_j^t + \sum_j w_{ij}^o h_j^{t-1}) \quad (27)$$

$$\tilde{C}_i^t = \tanh(b_i^C + \sum_j u_{ij}^C x_j^t + \sum_j w_{ij}^C h_j^{t-1}) \quad (28)$$

Der neue Speicher wird aus der Summe des alten Speichers berechnet, reduziert um die gelöschten Informationen (Gleichung 26), zuzüglich der neu extrahierten Informationen (Gleichung 27, 28) (vgl. Goldberg, 2017, S. 180; Goodfellow et al., 2018, S. 456; Olah, 2015).

$$C_i^t = f_i \cdot C_i^{t-1} + o_i^t \cdot \tilde{C}_i^t \quad (29)$$

Output-Gate

Die Ausgabe einer LSTM-Zelle basiert auf dem Zellstatus \mathbf{C}^t . Durch das Training wird erlernt, welche Informationen aus der vorherigen Zelle an die nächste Zelle weitergereicht werden. Im ersten Schritt wird der Anteil des Zellstatus an der Ausgabe q^t festgelegt (vgl. Gleichung 30) (vgl. Goldberg, 2017, S. 180; Goodfellow et al., 2018, S. 456). Anschließend wird der Anteil mit dem Zellstatus multipliziert, wobei der Zellspeicher mit dem *Tangens hyperbolicus* skaliert wird (vgl. Gleichung 31).

$$q_i^t = \sigma(b_i^q + \sum_j u_{ij}^q x_j^t + \sum_j w_{ij}^q h_j^{t-1}) \quad (30)$$

$$h_i^t = \tanh(C_i^t) q_i^t \quad (31)$$

4.3.3 Regularisierungsarchitektur

Beim Training von neuronalen Netzen ist zu beachten, dass die Modelle sich nicht zu stark an die Trainingsdaten anpassen und es zu *Overfitting* kommt. Hierzu werden verschiedene Regularisierungstechniken, genutzt. Die Regularisierungstechniken die in dieser Arbeit verwendet werden, werden im Folgenden beschrieben.

Dropping

Zur Verringerung der Gefahr von *Overfitting* beim Trainieren von *Machine-Learning*-Modellen hat sich eine Variation in den Trainingsdaten als wirksam erwiesen. In *Machine-Learning*-Algorithmen wie *Random Forest* wird dazu *Bagging* genutzt (vgl. Goldberg, 2017, S. 286). Für jeden Baum wird dabei eine zufällige Menge an Daten aus dem Trainingsdatensatz gezogen und auf diese Teilmenge wird dann trainiert. Der zweite Teil dient als Testdatensatz. Das Ergebnis über sämtliche Bäume hinweg wird gemittelt. Das Durchführen von *Bagging* beim Trainieren von neuronalen Netzen führt jedoch schnell zu einem hohen Rechenaufwand aufgrund einer großen Anzahl von parallel zu trainierenden Netzen bei gleichzeitigem Lernen mit *Backpropagation* (vgl. Goldberg, 2017, S. 287; Srivastava et al., 2014, S. 1930).

Beim Training von neuronalen Netzen wird daher das *Dropout*-Verfahren (Srivastava et al., 2014) genutzt, welches den Effekt einer hohen Anzahl an Netzen approximiert (vgl. Goldberg, 2017, S. 288). In die Architektur eines neuronalen Netzes wird eine *Dropout*-Schicht innerhalb der verdeckten Schicht (vgl. Abbildung 1) eingefügt. Diese Schicht besteht aus einer binären Matrix in der Größe der Eingabe (vgl. Goldberg, 2017, S. 288). In jedem Schritt einer Epoche wird die Verteilung der Einsen und Nullen in der Matrix zufällig festgelegt. Das Verhältnis von Nullen zu Einsen ist ein Hyperparameter. Die binäre Matrix wird mit den Daten multipliziert. Im nächsten Schritt der Epoche ist die Verteilung der Einsen und Nullen verändert und dadurch werden andere Daten zum Schätzen genutzt. Es entsteht wie beim *Bagging* eine zufällige Ziehung der Trainingsdaten. Jede Epoche kann als einzelnes neuronales Netz betrachtet werden, da jede Epoche mit einer anderen Datenkombination trainiert wird. Im *Bagging* sind die Bäume unabhängig voneinander. Beim *Dropout*-Verfahren hingegen werden Parameter Teilmengen zwischen den Netzen geteilt (vgl. Goldberg, 2017, S. 288). Obwohl theoretisch mehrere Netze nebeneinander entstehen, wird der Lernalgorithmus auf ein Netz angewendet. Durch die Vielzahl an möglichen Trainingsdatenkombinationen wird ein *Overfitting* des Modells erschwert (vgl. Srivastava et al., 2014, S. 1930).

Früher Abbruch

In jeder Epoche wird der Trainingsdatensatz zufällig in einen Trainings- und einen Testdatensatz geteilt (vgl. Goldberg, 2017, S. 275). Beide Teile werden nach jeder Epoche mit dem zu trainierenden Modell geschätzt. Während der Fehler des Trainingsdatensatzes kontinuierlich sinkt, ist der Verlauf beim Testdatensatz meistens abflachend (vgl. Goldberg, 2017, S. 273). Ein früher Abbruch erfolgt durch das Beenden des Trainings, wenn sich der Validierungsfehler einer festgelegten Anzahl an Epochen nicht verbessert hat (vgl. Goldberg, 2017, S. 273). Durch den frühen Abbruch wird ein zu genau auf den Trainingsdatensatz trainiertes Modell verhindert. Der Zeitpunkt des Abbruches kann auch nach dem Training eines Modelles bestimmt werden.

L^2 Parameter Regularisierung

Die Struktur einer Beobachtung kann stark von der Struktur anderer Daten abweichen. In diesem Fall handelt es sich um einen Ausreißer. Dieser kann einen natürlichen Ursprung haben. Andererseits kann es sich um einen Fehler in den Daten handeln. Ist die gewählte *Batch*-Größe (vgl. Kapitel 4.1.1) klein gewählt, hat ein Ausreißer einen starken Einfluss auf die Anpassung der Gewichte (vgl. Goldberg, 2017, S. 29).

Um einen Ausreißer richtig schätzen zu können, müssen die Gewichte hoch sein. Noch stärker ist der Effekt bei fehlerhaften Daten, da ein Zusammenhang hergestellt wird, der nicht vorhanden ist (vgl. Goldberg, 2017, S. 29). Um diesen Einfluss im Training zu reduzieren, wird die L_2 Funktion als Parameter Regularisierung genutzt. Durch Hinzufügen eines Wertes R bei der Minimierung der Verlustfunktion wird es dem Lernalgorithmus ermöglicht, die falsche Klassifizierung von stark abweichenden Beobachtungen zu akzeptieren und einer zu starken Anpassung entgegenzuwirken (vgl. Goldberg, 2017, S. 29). Durch die Wahl von $R_{L_2} = \sum_{ij} (\mathbf{W}_{ij})^2$ steigt der Fehlerterm quadratisch mit der Größe der Gewichte (vgl. Goldberg, 2017, S. 29). Ein hoher Fehlerterm führt im Lernalgorithmus zu sinkenden Gewichten (vgl. Kapitel 4.1.1). Eine Überanpassung an Ausreißer wird so verhindert.

4.4 Methodik der sozialen Netzwerkanalyse

Die von Twitter gesammelten Tweets werden durch *Top2vec* in Cluster unterteilt und jedem Cluster werden Wörter als Thema zugeordnet (vgl. Kapitel 4.2). Diese Tweets stehen ohne Verbindungen im Raum. Zur Analyse der Kommunikationsnetze und insbesondere zur Analyse der Regierungskommunikation auf Twitter werden die Antworten auf die Tweets als Verbindung genutzt. Im Raum wird jeder vorkommende Account durch einen Knoten dargestellt. $\mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_I)$ beschreibt die Position der Knoten im zweidimensionalen Raum. Entsprechend ist $\mathbf{v}_i = (v_{i1}, v_{i2})$. Von diesen Knoten gehen Linien mit den Gewichten w_{ij} zu den Accounts, auf die geantwortet wurde. Unbearbeitet ergibt sich eine nicht zu analysierende zufällige Verteilung der Knoten. Zur Analyse der Zusammenhänge werden die Abstände der Accounts zueinander durch die Antworten aufeinander definiert. Je häufiger ein Account auf einen anderen geantwortet hat, desto näher liegen die Accounts im Raum beieinander. Durch dieses Vorgehen kann analysiert werden, welche Accounts in einer Diskussion zu einem Thema eine besondere Stellung einnehmen und welche Accounts über die Antworten miteinander verbunden sind. Innerhalb dieser Struktur lässt sich die Regierungskommunikation auf Twitter analysieren.

Für die Berechnung des Netzwerks wird ein *Force-directed Graph*-Algorithmus verwendet. *Force-directed*-Graphen folgen einem physikalischen Ansatz und minimieren die Energie im System. *OpenOrd* (Martin et al., 2011) basiert auf dem Fruchterman-Reingold Algorithmus, der besonders für das Berechnen von großen Graphen konzipiert wurde (vgl. Martin et al., 2011, S. 2). Wie in Kapitel 4.2.2 beschrieben, werden während der Optimierung von *Force-directed*-Graphen die Knoten zusammen- oder

auseinandergezogen. In *OpenOrd* wird ein ungerichteter Graph berechnet und somit ist $w_{ij} = w_{ji}$ symmetrisch. D_{v_i} in Gleichung 32 ist die Dichte am Knoten v_i . Die Berechnung der Dichte für jeden Knoten bei jeder Verschiebung ist sehr rechenintensiv (vgl. Martin et al., 2011, S. 3). Anstelle einer Dichte pro Knoten werden daher Dichtefelder berechnet (vgl. Davidson et al., 2001, S. 5). Jeder Knoten wird von einem Kreis umgeben, indem der Dichtewert im Knoten am höchsten ist und von diesem aus zu den Rändern quadratisch sinkt (vgl. Davidson et al., 2001, S. 5). Die Werte in den überlappenden Kreisen werden summiert und ergeben somit Dichtefelder (vgl. Davidson et al., 2001, S. 5). Das Addieren von D_{v_i} in der Minimierung führt zu einem Auseinanderdriften der Knoten in Bereichen mit hoher Dichte. Der gegenüberstehende Effekt ist $\sum_j (w_{ij} d(v_i, v_j)^2)$. Die Distanz d zwischen zwei Knoten wird mit w_{ij} gewichtet. Im Laufe der Minimierung sinkt die Distanz von Knoten mit hohen Gewichten stärker als die Knotenkombination mit niedrigen Gewichten.

$$\min_{v_1, \dots, v_n} \sum_i (\sum_j (w_{ij} d(v_i, v_j)^2) + D_{v_i}) \quad (32)$$

Zur Optimierung des Graphen wird für jeden Knoten die innere Summe aus Gleichung 32 berechnet, wobei die anderen Knoten an den Positionen fixiert sind. Die Position des betrachteten Knotens wird angepasst und anschließend wird mit den anderen Knoten die Iteration fortgeführt (vgl. Martin et al., 2011, S. 3). Ein Schritt der Iteration ist beendet, wenn alle Knoten einmal in der Position verschoben wurden. Bei der Verschiebung eines Knotens sind zwei neue Positionen möglich. Die Verschiebung, die $(\sum_j (w_{ij} d(v_i, v_j)^2) + D_{v_i})$ minimiert, wird durchgeführt (vgl. Martin et al., 2011, S. 3). Die erste neue Position ergibt sich durch einen zufälligen Sprung von v_i , wobei die Weite des Sprunges in verschiedenen Phasen unterschiedlich beschränkt ist. Die zweite neue mögliche Position wird durch den *barrier jump* (Davidson et al., 2001, S. 4-5) berechnet. Die mögliche neue Position ist der gewichtete Mittelwert der verbundenen Knoten (vgl. Davidson et al., 2001, S. 4-5). Der Sprung wird mit einem Dämpfungsgewicht und die Energie im neuen Punkt durch ein Anziehungsgewicht erhöht (vgl. Martin et al., 2011, S. 3). Beide Gewichte sind zwischen null und eins. Je höher das Dämpfungsgewicht ist, desto größer werden die Sprünge. Je geringer das Anziehungsgewicht ist, desto geringer wird die Bestrafung bei einem Sprung. Ein hohes Dämpfungsgewicht in Kombination mit einem niedrigen Anziehungsgewicht führt zu weiteren Sprüngen der Knoten. Diese zwei Parameter bestimmen die Annäherung an ein Minimum beziehungsweise die Möglichkeit, lokale Minima zu überspringen

(vgl. Davidson et al., 2001, S. 4-5).

OpenOrd wird in fünf Phasen unterteilt: *liquid*, *expansion*, *cool-down*, *crunch* und *simmer* (vgl. Martin et al., 2011, S. 3). In den Phasen *liquid*, *expansion* und *cool-down* variieren das Dämpfungs- und Anziehungsgewicht, wobei in den Phasen *expansion* und *cool-down* die höchste Bewegungsfreiheit erlaubt ist (vgl. Martin et al., 2011, S. 3). Die anfänglich zufällige Verteilung der Knoten wird aufgebrochen und die ersten vorläufigen Cluster werden in den *liquid*, *expansion* und *cool-down* gebildet. In den Phasen *crunch* und *simmer* wird die Weite eines möglichen zufälligen Sprungs auf $\frac{1}{4}$ reduziert (vgl. Martin et al., 2011, S. 3). Gleichzeitig sinken die Dämpfung und Anziehungskraft. In diesen Phasen werden bereits entstandene Strukturen optimiert.

Ein weiterer Parameter innerhalb des *OpenOrd*-Algorithmus ist das *Edge-cutting* (vgl. Martin et al., 2011, S. 3-4). Der dabei verwendete Parameter liegt zwischen null und eins. Im *Edge-cutting* wird die Distanz sämtlicher verbundener Knoten berechnet. Anschließend wird die größte Distanz ausgewählt und mit dem *Edge-cutting*-Parameter als Schwellwert genutzt (vgl. Martin et al., 2011, S. 4). Ein Wert von null entspricht 100% der längsten Distanz und ein Wert von eins entspricht 0%. Die Verbindung zwischen zwei Knoten wird getrennt, wenn die Distanz größer als der berechnete Teil der längsten Distanz ist (vgl. Martin et al., 2011, S. 4). *Edge-cutting* wird in den Phasen *expansion* und *cool-down* genutzt. Knoten mit einer hohen Distanz und einem hohen Gewicht können die Clusterbildung erschweren, da die hohen Gewichte den Mittelpunkt des Clusters verzerren. *Edge-cutting* führt zu einer Trennung bei einer hohen Distanz und so zu vermehrter Clusterbildung (vgl. Martin et al., 2011, S. 4). Ein Übermaß an *Edge-cutting* führt zu einem Verlust an Informationen im Netzwerk. Durch die Beschränkung des *Edge-cutting* auf die Phasen *expansion* und die *cool-down* können jedoch bereits in der Phase *liquid* vorläufige Cluster gebildet werden und häufig verbundene Cluster befinden sich bereits in einer Umgebung. Trotzdem muss der Schwellwert im *Edge-cutting* vorsichtig gewählt werden.

4.5 Daten

Die Datenbeschaffung und die Datenaufbereitung stellen die Grundlage einer empirischen Analyse dar. Die drei unterschiedlichen Datensätze dieser Arbeit stehen sowohl in der Beschaffung als auch in der Datenaufbereitung in Abhängigkeiten zueinander. Um für diese Prozesse einen Überblick zu ermöglichen, wird in diesem Kapitel im ersten Schritt die Datenbeschaffung erläutert und anschließend die Datenaufbereitung für die drei Datensätze beschrieben.

4.5.1 Datenerhebung

Zum Erstellen des Datensatzes wurden sämtliche deutschsprachigen Tweets vom 01.01.2020 bis zum 31.08.2020, welche die Wörter oder Teilwörter *Corona*, *Covid* oder *SARS-CoV-2* enthalten, Ende Oktober 2020 innerhalb von zwei Wochen heruntergeladen. Insgesamt wurden so 2.946.014 Tweets mithilfe von *Tweepy* (Roesslein, 2009) von Twitter extrahiert. Die geladenen Informationen enthalten neben dem Tweet die Tweet-ID, das Datum, die Uhrzeit und den Nutzernamen.

Für die Analyse der Interaktion innerhalb der extrahierten Tweets wurden die Antworten auf die Tweets mit einer Kombination von *sns scrape* (JustAnotherArchivist, 2020) und *Tweepy* heruntergeladen. Eine direkte Abfrage der Antworten auf Tweets ist nicht möglich. Um diese Antworten erhalten zu können, wurde mit *sns scrape* jeder Nutzeraccount des originalen Tweets aufgerufen und die ID der Antworten auf sämtliche Tweets des Accounts gescraped. Es wurden Antworten inkludiert, die fünf Tage nach dem Post des originalen Tweets auf den Account geantwortet hatten. Die Antworten auf dem Account wurden anschließend über *Tweepy* geladen. Jede geladene Antwort enthält den Text, den Usernamen, die Antwort-ID, das Datum, die Uhrzeit und die Tweet-ID, auf die geantwortet wurde. Durch Abgleichen der geladenen Antworten mit den Tweet-IDs wurden die Antworten auf die Tweets extrahiert.

Für das Trainieren eines Modells zur Schätzung des Sentiments der Tweets ist ein Trainingsdatensatz notwendig. Hierzu wurde ein von Mozetič et al. (2016) bereit gestellter Datensatz genutzt. Aufgrund der Twitter-Richtlinien ist das Veröffentlichen eines vollständigen Twitter-Datensatzes nicht gestattet (vgl. Twitter, 2020b). Der Datensatz wird durch die Veröffentlichung der Tweet-IDs bereitgestellt. Die Tweets wurden auf Basis der IDs geladen. Jedem Tweet wurde von Mozetič et al. (2016) manuell ein Label zugeordnet. Die Tweets wurden von zwei verschiedenen Personen in positiv, negativ und neutral gelabelt, wobei ein deutlicher Qualitätsunterschied in den Labeln existiert. Das Erstellen eines Modells zur Klassifizierung wird dadurch erschwert. Der Datensatz wurde im Original für 109.130 Tweets erstellt. Durch den Effekt von beispielsweise gelöschten Accounts reduzierte sich die Zahl der geladenen Tweets auf 102.273. Ein anderer frei zugänglicher Datensatz ist von Cieliebak et al. (2017). Dieser enthält 10.000 gelabelte Tweets und ist aufgrund seiner geringen Anzahl an Tweets als Datensatz für eine Verwendung mit neuronalen Netzen nicht geeignet.

Sämtliche erhobenen Daten in dieser Arbeit waren zum Zeitpunkt der Erhebung frei zugänglich und öffentlich einsehbar. Auf Twitter besteht die Möglichkeit, einen Account auf nicht öffentlich zu stellen. Des Weiteren werden keine Account-Namen und keine dahinterstehenden Personen genannt, wenn diese nicht Personen des öffentlichen Lebens, Unternehmen, staatliche Stellen oder Personen in Ausübung eines öffentlichen Amtes sind.

4.5.2 Datenaufbereitung

Beim Rechnen mit Textdaten stellt die Datenaufbereitung einen wichtigen Schritt zur Verbesserung der Genauigkeit dar (Effrosynidis et al., 2017; Jianqiang & Xiaolin, 2017; Symeonidis et al., 2018). Tweets weisen im Vergleich zu anderen Textdatensätzen ein besonders hohes Rauschen auf. Zeitungs- oder auch Wikipedia-Artikel enthalten wenige Rechtschreib- oder Grammatikfehler, da sie vor einer Veröffentlichung überarbeitete und lektoriert werden - im Fall von Zeitungsartikeln durch professionelle Strukturen und bei Wikipedia-Artikel durch die freie Diskussion der Mitwirkenden. Bei Tweets von Regierungen und Unternehmen ist mit einer ähnlich geringen Fehlerrate zu rechnen. Die Mehrzahl der Tweets wird jedoch von Privatpersonen verfasst und unterliegt keinem Korrekturprozess. Eine weitere Besonderheit von Tweets ist die Begrenzung auf 280 Zeichen. Durch diese Begrenzung kommt es zu Abkürzungen und Verkürzungen von Aussagen. Diese Besonderheiten von Tweets führen dazu, dass in dieser Arbeit keine vortrainierten *Word-Embedding*-Modelle genutzt, sondern die Modelle selbstständig auf die erhobenen Textdaten trainiert werden. Die bekannten deutschen vortrainierten *Word-Embedding*-Modelle wurden umfangreich auf deutsche Wikipedia-Artikel trainiert. Das Nutzen von diesen *Word-Embeddings* führt bei der Bearbeitung von Tweets zu Verzerrungen.

Wie oben beschrieben, wird in dieser Arbeit mit drei verschiedenen Datensätzen gearbeitet. Durch die Bearbeitung der Tweets werden die Themen extrahiert, die während des erhobenen Zeitraums auf Twitter behandelt wurden. Durch die Antworten auf die Tweets werden die Kommunikationsnetze zu den verschiedenen Themen erstellt. Abschließend werden auf der Grundlage des Datensatzes von Mozetič et al. (2016) Modelle zum Schätzen des Sentiments trainiert. Da die Modelle die Strukturen der Tweets aus dem Datensatz von Mozetič et al. (2016) extrahieren, welche zu einer präzisen Schätzung des Sentiments führen, ist es notwendig, die gleichen Datenaufbereitungsschritte für die Antworten auf die Tweets und den Trainingsdatensatz durchzuführen. Eine andere Datenaufbereitung der Antworten würde zu einer Verz-

errung der Schätzung führen.

Auf jedem Tweet in den drei Datensätzen wird eine Reihe von Bearbeitungsschritten durchgeführt. Im ersten Schritt werden in jedem Tweet die Umlaute und der Buchstabe „ß“ durch Schreibweisen ersetzt, die robuster gegenüber verschiedene Datenformaten sind. Ein „ä“ wird zu einem „ae“ und das „ß“ zu einem „ss“. Werden in den Tweets Zahlen verwendet, sind die Modelle nicht in der Lage, die nummerische Darstellung im semantischen Raum einzuordnen. Um dieses Problem umgehen zu können, wird jede Zahl in eine ausgeschriebene Textform umgewandelt (vgl. Effrosynidis et al., 2017, S. 396). Wird ein anderer Account in einem Tweet verlinkt, dann wird dem Nutzernamen ein „@“ vorangestellt. Das „@“ wird aus dem Tweet gelöscht und der Nutzername durch das Wort „Name“ ersetzt (vgl. Effrosynidis et al., 2017, S. 397; Symeonidis et al., 2018, S. 299). Für die Bestimmung des Sentiments oder des Themas des Tweets ist es unerheblich, welcher Nutzer genannt wurde. Entscheidend ist, dass ein anderer Nutzer erwähnt wurde. Ein weiteres häufig vorkommendes Zeichen in Tweets ist „#“. Das Zeichen „#“ wird am Ende eines Tweets genutzt, um den Tweet einer selbst gewählten Themengruppe zuzuweisen. Durch die Verbindung des „#“ mit Wörtern haben diese Wörter jedoch nicht mehr die gleiche Struktur wie dasselbe Wort im Text. Um eine Vergleichbarkeit erreichen zu können, wird das „#“ aus den Tweets gelöscht (vgl. Effrosynidis et al., 2017, S. 397; Symeonidis et al., 2018, S. 299). Ein häufiges Vorgehen ist das Verlinken von Webseiten in Tweets durch die Nutzer. In den Textdaten sind diese Verlinkungen durch die URLs im Text zu erkennen. Die URL enthält keine Information über die Stimmung in einem Text und auch ein direkter Bezug zum Inhalt der Verlinkung ist nicht zu entnehmen. Da in der URL die verlinkte Webseite genannt ist, können darüber hinaus Verzerrungen bei dem Erstellen von Themenclustern entstehen. Eine mögliche Verzerrung resultiert, indem die genutzten Modelle die URLs als eine Gemeinsamkeit der Texte auffassen. Um diese Problematiken zu umgehen, werden die URLs aus den Textdaten entfernt. Jedes Wort im Satz enthält unterschiedlich viel Informationsgehalt. Besonders Artikel oder Anreden liefern keine Informationen über das Sentiment oder das Thema eines Tweets, gleichzeitig erhöhen diese jedoch die Datenmenge (vgl. Symeonidis et al., 2018, S. 301). Diese Wörter werden Stopwörter genannt. Durch das Löschen eines jeden Wortes mit weniger als vier Zeichen aus den Tweets wird die Datenmenge bei geringem Informationsverlust deutlich reduziert. Abschließend werden alle Satzzeichen aus den Tweets entfernt und für alle Buchstaben wird die kleine Schreibweise verwendet, um eine verbesserte Vergleichbarkeit der Wörter zu erreichen (vgl. Effrosynidis et al., 2017, S. 396; Symeonidis et al., 2018, S. 301).

Während für das Entfernen von Stoppwörtern, das Umwandeln von Großbuchstaben, das Löschen von Sonderzeichen und das Umwandeln von Zahlen in verschiedenen Experimenten ein positiver Effekt auf die Genauigkeit der Modelle festgestellt werden konnte, sind die Ergebnisse für die Durchführung von Lemmatisierung und Stemming nicht eindeutig (vgl. Effrosynidis et al., 2017, S. 404; Jianqiang & Xiaolin, 2017; Symeonidis et al., 2018, S. 304-305). Lemmatisierung und Stemming sind zwei unterschiedliche Techniken, um die Vergleichbarkeit von Wörtern erhöhen zu können. Durch Stemming werden Wörter auf ihren Wortstamm reduziert, während Lemmatisierung die Wörter in die Wortform umwandelt, die im Wörterbuch zu finden ist (vgl. Symeonidis et al., 2018, S. 305). Durch das Anwenden der beiden Techniken steigt die Vergleichbarkeit der Wörter, jedoch gehen die Informationen über die genaue Wortform verloren. Symeonidis et al. (2018) fanden einen durchgehend positiven Effekt von Lemmatisierung und Stemming, während Effrosynidis et al. (2017) in unterschiedlichen Datensätzen entgegengesetzte Effekte feststellten. In beiden Untersuchungen wurden die Tweets mit der *Bag of Words*-Technik bearbeitet. In dieser Arbeit wird aufgrund der unklaren Ergebnisse auf die Bearbeitung der Tweets mit Lemmatisierung und Stemming verzichtet. Es ist weiterhin davon auszugehen, dass die möglichen Nachteile durch unterschiedliche Wortformen durch das Berechnen eines semantischen Raumes ausgeglichen werden. Die ähnlichen Wortformen haben im optimalen semantischen Raum einen geringen Abstand zueinander.

Im Anschluss an die Bearbeitung der Tweets werden die Datensätze weiter verarbeitet. Der Datensatz der Tweets zur COVID-19-Pandemie wird auf doppelt auftretende Tweets untersucht. Durch die Anwendung von drei Suchwörtern zum Herunterladen der Tweets können Tweets häufiger im Datensatz enthalten sein, wenn mehrere Suchwörter im Tweet verwendet worden sind. Nach der Bereinigung der doppelten Tweets wurde die Anzahl der Tweets von 2.946.014 auf 2.453.280 reduziert.

Der von Mozetič et al. (2016) zur Verfügung gestellte Datensatz wird nach der Textaufbereitung um die neutralen Tweets bereinigt. Dieser Schritt wird durchgeführt, um eine Eindeutigkeit in der Analyse der Antworten auf die Regierungskommunikation zu erreichen. Dieser Datensatz wird anschließend in einen Trainings und in einen Testdatensatz unterteilt. Mit dem Testdatensatz wird nach dem erfolgreichen Training auf dem Trainingsdatensatz die Genauigkeit der Modelle bei unbekannten Daten untersucht. Die Ziehung erfolgt zufällig unter der Bedingung, dass in beiden Datensätzen das gleiche Verhältnis der positiven und negativen Tweets enthalten ist. Aus dem Trainingsdatensatz wird für das Hyperparametertraining wiederum ein zufälliger

Validierungsdatensatz gezogen. Bei dieser Ziehung wird ebenfalls ein gleiches Verhältnis zwischen positiven und negativen Tweets sichergestellt. Während des Trainings von *Machine-Learning* können umbalancierte Datensätze zu einer Verzerrung der Schätzung des Modells führen. Ein Datensatz ist umbalanciert, wenn die Klassen ungleich im Datensatz verteilt sind. Das Modell erlernt die Eigenschaften zur Schätzung der häufiger vertretenen Klasse stärker und die Gewichte passen sich an diese Klasse an. Dieses auf die eine Klasse angepasste Modell weist eine geringere Genauigkeit in der Schätzung der anderen Klassen auf. Um eine Überanpassung des Modells auf eine Klasse zu vermeiden, werden einer Klasse mit einem geringeren Anteil am Datensatz zufällig Beobachtungen kopiert und dem Datensatz hinzugefügt, bis ein Gleichgewicht erreicht wird. Der Trainingsdatensatz exklusive des Validierungsdatensatzes weist einen Anteil von 58 % an positiven Tweets und einen Anteil von 42 % an negativen Tweets auf. Die negativen Tweets werden wie beschrieben bis zum Erreichen eines ausgewogenen Verhältnisses erhöht.

5 Empirische Analyse

Die empirische Analyse dieser Arbeit unterteilt sich in zwei Bereiche. Im ersten Schritt werden die im vorherigen Kapitel vorgestellten Methoden genutzt, um eine themenspezifische Analyse zu ermöglichen. Im anschließenden zweiten Schritt wird die Regierungskommunikation analysiert, um für spezifische aus den gesamten Themen extrahierte Themen die Integration der Regierung innerhalb des Kommunikationsnetzes zu untersuchen.

5.1 Untersuchung der Modelle

Die Analyse der Regierungskommunikation basiert auf der Anwendung von zwei Modellen. Durch das Nutzen von *Top2Vec* werden die behandelten Themen aus dem COVID-19-Pandemie Datensatz extrahiert. Das für die Themenbestimmung verwendetet *Top2Vec*-Modell verfügt über verschiedene Parameter, die einen Einfluss auf die Genauigkeit des Modells haben. Über die Antworten auf die Tweets werden die Kommunikationsnetzwerke analysiert. Um eine präzisere Analyse der Kommunikationsnetzwerke durchführen zu können, wird ein Sentiment-Modell trainiert. Für das Sentiment-Modell kommen verschiedene Netzarchitekturen infrage, die miteinander verglichen werden. In dem folgenden Kapitel wird die Auswahl der Themen und des Sentiment-Modells erläutert.

5.1.1 Untersuchung der *Topic Modeling*-Modelle und Bestimmung der Themen

Zur Bestimmung der Themen während der COVID-19-Pandemie wurden vier verschiedene *Top2Vec*-Modelle berechnet. Variiert wurde zum einen die bei der Bestimmung des *Doc2vec* nötige Mindestanzahl der Häufigkeit eines Wortes. Hier wurde als Auswahlmöglichkeit eine Mindestanzahl von 30 oder 50 gewählt. Zum anderen wurde die Mindestclustergröße zwischen 200 und 300 Tweets pro Cluster unterschieden. Aufgrund des hohen Rechenaufwands war eine Bestimmung einer höheren Kombination an Parametern nicht möglich. Für alle vier berechneten Modelle wurde ein Silhouettenkoeffizient nahe null festgestellt, wobei das *Top2vec*-Modell mit einer Mindesthäufigkeit von 30 Wörtern und einer Mindestclustergröße von 300 Tweets die höchste Differenzierung zwischen den Clustern zeigte. Während für die Kombinationen 200 und 30, 200 und 50 und 300 und 50 Silhouettenkoeffizienten von $-0,085$, $-0,086$ und $-0,089$ bestimmt wurden, zeigte die Kombination mit einer Mindesthäufigkeit von 30 Wörtern und einer Mindestclustergröße von 300 Tweets einen Silhouettenkoeffizienten von $-0,078$.

Mit dieser Kombination an Parametern wurden 804 verschiedene Themen aus dem COVID-19-Pandemie-Datensatz extrahiert. Die Betrachtung der sich im semantischen Raum befindenden, auf fünf Dimensionen reduzierten Tweets in einer zweidimensionalen Darstellung zeigt, dass die Mehrzahl der Tweets einen geringen Abstand zueinander aufweisen (vgl. Abbildung 5 links). Die geringe Clusterbildung innerhalb des semantischen Raumes entsteht durch die geringe Länge der Tweets und die somit geringere Differenzierbarkeit des Themas. Die geringe Differenzierbarkeit äußerte sich ebenfalls mit dem Silhouettenkoeffizienten nah null. In der zweidimensionalen Ausgabe von UMAP sind jedoch bereits einzelne Cluster erkennbar.

Durch die Anwendung von HDBSCAN auf den auf fünf Dimensionen reduzierten Datensatz werden die Themen innerhalb des semantischen Raums bestimmt. In der zweidimensionalen Darstellung in Abbildung 5 (rechts) sind die verschiedenen Cluster dargestellt. Die hellblaue Punktwolke ist die Darstellung der als Rauschen identifizierten Tweets. Insgesamt sind 1.538.393 der 2.453.280 Tweets Rauschen.

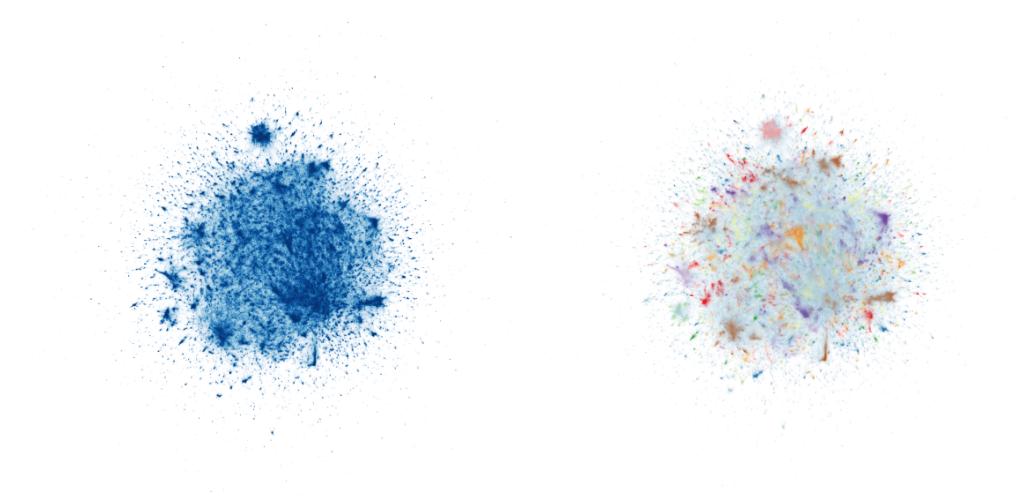


Abbildung 5: Zwei dimensionale Darstellung der auf fünf Dimensionen reduzierten Tweets (links) und zwei dimensionale Darstellung mit Clustern (rechts) (eigene Darstellung)

Das *Top2Vec*-Modell bestimmt für jedes Thema einen Themenvektor, der durch den semantischen Raum einem Wort entspricht. Zur besseren Einordnung der Themen wird nicht nur ein Wort als Beschreibung bestimmt, sondern die fünfzig Wörter, welche die geringste Distanz zum Themenvektor aufweisen. Während des Untersuchungszeitraums wurde eine große Bandbreite an Themen auf Twitter diskutiert, von Krankheitssymptomen (Thema 24) und aktuellen Infektionszahlen (Thema 1) über das Arbeitsrecht (Thema 40) und Wirtschaftshilfen (Thema 13) bis hin zu Hamsterkäufen (Thema 10) und Fußball (Thema 0). Die behandelten Themen ergeben eine breite Abbildung über die in der Gesellschaft diskutierten Themen. Mit der Höhe der Nummer eines Themas sinkt die Anzahl der Tweets, die dem Thema zugeordnet sind.

Der Fokus dieser Arbeit liegt auf dem Untersuchen der Regierungskommunikation zu wirtschaftlichen Themen. Um die Themen in der Analyse auf diese Themen beschränken zu können, wird wie bei der Themensetzung der Abstand zwischen Wörtern, Tweets und Themen im semantischen Raum genutzt. Um die Themencluster zu bestimmen, in denen Wirtschaftsthemen diskutiert werden, werden die fünf Themen ausgewählt, in denen der Themenvektor am nächsten an dem Wort „wirtschaft“ ist. Durch diese Berechnung konnten die Themen 52, 48, 76, 253 und 39 bestimmt werden (absteigend mit der Entfernung). Im Folgenden werden die Themen durch

Wordclouds untersucht. Die Größe der Wörter richtet sich nach der Entfernung des Wortvektors zum Themenvektor. Mit geringer Distanz steigt die Größe des Wortes in der Abbildung. In den fünf extrahierten Themen zum Wort „wirtschaft“ zeigen die Themen einen unterschiedlichen Schwerpunkt (vgl. Abbildung 6). Thema 52 und 48 liegen thematisch nah beieinander und eine eindeutige Differenzierung ist nicht möglich. Beide Themen behandeln den Einbruch der Nachfrage, wobei das Thema 52 in der Tendenz eine kurzfristigere Orientierung zeigt, während das Thema 48 tendenziell die fallende Nachfrage in der Weltwirtschaft behandelt (vgl. Feld et al., 2020, S. 15). Die Tweets des Themas 76 beschäftigen sich mit den Auswirkungen der Pandemie auf die Lieferketten und die entstehenden wirtschaftlichen Folgen. Im Rahmen der Unterstützung der Wirtschaft wurde in der Presse und in der Öffentlichkeit eine Kaufprämie für Autos diskutiert (vgl. Schlitt, 2020). Diese Thematik wird in dem Thema 253 zusammengefasst. Die vorgestellten Themen legen den Fokus auf die beobachteten Effekte und Maßnahmen, während die Tweets in dem Thema 39 die möglichen Folgen durch die COVID-19-Pandemie behandeln.



Abbildung 6: Die fünf Themen mit dem geringsten Abstand im semantischen Raum zum Wort "wirtschaft"

Mit der Annahme, dass Maßnahmen mit direkten wirtschaftlichen Effekten mit der Schließung von Geschäften oder Grenzen zusammenhängen, werden ebenfalls die fünf Themen (761, 459, 25, 514, 166), die am nächsten an dem Wortvektor „schlussungen“ liegen, in die Untersuchung mit eingeschlossen (vgl. Anhang: A.3). Von den fünf extrahierten Themen sind vier dem Oberthema Wirtschaft zuzuordnen. Das

Thema 459 hingegen behandelt die Schließung von Kindertagesstätten. Das Thema 761 weist den geringsten Abstand zum Wortvektor auf. Die Tweets dieses Themas behandeln die Schließungen des Einzelhandels und Insolvenzen beziehungsweise das Schutzschirmverfahren von GALERIA Karstadt Kaufhof GmbH (vgl. GmbH, 2020). Wie der Einzelhandel war auch die Gastronomie von Schließungen während des Untersuchungszeitraums betroffen. Die Themen 761 und 166 umfassen Tweets, welche diese Schließungen thematisieren. Innerhalb des Themas 514 werden die Tweets zusammengefasst, die sich mit Beratungskosten und Entschädigungen beschäftigen. Durch die Nähe im semantischen Raum zu den anderen vier Themen und zu dem Wort „schliessungen“ ist anzunehmen, dass dieses rechtliche Thema die Folge der Schließungen behandelt.

Die direkten wirtschaftlichen Maßnahmen führten zu Verlusten in den betroffenen Betrieben und deshalb beschloss die Bundesregierung die finanzielle Unterstützung von Betrieben, um die negativen Effekte der direkten Maßnahmen wie auch die der gesamtwirtschaftlichen Situation abzumildern (vgl. Feld et al., 2020, S. 93-94). Um die Kommunikation zu diesen Unterstützungen in der Analyse berücksichtigen zu können, werden die fünf Themen untersucht, die am nächsten an dem Vektor des Wortes „hilfen“ sind. Diese Themen sind 13, 209, 70, 38 und 260 (vgl. Anhang: A.3). Die fünf Themen können in drei Gruppen unterteilt werden. Die Themen 13 und 209 beschäftigen sich mit den von der Regierung zur Verfügung gestellten Soforthilfen für Selbständige und für kleine Betriebe (Thema 13) und dem Auftreten von Betrugsverdachtsfällen bei der Beziehung der Soforthilfen (Thema 209). Die Hilfspakete zur Unterstützung der Wirtschaft wurden nach den Beratungen der Bundesregierung bekannt gegeben. Die Kommunikation und Diskussion zu diesen Beratungstreffen und den veröffentlichten Unterstützungsmaßnahmen werden in dem Thema 70 zusammengefasst. Die Kommunikation zu den Hilfspaketen und zu ihrer Finanzierung betreffen die Tweets des Themas 38. Während die vier vorgestellten Themen Hilfsangebote für Unternehmen behandeln, fasst das Thema 260 die Diskussion über Hilfen auf individueller Ebene zusammen, wie beispielsweise das Anrecht auf Hartz IV nach einer Selbständigkeit.

Durch die Anwendung von *Top2Vec* auf den COVID-19 Datensatz konnte nicht nur eine Reihe von Themen extrahiert werden, in denen Maßnahmen der Regierung oder Folgen der Covid-19 Pandemie diskutiert werden, sondern auch ein Cluster (Thema 28), in dem eine maßnahmenunspezifische Ablehnung der Maßnahmen den Inhalt der Tweets bestimmt (vgl. Abbildung 7). Die Gründe der Ablehnung von Maßnah-

men und die darauf folgenden Demonstrationen und Ausschreitungen sind komplex und die Untersuchung derer ist nicht das Ziel dieser Arbeit (vgl. [tagesschau.de](https://www.tagesschau.de), 2020). Als ein Cluster von Tweets, in dem grundsätzlich ablehnende Äußerungen gegenüber den Maßnahmen der Regierung fallen, werden die mögliche Interaktion der Regierungskommunikation innerhalb dieses Clusters und das Kommunikationsnetzwerk jedoch untersucht.

Topic 28



Abbildung 7: Darstellung der Themen zur Demonstrationen gegen Einschränkungen der Grundrechten

5.1.2 Untersuchung verschiedener künstlicher neuronaler Netze zur Sentiment-Analyse

Zum Schätzen des Sentiments der Antworten innerhalb der extrahierten Themen soll ein künstliches neuronales Netz genutzt werden. Um eine hohe Übereinstimmung in der Struktur der Trainingsdaten und der Antworten erhalten zu können, werden verschiedene künstliche neuronale Netze trainiert und auf Basis der *Accuracy* miteinander verglichen. Trainiert und durch umfangreiches Hyperparametertraining untersucht werden ein CNN, ein LSTM, ein CNN-LSTM und ein LSTM-CNN. Aufgrund der hohen Hyperparametertrainingsdauer von CNN-LSTM und LSTM-CNN werden CNN und LSTM besonders umfangreich trainiert und die Parameteranzahl von CNN-LSTM und LSTM-CNN wird auf Basis der gewonnenen Erkenntnisse reduziert.

Wie in Kapitel 3.4 beschrieben wird zur Vermeidung von *Overfitting* und *Underfitting* Hyperparametertraining durchgeführt. Der Trainingsdatensatz wird vor der Durchführung des Hyperparametertrainings in einen Test- und in einen Trainingsdatensatz unterteilt. Der Testdatensatz wird nicht im Hyperparametertraining genutzt und bleibt dem Modell unbekannt. Nach abgeschlossenem Training wird der Testdatensatz verwendet, um die Modelle abschließend vergleichen zu können. Der Trainingsdatensatz wird im Hyperparametertraining eingesetzt. Um während des Hyperparametertrainings keine Parameterkombination, die auf den Trainingsdatensatz und somit auf

dem Modell bekannte Daten zugeschnitten ist, zu wählen, wird der Trainingsdatensatz in einen eigentlichen Trainings- und in einen Validierungsdatensatz unterteilt. Anhand des Validierungsdatensatzes werden die Parameterkombinationen verglichen. Der Testdatensatz, auf dem weder das Modell trainiert, noch die Parameter angepasst wurden, wird zur Überprüfung der Genauigkeit des Modells genutzt.

Bei der Anwendung eines CNN auf Textdaten werden die Bedeutungscluster extrahiert, die zur möglichst genauen Schätzung des wahren Werts beitragen (vgl. Kapitel 4.3.1). Die Architektur des trainierten CNN-Modells folgt Kim (2014). Anders als bei Kim (2014) und den Ergebnissen von Sosa (2017) folgend wird die *Word-Embedding*-Schicht in den Lernalgorithmus integriert. Nach der *Word-Embedding*-Schicht werden die Daten in drei parallel verlaufende Berechnungsabfolgen gegeben. Jede Berechnungsabfolge besteht aus einer *Convolutions*-Schicht, einer ReLU Schicht und einer folgenden *Maxpooling*-Schicht. Der Unterschied der Berechnungsabfolgen besteht in der Wahl der Breite der Filter in der *Convolution*-Schicht. Es werden Filter mit der Größe drei, vier und fünf gewählt (vgl. Kim, 2014, S. 3). Durch die Anwendung von *Maxpooling* auf die Ausgabe der *Convolution*-Schicht können die Daten in einem Tensor zusammengefasst und anschließend durch *Flattening* ein Vektor berechnet werden. Dieser Vektor wird zur Regularisierung in eine *Dropout*-Schicht gegeben. Die Ergebnisse des *Dropouts* werden in der *Dense*-Schicht¹⁶ zur zwei Neuronen zusammengefasst. Die Ausgabe der *Dense*-Schicht wird in die Softmax-Funktion gegeben und es wird die Wahrscheinlichkeit berechnet, ob die Beobachtung positiv oder negativ ist. Während des Trainings werden die Gewichte der drei *Convolution*-Schichten und die Gewichte der *Dense*-Schicht trainiert. Auf die Gewichte der *Dense*-Schicht wird zur Bestrafung von hohen Gewichtungen während des Lernalgorithmus die L_2 Parameter Regularisierung angewendet.

Beim Hyperparametertraining des CNN sind die Parameter die Anzahl der Filter der *Convolution*-Schicht, die Lernrate, die Batchgröße, die Höhe des Dropouts und die Höhe des genutzten L_2 -Regularisierungs-Parameters. Eine zu hohe Anzahl an Filtern führt zu einer Überanpassung des Netzes an die Trainingsdaten und erhöht die Trainingszeit. Dagegen führt eine zu geringere Anzahl an Filtern zu einer oberflächlichen Extraktion der Struktur der Daten. Die Lernrate bestimmt die Anpassung an das Optimum während des Lernalgorithmus. Eine zu starke Anpassung kann zu einem Überspringen des Optimums führen, während durch eine zu niedrige Anpassung das Optimum nicht erreicht wird (vgl. Kapitel 4.1.1). Die Lernrate bestimmt die

¹⁶Die *Dense*-Schicht entspricht der Schicht z (vgl. Kapitel 4.1)

Geschwindigkeit der Anpassung der Gewichte, die Batchgröße hingegen die Varianz in den Trainingsdaten. Eine hohe Batchgröße führt zu einer Verallgemeinerung der Informationsstruktur und zu einer Verzerrung der Schätzung. Eine geringe Batchgröße hingegen erhöht die Varianz der Daten und führt somit zu einer höheren Robustheit bei der Schätzung von bisher unbekannten Daten. Die Wahl einer zu niedrigen Batchgröße resultiert jedoch in einer starken Varianz der Daten und das Modell ist dann nicht in der Lage Strukturen zu extrahieren. Wie die Batchgröße wirkt auch die *Dropout*-Schicht auf die Varianz der Daten. Eine hoher *Dropout*-wert führt zu einer höheren Varianz in den Trainingsdaten während des Lernalgorithmus. Ist die Varianz der Trainingsdaten jedoch zu hoch, können keine Informationen über die Struktur der Daten extrahiert werden (vgl. Kapitel 4.3.3). Die durch die L_2 -Regularisierung durchgeführte Bestrafung von hohen Gewichten muss ebenfalls im Hyperparameter-training berücksichtigt werden. Ist die Bestrafung zu hoch, so ist das Modell während des Lernalgorithmus nicht in der Lage, die Varianz in den Daten abzubilden. Eine zu niedrige Bestrafung von hohen Gewichten führt im Gegensatz dazu jedoch zu einer zu genauen Anpassung des Modells an die Varianz in den Trainingsdaten und das Modell ist nicht mehr in der Lage Ausreißer als Einfluss auf die Gewichtung auszuschließen.

Durch das Hyperparametertraining soll die optimale Wahl der Parameterkombinationen angenähert werden. Das Finden der optimalen Kombination ist in der Theorie möglich, jedoch nicht praktikabel. Die Anzahl der möglichen Kombinationen an Parametern ist zu hoch und das Trainieren von neuronalen Netzen rechenintensiv. Beim Hyperparametertraining des CNN wird mit jeder möglichen Kombination der folgenden Werte ein Modell in 15 Epochen trainiert: Anzahl an Filtern (120, 150, 200), Lernrate (0, 000001, 0, 0001, 0, 001, 0, 01), Batchgröße (32, 68, 128), Dropout (0,5, 0,8) und L_2 -Regularisierungs-Parameters (0, 0001, 0, 001) (vgl. Abbildung 8). Die Validierungs-*Accuracy* der trainierten Modelle liegt zwischen 0,41678 und 0,73849. Die Auswertung der verschiedenen Kombinationen zeigt auf, dass das Modell sensibel auf eine Veränderung der Lernrate reagiert. Eine Lernrate oberhalb von 0,001 führt in allen trainierten Modellen zu einer Validierungs-*Accuracy* oberhalb von 70. Durch eine Lernrate kleiner 0,001 wird jedoch unabhängig von den anderen Parametern eine Validierungs-*Accuracy* unterhalb von 70 erreicht. Die Kombination von 120 Filtern, einer Lernrate von 0,001, einer Batchgröße von 32, einem Dropout von 0,5 und einem L_2 -Regularisierungs-Parameter von 0,0001 führt im Hyperparametertraining zur höchsten Validierungs-*Accuracy* von 0,73849.

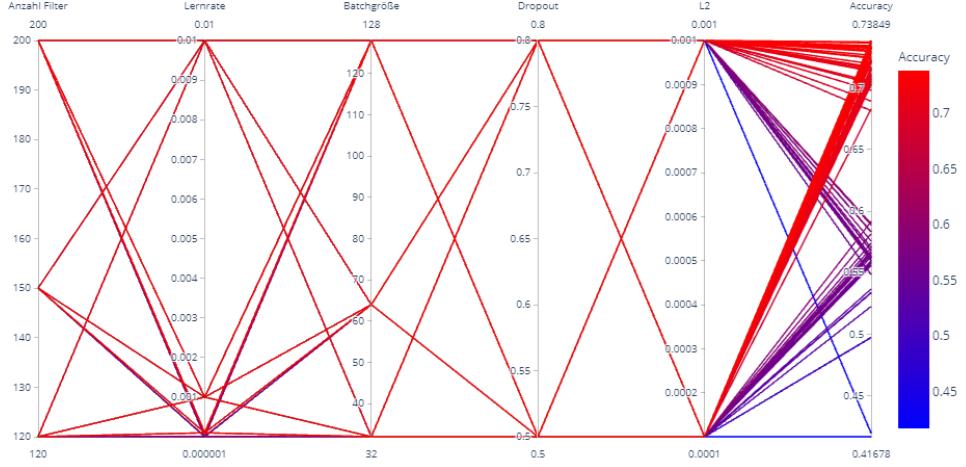


Abbildung 8: Hyperparametertraining CNN (eigene Darstellung)

Der Aufbau des in dieser Arbeit genutzten LSTM-Netzes orientiert sich an dem von Sosa (2017) oder auch Tang et al. (2016) genutzten Aufbaus. Nach der *Word-Embedding*-Schicht wandern die Daten in eine *Dropout*-Schicht. Von dieser aus werden die Daten in die LSTM-Zellen gegeben und anschließend durch die *Dense*-Schicht zusammengefasst. Die *Output*-Schicht gibt durch das Nutzen der Softmax-Funktion die Wahrscheinlichkeiten aus, dass die Tweets positiv oder negativ sind. Die Gewichte der LSTM-Schicht werden mit einer L_2 -Regularisierung bestraft. Im Vergleich zum Hyperparametertraining des CNN wird beim LSTM-Netz die Anzahl der LSTM-Zellen und nicht die Anzahl der Filter untersucht. Die Anzahl der LSTM-Zellen bestimmt, wie kleinteilig die Sequenz der Eingabe untersucht wird. Eine hohe Anzahl von LSTM-Zellen führt zu einer genaueren Anpassung an den Trainingsdatensatz. Eine niedrige Zahl erhöht die Robustheit der Schätzung gegenüber einer Varianz in den Daten. Das Hyperparametertraining des LSTM-Netzes wird mit 20, 60 oder 200 LSTM-Zellen durchgeführt, einer Lernrate von 0,0001 und 0,00001, einem *Dropout*-wert von 0,2, 0,4 und 0,6, einer möglichen Batchgröße von 32, 64, 128, 256 und 312 und einem L_2 -Regularisierungs-Parameter von 0,01 oder 0,0001 (vgl. Anhang A.2). Besonders eine geringe Lernrate führt zu einem Abfallen der Validierungs-*Accuracy*. Die höchste Validierungs-*Accuracy* von 0,74646 wird mit 20 LSTM-Zellen, einer Lernrate von 0,0001, einer Batchgröße von 312, einen *Dropout*-wert von 0,6 und einem L_2 -Regularisierungs-Parameter von 0,01 erreicht.

Wie in Kapitel 4.3 beschrieben, werden durch die Anwendung von CNN die Bedeutungscluster der Wörter extrahiert. Hierbei bleibt die Reihenfolge der Wörter unberücksichtigt. In LSTM-Schichten wird hingegen die Reihenfolge der Wörter zur

Extraktion von Informationen aus Texten genutzt. Durch die Kombination von CNN mit LSTM sollen beide Schwerpunkte der Architekturen verbunden und eine höhere Accuracy erreicht werden. In einem CNN-LSTM-Netz werden die Bedeutungscluster der Wörter extrahiert und anschließend durch die LSTM-Schicht die Zusammenhänge der Wörter über die Sequenz hinweg zur Schätzung des Sentiments genutzt (vgl. Sosa, 2017, S. 4). Der oben beschriebene Aufbau des CNN bleibt erhalten. Anstatt jedoch nach dem Maxpooling die Daten zu aggregieren, werden die Daten in eine *Dropout*-Schicht gegeben und anschließend die Zusammenhänge über die Sequenz der Sätze hinweg durch die LSTM-Schicht extrahiert. Die Ausgabe aus der LSTM-Schicht wird zusammengefasst und anschließend wird die Ausgabe mit der Softmax-Funktion berechnet. Der Vorteil einer getrennten Berechnung der LSTM-Schicht für die drei verschiedenen Filtergrößen der *Convolution*-Schicht ist, dass auch die Effekte der verschiedenen Längen der Sequenzen im Lernalgorithmus berücksichtigt werden. Im Hyperparametertraining werden die Anzahl der Filter (120, 150), die Anzahl der LSTM-Zellen (20, 30), die Lernrate (0, 0001, 0, 00001), der *Dropout*wert, die Batchgröße (32, 64) und der L_2 -Parameter untersucht (vgl. Anhang: A.2). Mit 120 Filtern je *Convolution*-Schicht, 20 LSTM Zellen, einer Lernrate von 0, 0001, einem *Dropout*wert von 0, 8, einer Batchgröße von 32 und einem L_2 Parameter von 0, 01 wird eine Validierungs-*Accuracy* von 0, 73809 erreicht.

Die Ausgabe einer LSTM-Schicht \mathbf{h}^t beinhaltet Informationen über die Ausgabe zu diesem Zeitpunkt der Sequenz, jedoch auch Informationen zu den vorherigen Zeitpunkten der Sequenz (vgl. Kapitel 4.3.2) (vgl. Sosa, 2017, S. 4). Durch die Kombination dieser Ausgabe mit einem CNN können die Bedeutungscluster der Wörter unter Berücksichtigung vorheriger Wörter extrahiert werden. Verschiebungen der Bedeutungen der Wörter durch vorherige Wörter im Satz werden durch ein LSTM-CNN berücksichtigt. Im LSTM-CNN Netz werden die *Word-Embeddings* als Eingabe in eine LSTM-Schicht gegeben. Die Ausgabe dieser Schicht wird dann in das bereits bekannte parallel verlaufende CNN gegeben. Wie beim CNN folgt auf das Zusammenfassen und Vektorisieren der Ausgabe eine *Dropout*-Schicht. Für die Berechnung der Ausgabe wird durch die *Dense*-Schicht zusammengefasst und anschließend mit der Softmatrix-Funktion die Ausgabe berechnet. Während des Hyperparametertrainings werden dieselben Parameterkombinationen untersucht wie beim CNN-LSTM. Im Unterschied zu CNN-LSTM zeigt sich, dass besonders eine niedrige Lernrate zu einer hohen Validierungs-*Accuracy* führt. Außerdem kommt es zu einer deutlich geringeren Streuung der Validierungs-*Accuracy*. Die Architektur des LSTM-CNN führt bei sämtlichen untersuchten Parameterkombinationen zu einer Validierungs *Accuracy*

größer 0,69. Die höchste Validierungs-*Accuracy* von 0,74407, wird mit der Kombination von 120 Filtern, 20 LSTM Zellen, einer Lernrate von 0.00001, einer Batchgröße von 32, einem *Dropout*-wert von 0,5 und einem L_2 -Parameter von 0,0001 erreicht.

Innerhalb des Hyperparametertrainings wird nicht die Methode des frühen Abbruches verwendet, um die Vergleichbarkeit der Modelle zu erhalten. Die Modelle aus dem Hyperparametertraining werden nun manuell untersucht, um einen Zeitpunkt für das frühe Abbrechen der Modelle festzulegen. Hierzu werden die Trainings- und Validierungs-*Accuracy* über die Epochen verglichen. Steigt die Trainings-*Accuracy* bei einem Abflachen der Steigung der Validierungs-*Accuracy* wird das Training beendet. In der Abbildung 9 ist das Training von CNN, LSTM, CNN-LSTM und LSTM-CNN (von links oben nach rechts unten) dargestellt.

Während des Trainings des CNN steigen die Trainings- und Validierungs-*Accuracy* bis zu der Epoche drei parallel zueinander an. Mit Begin der Epoche vier flacht die Steigung der Validierungs-*Accuracy* ab. Mit Epoche sechs steigt die Differenz der Trainings- und Validierungs-*Accuracy* deutlich an. Das Modell extrahiert nun weniger allgemeingültige Strukturen und das Training wird im finalen Modell hier abgebrochen. Beim Training des LSTM steigt die Trainings- wie auch Validierungs-*Accuracy* nach drei Epochen stark an. Nach vier Epochen wird bereits ein gleichbleibendes Niveau der Validierungs-*Accuracy* erreicht. Während des Trainings vom CNN-LSTM steigen die *Accuracys* vier Epochen simultan an. Ab der Epoche vier steigt die Trainings-*Accuracy* kontinuierlich bis zum Ende des Trainings an. Die Validierungs-*Accuracy* steigt von Epoche vier zu Epoche sechs weiter an. Aufgrund der großen Differenz in der Steigung zwischen Trainings- und Validierungs-*Accuracy* wird bereits bei Epoche vier das Training abgebrochen, um ein Overfitting zu vermeiden. Abschließend wird das Training des LSTM-CNN analysiert. Während des Trainings steigt die Trainings-*Accuracy* des CNN-LSTM-Modells kontinuierlich an, dahingegen fällt die Validierungs-*Accuracy* leicht über die Epochen hinweg. Das LSTM-CNN-Modell lernt zwar die Strukturen des Trainingsdatensatzes, jedoch keine allgemeingültigen Zusammenhänge.

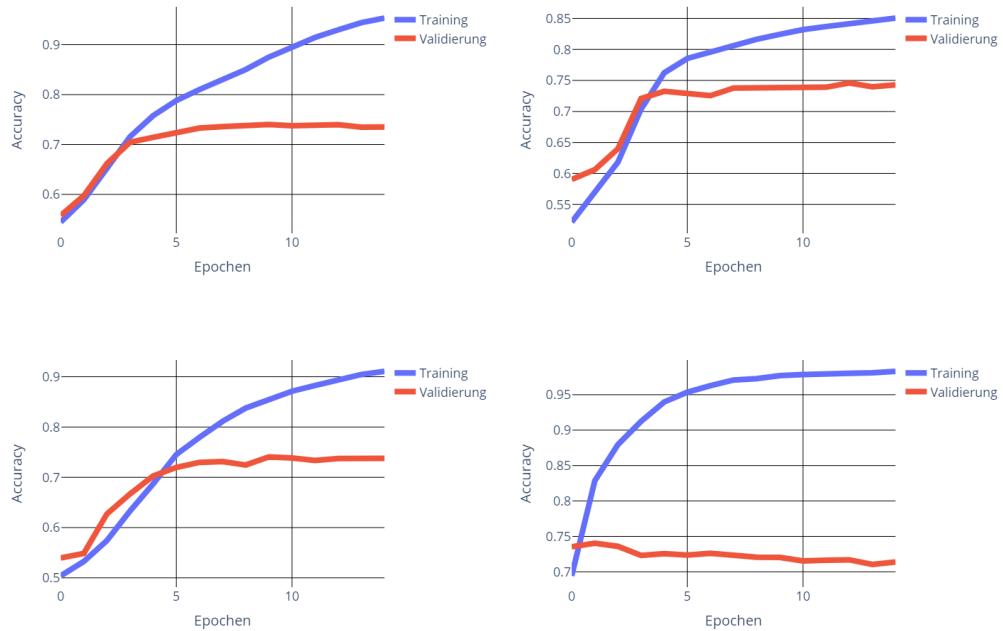


Abbildung 9: Vergleich der Trainings- und Validierungs-Accuracy der künstlichen neuronalen Netze zum Schätzen des Sentiments (eigene Darstellung)

Die Modelle werden auf dem Trainingsdatensatz trainiert und die Hyperparameter auf Basis des Validierungsdatensatzes angepasst. Um zu überprüfen, wie hoch die Genauigkeit der Modelle auf einem unbekannten Datensatz ist, wird die *Accuracy* bei der Anwendung der Modelle auf den Testdatensatz berechnet. Die vier trainierten Modelle erreichen nach umfangreichem Hyperparametertraining und einem guten Ergebnis auf den Validierungsdatensatz eine *Test-Accuracy* die deutlich unterhalb einer zufälligen Schätzung liegt (vgl. Tabelle 1). Diese trainierten Modelle können zum Bestimmen des Sentiments der Antworten nicht genutzt werden. Im folgenden Verlauf der Arbeit wird auf das Schätzen des Sentiments verzichtet.

Tabelle 1: Test-Accuracy der Modell zur Sentiment-Schätzung

Modell	Testaccuracy
CNN	0.2866
LSTM	0.2818
CNN-LSTM	0.2740
LSTM-CNN	0.3369

5.2 Regierungskommunikation

Zur Erfassung der Regierungskommunikation auf Twitter werden die Tweets der Regierungsaccounts aus dem erhobenen Twitter-Datensatz zur COVID-19-Pandemie extrahiert. Die Regierungsaccounts teilen sich in die 14 Accounts der Bundesministerien der Bundesrepublik Deutschland, in die Accounts der Bundesminister und Bundesministerinnen, in die des Robert Koch-Instituts und des Paul-Ehrlich-Instituts und in den der Kreditanstalt für Wiederaufbau (KfW) auf. Diese Liste wird durch Steffen Seibert und Ulrike Demmer in der Funktion als Sprecher und stellvertretende Sprecherin der Bundesregierung ergänzt. Ebenfalls in der Liste enthalten ist Helge Braun als Bundesminister für besondere Aufgaben und Leiter des Bundeskanzleramtes. Keinen persönlichen Twitter-Account hatten zum Zeitpunkt der Erhebung die Bundeskanzlerin und die folgenden Bundesminister und Bundesministerinnen: Christine Lambrecht¹⁷, Franziska Giffey¹⁸, Horst Seehofer¹⁹ und Gerd Müller²⁰. Das Robert Koch-Institut ist ein dem Bundesgesundheitsministerium untergeordnetes Bundesinstitut zur Überwachung und Erforschung von Infektionskrankheiten in Deutschland (vgl. Robert-Koch-Institut, 2020). In dieser Funktion stellt das Robert Koch-Institut den wissenschaftsbasierten Kommunikationskanal der Regierung dar. Ebenfalls dem Gesundheitsministerium untergeordnet ist das Paul-Ehrlich-Institut in Funktion als Bundesinstitut für Impfstoffe und biomedizinische Arzneimittel (vgl. BASIG, 1972). Die besondere Funktion des Paul-Ehrlich-Instituts liegt in der Zulassung und Beurteilung von Medikamenten und Impfstoffen. Dem Bundesfinanzministerium ist die Kreditanstalt für Wiederaufbau untergeordnet. Als Förderbank ist die KfW beispielsweise für das Anbieten von Krediten mit niedrigen Zinsen im Rahmen von Hilfsprogrammen während der COVID-19-Pandemie zuständig und wird deshalb in der Analyse der Regierungskommunikation berücksichtigt (vgl. Bundesministerium-für-Wirtschaft-und-Energie, 2020).

Im ersten Schritt der Analyse werden die unterschiedlichen Regierungsaccounts und die von der Regierung angesprochenen Themen untersucht. Anschließend wird die Interaktion der Regierung innerhalb der Themen analysiert.

¹⁷Bundesministerin der Justiz und für Verbraucherschutz (vgl. Bundespressamt, 2019a).

¹⁸Bundesministerin für Familie, Senioren, Frauen und Jugend (vgl. Bundespressamt, 2019b).

¹⁹Bundesminister des Innern, für Bau und Heimat (vgl. Bundespressamt, 2017a).

²⁰Bundesminister für wirtschaftliche Zusammenarbeit und Entwicklung (vgl. Bundespressamt, 2017b).

5.2.1 Kommunikation der Regierung

Im Zeitraum der Erhebung konnten für die oben beschriebenen Regierungsaccounts insgesamt 2.176 Tweets aus den gesammelten Tweets zur COVID-19-Pandemie extrahiert werden. Durch die Anwendung von *Top2Vec* wird jeder Tweet der Regierung als Rauschen erkannt oder einem Thema zugeordnet. Von den 2.176 Tweets kann für 2.045 Tweets ein Thema bestimmt werden. Von den insgesamt 804 Themen kommunizierte die Regierung zu 391 Themen. Wird die Anzahl der Tweets über alle Accounts hinweg kumuliert, dann zeigen sich die Schwerpunkte der Regierungskommunikation (vgl. Tabelle 2). Zum Thema 251 und zum Einsatz der Bundeswehr während der COVID-19-Pandemie wurden von der Regierung 104 Tweets veröffentlicht. Die Soforthilfen für kleine Betriebe und Selbständige wurden von der Regierung am zweithäufigsten angesprochen. Neben dem Thema 13 befindet sich keines der in Kapitel 5.1.1 extrahierten Themen unter den Themen mit der größten Beteiligung der Regierung. Die Regierung kommunizierte des Weiteren zu den aktuellen Infektionszahlen (Thema 2), zu den Fortschritten in der Impfstoffentwicklung (Thema 9), zu der Entwicklung der Corona-Warn-App (Thema 5), zu Handlungshilfen für die Landkreise (Thema 18), zur Unterstützung von Digitalisierungsprozessen, mit Aufrufen zu Plasma- oder Blutspenden (Thema 41), zu den Grenzschließungen und Sondergenehmigungen zur Einreise (Thema 283) und zu den Bund- und Länder-Konferenzen (Thema 135).

Tabelle 2: Die Zehn Themen mit der höchsten Anzahl an Regierungstweets

Thema	251	13	2	9	5	18	6	41	283	135
Anzahl	104	87	73	53	42	41	39	36	35	34

Die drei Regierungsaccounts mit den meisten Tweets sind die Accounts des Bundesministeriums für Gesundheit (BMG) mit 354 Tweets, des Bundesministeriums für Bildung und Forschung (BMBF) mit 216 Tweets und des Bundesministeriums der Verteidigung mit 185 Tweets (BMVg) (vgl. Tabelle 3). Unter den zehn Regierungsaccounts mit der höchsten Anzahl an Tweets sind acht offizielle Accounts von Ministerien, der Account des Regierungssprechers und der Account der KfW. Insgesamt weisen die persönlichen Accounts der Minister und Ministerinnen eine geringere Anzahl an Tweets zur COVID-19-Pandemie als die jeweiligen Accounts der Ministerien auf.

Tabelle 3: Die Zehn Regierungsaccounts mit den meisten Tweets

Twitteraccounts	Anzahl
BMG_Bund	354
BMBF_Bund	216
BMVg_Bundeswehr	185
BMWi_Bund	136
BMF_Bund	132
RegSprecher	132
bmel	128
AuswaertigesAmt	119
KfW	74
BMFSFJ	72

Werden die einzelnen Themen und die Anzahl der Tweets pro Regierungsaccount betrachtet, dann kann analysiert werden, welche Themen in der Kommunikation der einzelnen Accounts einen hohen Stellenwert einnehmen (vgl. Tabelle 4). Darüber hinaus ist anzunehmen, dass die Regierungsaccounts in diesen Themen eine höhere Integration im Kommunikationsnetz aufweisen. Einen eindeutigen Fokus in der Kommunikation zeigt das BMVg: 53 % der veröffentlichten Tweets des BMVg gehören dem Thema 251 an. Das Thema 251 umfasst Tweets, die sich während der COVID-19-Pandemie mit der Bundeswehr befassen, und ist das Thema mit der höchsten Anzahl an Tweets der Regierung (vgl. Tabelle 2). Ein Thema, welches über mehrere Regierungsaccounts hinweg eine hohe Anzahl an Tweets zeigt, ist das Thema 13. Das Thema 13, welches Tweets enthält, die sich mit Soforthilfen der Bundesregierung für Unternehmen beschäftigen, ist ein Schwerpunkt in der Kommunikation des Bundesministeriums der Finanzen (BMF), der KfW und des Bundesministeriums für Wirtschaft und Energie (BMWI). Das BMF veröffentlichte 27 % der Tweets in dem Thema 13. In der Kommunikation der KfW ist das Thema 13 mit einem Anteil von 31 % vertreten. Die Tweets des BMWIs sind zu 13 % dem Thema 13 zuzuordnen. Nach dem Thema 251 ist es das Thema mit der zweithöchsten Anzahl von Tweets der Regierung. Weitere Themen, die einen besonderen Fokus von einzelnen Regierungsaccounts zeigen, sind die Themen 9, 69, 38, 283 und 180. Das BMBF veröffentlichte 20 Tweets zum Thema 9, welches die Entwicklung des Impfstoffes betrifft, einen Anteil von 9 % in der Kommunikation des BMBF einnimmt und insgesamt das viertgrößte Thema in der Kommunikation der Regierung ist. Der Regierungssprecher, der in Personalunion auch der Sprecher der Bundeskanzlerin ist, kommunizierte in 14 % seiner Tweets zum Thema 69, welches Pressekonferenzen und Ansprachen der Bundeskanzlerin umfasst. Das BMF kommunizierte neben Thema 13 und der Umsetzung der

Soforthilfen mit einem Anteil von 7 % zu den Konferenzen, auf denen die Hilfspakete beschlossen wurden (Thema 38). Das Informieren über die zu Beginn der Pandemie beschlossenen Einreisebeschränkungen war ein Schwerpunkt der Kommunikation des Bundesministeriums für Inneres (BMI) und des Auswärtigen Amtes, welche jeweils 10 % und 13 % der Tweets zum Thema 283 veröffentlichten. Die gemeinsame Kommunikation des BMIs und des Auswärtigen Amtes führte dazu, dass das Thema 283 die neunthöchste Beteiligung der Regierung hat. Die Einreisebeschränkungen führten zu einer Verringerung des Angebots an Arbeitskraft in der Landwirtschaft, was in Thema 180 behandelt wird und einen Anteil von 11 % an der Kommunikation des Bundesministeriums für Ernährung und Landwirtschaft (BMWL) hat.

Tabelle 4: Die Zehn Regierungsaccounts mit den meisten Tweets zu einem Thema

Twitteraccounts	Anzahl
BMG_Bund	354
BMBF_Bund	216
BMVg_Bundeswehr	185
BMWi_Bund	136
BMF_Bund	132
RegSprecher	132
bmel	128
AuswaertigesAmt	119
KfW	74
BMFSFJ	72

Eine hohe Konzentration der Kommunikation eines Regierungsaccounts erhöhte die Wahrscheinlichkeit, im Kommunikationsnetzwerk dieses Themas einen Einfluss aufzuweisen. In der Abbildung 10 ist die Anzahl der Tweets jedes Accounts für die Themen dargestellt. Insgesamt zeigen einzelne Regierungsaccounts einen Themenschwerpunkt in der Kommunikation, jedoch ist mit Ausnahme der Kommunikation des BMVg die Kommunikation der Regierungsaccounts über die Themen stark gestreut. Hierbei ist zu berücksichtigen, dass das BMF und die KfW einen Schwerpunkt der Kommunikation auf das Thema 13 legen. Eine starke Streuung in der Kommunikation zeigen die Accounts des BMGs und des BMBFs. Diese weisen eine hohe Anzahl an Tweets auf, zeigen jedoch keinerlei Konzentration in der Kommunikation.

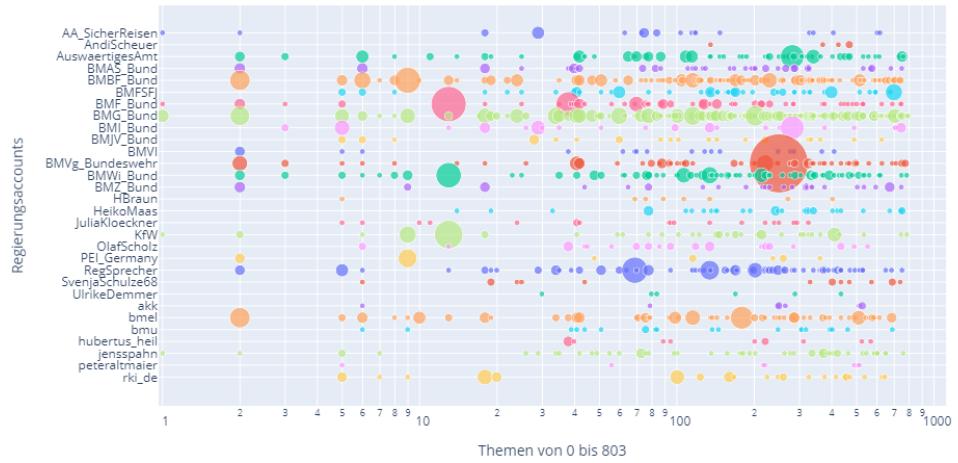


Abbildung 10: Verteilung der Regierungstweets (eigene Darstellung)

5.2.2 Integration der Regierung im Netzwerk

In Kapitel 5.1.1 wurden anhand des Abstandes der Themen im semantischen Raum zu den Wörtern „wirtschaft“, „schliessungen“ und „hilfen“ die Themen bestimmt, in denen die Kommunikation auf Twitter zu wirtschaftlichen Maßnahmen und Unterstützungen erfolgt. Um die Interaktion der Regierung zu diesen Themen untersuchen zu können, wird im Vorlauf analysiert, ob die Regierung zu diesen Themen kommuniziert hat. Fand keine oder eine geringe Kommunikation statt, so wird die Interaktion innerhalb des Kommunikationsnetzwerkes nicht untersucht. Eine Ausnahme bildet hier wie in Kapitel 5.1.1 beschrieben das Thema 28. Von den 16 extrahierten Themen kommunizierte die Regierung zu 13 Themen (vgl. Tabelle 5). Die untersuchten Regierungsaccounts kommunizierten nicht zum Thema 209 und zu den Betrugsverdachtsfällen im Rahmen der Soforthilfe, nicht zum Thema 761 und zur Schließung von Geschäften und zur Insolvenz der GALERIA Karstadt Kaufhof GmbH und nicht zum Thema 52, welches den Rückgang der Nachfrage beschreibt.

Tabelle 5: Anzahl an Regierungstweets in den ausgewählten Themen

Thema	13	38	70	76	25	48	260	39	28	459	166	253	514
Anzahl	87	30	19	15	6	5	4	4	3	2	1	1	1

Von den extrahierten Themen sind die Themen 13 und 38 diejenigen, welche die höchste Regierungskommunikation aufweisen (vgl. Abbildung 11). Wie bereits in Kapitel 5.1.1 beschrieben, stellt die Kommunikation zur Unterstützung von Unternehmen

(Thema 13) mit 87 Tweets einen Schwerpunkt in der Regierungskommunikation dar. Das Thema 38, die Finanzierung der Unterstützungen für die Wirtschaft, ist für das BMF das Thema mit der zweithöchsten Beteiligung und weist insgesamt 30 Tweets der Regierung auf (vgl. Kapitel 5.1.1). Die restlichen Themen zeigen eine geringe Anzahl an Tweets von Regierungsaccounts und sind aufgrund dessen wenig im Kommunikationsnetzwerk integriert und werden im Folgenden nicht weiter analysiert. Die Kommunikation zu den ausgewählten Themen wird hauptsächlich durch die Accounts des BMFs, des BMWIs, des BMGs, von Olaf Scholz und des KfK durchgeführt.

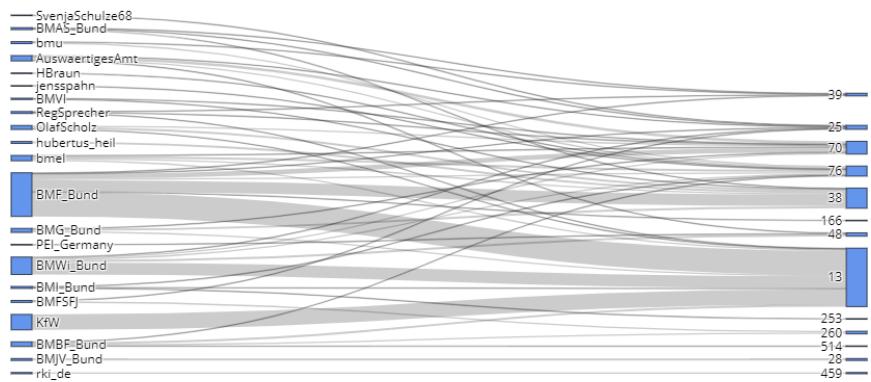


Abbildung 11: Verteilung der Regierungstweets auf die extrahierten Themen (eigene Darstellung)

Für die Netzwerkanalysen werden die Antworten auf die Tweets für jeden Account kumuliert. Das Sentiment jeder Antwort kann aufgrund der geringen Test-Accuracy der Sentiment-Modelle nicht geschätzt werden. Die Größe der Knoten und der Beschriftung ist von der gewählten Metrik abhängig (vgl. Kapitel 3.5). Wie bereits erwähnt, werden Privatpersonen in dieser Arbeit anonymisiert. Personen außerhalb des öffentlichen Lebens erhalten den Namen „Privatperson“. Handelt es sich bei einem Account um einen kleinen Betrieb wird der Accountname in „Betrieb“ verändert.

Aus dem COVID-19-Pandemie-Datensatz wurden durch *Top2vec* 15.794 Tweets dem Thema 13 zugeordnet. Nach fünf Tagen wurden kumuliert über alle Tweets hinweg 9.162 Antworten erfasst. In Abbildung 12 ist das Kommunikationsnetzwerk zu Thema 13 dargestellt. Es wurde mit einem *Edge-Cut* von 0,3 und einer Anzahl von 2.000 Iterationen erstellt (vgl. Kapitel 4.4). Für die Aufteilung der unterschiedlichen Phasen innerhalb der Optimierung des *OpenOrd*-Algorithmus wurden für *liquid*, *expansion*, *cool-down*, *crunch* und *simmer* die Anteile von 25 %, 25 %, 25 %, 10 % und

15 % gewählt. Diese Anteile entsprechen dem von Martin et al. (2011) empfohlenen Verhältnis. Durch den verwendeten Algorithmus befinden sich Knoten, auf die von ähnlichen Nutzern geantwortet wurde, im Netzwerk nah beieinander.

Im linken Netzwerk wurde die *Eigenvector-Centrality* verwendet, um die Bedeutung der einzelnen Accounts darzustellen zu können. Im rechten Netzwerk hingegen wurde die *Degree-Centrality* genutzt. In beiden Netzwerken zeigen die Accounts der Investitionsbank Berlin (PR_ibb), des Ministeriums für Wirtschaft, Innovation, Digitalisierung und Energie des Landes Nordrhein-Westfalen (WirtschaftNRW) und der Account der Tagesschau (tagesschau) den höchsten Einfluss. Die Rangfolge der Bedeutung veränderte sich jedoch durch die Metrik. Die Tagesschau verzeichnete im Netzwerk die meisten Antworten, zeigt jedoch eine geringere *Eigenvector-Centrality* als PR_ibb und WirtschaftNRW. Aus dieser Verschiebung kann geschlussfolgert werden, dass auf die Tweets der Tagesschau von einer Vielzahl verschiedener Accounts geantwortet wird, diese jedoch eine geringere Aktivität in der gesamten Diskussion zu der Thematik der Soforthilfen für kleine Unternehmen und Selbständige zeigten. Insgesamt bedeutet ein höherer Rang in der *Eigenvector-Centrality* als in der *Degree-Centrality* eine größere Bedeutung in der Kommunikation in den einzelnen Clustern des Kommunikationsnetzwerkes. Ein Account mit einer hohen *Degree-Centrality* hat im gesamten Netzwerk eine hohe Anzahl an Verbindungen. Die Accounts mit einer hohen *Eigenvector-Centrality* hingegen sind besonders einflussreich im Netzwerk.

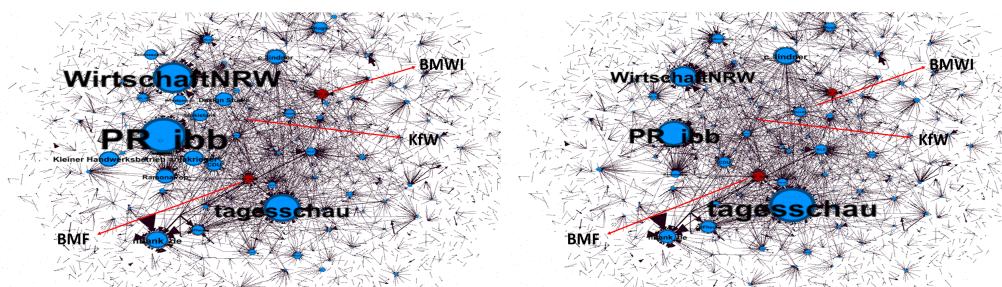


Abbildung 12: Kommunikationsnetzwerk des Themas 13 nach Eigenvector-Centrality und Degree-Centrality (eigene Darstellung)

Die zehn einflussreichsten Accounts pro Metrik im Kommunikationsnetzwerk des Themas 13 unterteilen sich in vier verschiedene Gruppen. Die erste Gruppe der Accounts sind staatliche oder halbstaatliche Accounts wie die Förderbanken Berlins (PR_ibb) oder Niedersachsens (nbank_de) und Accounts von Landes- oder Bundesministerien wie das BMF, das Ministerium für Wirtschaft des Landes Nordrhein-Westfalen oder das bayrische Wirtschaftsministerium (BayStMWi). Bei der zweiten Gruppe handelt es sich um Accounts von Personen im öffentlichen Leben wie Christian Lindner²¹ (c_lindner), Ramona Pop²² (RamonaPop) oder Klaus Lederer²³ (klauslederer). Die dritte Gruppe von Accounts innerhalb des Netzwerkes sind Accounts der Presse wie die Tagesschau, ZDF heute (ZDFheute) oder die Bild Zeitung (BILD). Die letzte der vier Gruppen fasst die Accounts von Privatpersonen und kleinen Betrieben zusammen.

Im Thema 13 wurden hauptsächlich Regierungstweets des BMFs, des BMWIs und der KfW zugeordnet. Für die Accounts des BMFs und des BMWIs konnte ein relevanter Einfluss auf das Kommunikationsnetzwerk festgestellt werden. Der Account des BMFs weist die achthöchste *Degree-Centrality* und den Rang 17 in der *Eigenvector-Centrality* auf (vgl. Tabelle: 8). Das BMWI erlangte mit Rang 16 eine geringere Bedeutung bei der *Degree-Centrality*, jedoch mit Rang 16 bei der *Eigenvector-Centrality* eine ähnliche Bedeutung wie das BMF. Das BMWI zeigt ein ausgeglichenes Verhältnis der *Degree-Centrality* und der *Eigenvector-Centrality*. Das BMF hingegen hat einen geringeren Einfluss auf Accounts, die sich intensiver mit der Thematik auseinandersetzen, als aufgrund der *Degree-Centrality* erwartbar. Die KfW zeigt eine deutlich geringere Integration in das Netzwerk als das BMF oder das BMWI und wird deshalb nicht genauer untersucht.

Tabelle 6: Die Zehn Accounts im Kommunikationsnetzwerk des Thema 13 nach Metrik

Account Eigenktor	Eigenvector centrality	Account Degree	Degree centrality	Account Betweenes	Betweeness centrality
PR_ibb	1.000000	tagesschau	404	PR_ibb	2003.5
WirtschaftNRW	0.880915	PR_ibb	290	Betrieb Holz	1138.5
tagesschau	0.792434	WirtschaftNRW	237	WirtschaftNRW	740.5
nbank_de	0.481416	c_lindner	201	WKogler	111.0
anjakrieger	0.439675	nbank_de	154	BayStMWi	110.0
Betrieb Holz	0.439675	ZDFheute	141	BMWi_Bund	101.0
c_lindner	0.403478	CDU	117	Privatperson	96.0
RamonaPop	0.401809	BMF_Bund	115	klauslederer	66.0
Betrieb Design	0.393141	BILD	112	sarfeld	65.0
CDU	0.359497	drumheadberlin	110	RegierungBW	63.0

²¹Christian Lindner: Bundesvorsitzender der Freien Demokratischen Partei (vgl. Deutscher-Bundestag, 2018).

²²Ramona Pop: Senatorin für Wirtschaft, Energie und Betriebe in Berlin (vgl. Berlin, 2019).

²³Klaus Lederer: Kultur- und Europasenator von Berlin (vgl. Berlin, 2017).

Für die Analyse der Kommunikation von BMF und BMWI im Thema 13 wird das Netzwerk auf die Kommunikation beschränkt, die bis zu drei Knoten von BMF und BMWI entfernt ist. Es werden also die Verbindungen bis zu einem dritten Grad betrachtet (vgl. Abbildung 14). Die Größe der Knoten ist durch die *Eigenvector-Centrality* bestimmt. Die Integration der Accounts im Netzwerk zeigt ähnliche Strukturen. Accounts, die mit dem BMF und dem BMWI kommunizierten, kommunizierten auch mit der Tagesschau, den Investitionsbanken von Niedersachsen und dem Ministerium für Wirtschaft des Landes NRW. Der auffälligste Unterschied liegt in dem Fehlen der Investitionsbank Berlins im Netzwerk des BMWIs. Wird die Position der Accounts im Netzwerk betrachtet, ist auffällig, dass der Account des BMF sich mittiger im Netzwerk befindet und räumlich den verschiedenen Accounts der Banken näher steht als der Account des BMWIs. Dieser befindet sich am Rand des Kommunikationsnetzwerkes.

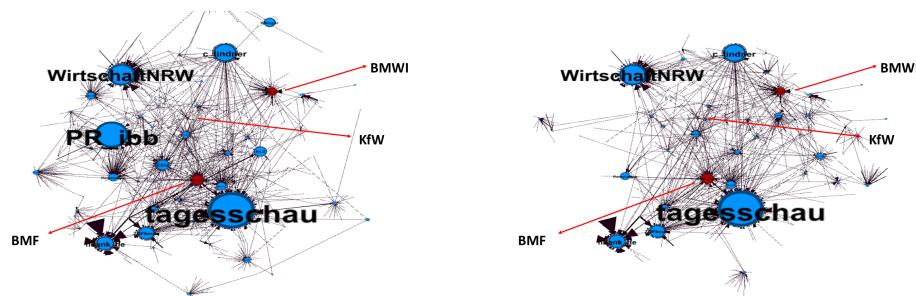


Abbildung 13: Kommunikationsnetzwerk von BMF und BMWI innerhalb des Themas 13 (eigene Darstellung)

Das Thema 38 enthält 8.676 Tweets und 5.965 Antworten wurden innerhalb von 5 Tagen nach dem Veröffentlichen der Tweets kommentiert. Zum Erstellen des Kommunikationsnetzwerkes wurden die gleichen Parameter wie zur Erstellung des Kommunikationsnetzwerkes des Themas 13 gewählt. In Abbildung 14 wird das Netzwerk des Themas 38 dargestellt. Die Größe der Knoten ist durch die *Eigenvector-Centrality* bestimmt. Innerhalb des Netzwerkes kann im Vergleich zum Thema 13 eine verstärkte Clusterbildung festgestellt werden. Die Accounts mit hoher *Eigenvector-Centrality* verteilen sich über das Netzwerk und bilden eine Art Kommunikationsmittelpunkt in

ihren Clustern. In einem Cluster können sich auch mehrere einflussreiche Accounts befinden. Einen hohen Einfluss im Netzwerk zeigen Accounts der Presse wie die Tagesschau, ZDF heute und das heute journal (heutejournal). Neben der Presse sind politische Accounts wie beispielsweise der Twitteraccount von Markus Söder²⁴ (Markus_Soeder) einflussreich. Diese zwei Accountarten sind im Netzwerk in Cluster unterschieden, wobei die Presse in einem größeren Teil des Netzwerkes verbreitet ist und das zentrale Cluster in der Mitte darstellt. Das Pressecluster kann in Fernsehen, Zeitung und Onlinemedien weiter unterteilt werden. Der Account von Markus Söder befindet sich außerhalb des politischen Clusters im Pressecluster und bildet dort mit dem Account von Andreas Scheuer²⁵ zusammen ein kleines politisches Cluster zwischen einem Cluster der Springer-Presse unterhalb und des Accounts der Frankfurter Allgemeinen Zeitung oberhalb. Im politischen Cluster befindet sich ein kleineres leicht abgegrenztes Cluster der Diskussion in Österreich um Sigrid Maurer²⁶. Rechts daneben liegt ein Teil des politischen Clusters mit der CDU und Olaf Scholz als einflussreichste Accounts. Diesem Cluster wird auch der Account des BMF zugeordnet. Rechts oberhalb dieses Clusters befindet sich die Kommunikation um den Account der Bundestagsfraktion der AfD (AfDimBundestag) mit Malte Kaufmann²⁷ (MalteKaufmann) und Beatrix von Storch²⁸ (Beatrix_vStorch). Zwischen diesen Kommunikationsnetzwerken und dem Cluster der Presse liegen die Accounts des BMGs und von Paul Ziemiak²⁹ (PaulZiemiak). Von den ausgewählten Regierungsaccounts befinden sich wie beschrieben die Accounts des BMGs, des BMFs und von Olaf Scholz³⁰ mit einem relevanten Einfluss im Netzwerk. Während die Accounts des BMGs und des BMFs gut im Netzwerk integriert sind, liegt der Account von Olaf Scholz außerhalb des Netzwerkes und ist nicht deutlich integriert. Die Accounts des BMGs und des BMFs befinden sich räumlich in der Nähe der drei Accounts der Nachrichtensendungen heute journal, Tagesschau und ZDF heute. Entsprechend besteht eine Überschneidung in der angesprochenen Nutzergruppe. Anders als bei Thema 13 wird hier das Netzwerk mit Knoten nach *Degree-Centrality* nicht dargestellt, da es nicht zu deutlichen Verschiebungen kommt.

²⁴Bayerische Ministerpräsident (vgl. Bayerische-Staatskanzlei, 2020).

²⁵Bundesminister für Verkehr und digitale Infrastruktur (vgl. Bundespressamt, 2020b).

²⁶Abgeordnete vom Nationalrat (Die Grünen) (vgl. Republik-Österreich, 2020).

²⁷Oberbürgermeisterkandidat der AfD in Stuttgart (vgl. Südwestrundfunk, 2020).

²⁸Mitglied des Bundestages (AfD) (vgl. Deutscher-Bundestag, 2020a).

²⁹Generalsekretär der CDU (vgl. Christlich-Demokratische-Union-Deutschlands, 2018).

³⁰Bundesminister der Finanzen (vgl. Bundespressamt, 2019c).

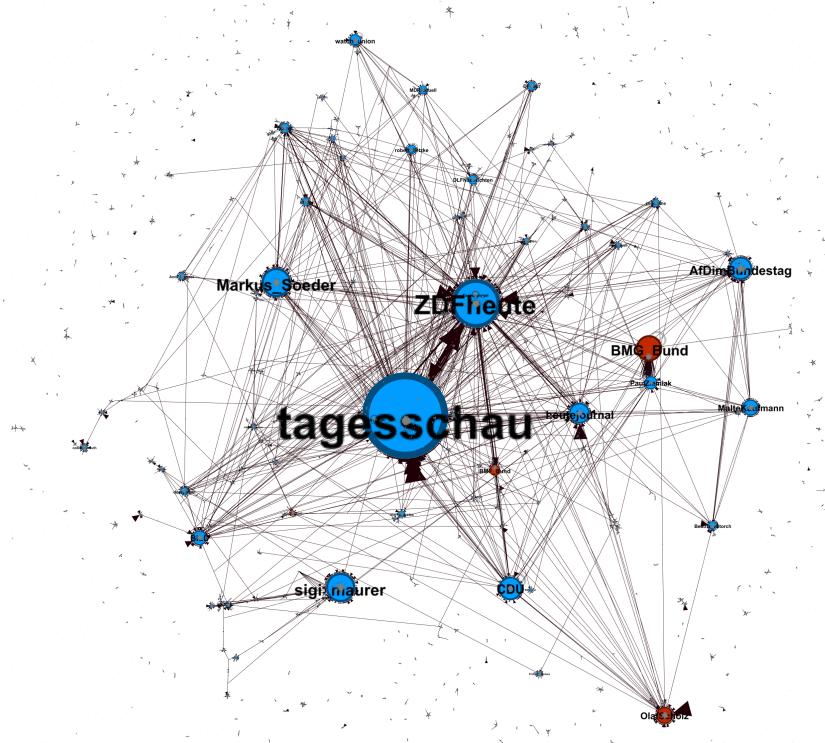


Abbildung 14: Kommunikationsnetzwerk des Themas 38 nach Eigenvector-Centrality (eigene Darstellung)

Wie bereits in der visuellen Darstellung festgestellt, zeigen die Accounts der Tagesschau und von ZDF heute den höchsten Einfluss im Netzwerk. Die Kombination aus einem hohen Rang im *Eigenvector-Centrality*- und *Degree-Centrality*-Ranking führt zu der herausgehobenen Stellung im größten Cluster, welches Verbindungen zu allen anderen Clustern aufweist. Die Accounts von Markus Söder und Sigrid Maurer weisen die dritt- beziehungsweise viertgrößte *Eigenvector-Centrality* und in der Rangfolge umgekehrte *Degree-Centrality* auf. Diese hohen Ränge bestätigen die Bedeutung dieser Accounts in ihren Unterclustern.

Insgesamt zeigen die zehn Accounts mit der höchsten *Eigenvector-Centrality* eine geringere *Eigenvector-Centrality* als die des Themas 13 (vgl. Tabelle 7). Die Accounts mit einem Einfluss innerhalb des Netzwerkes liegen weiter auseinander als im Thema 13. Daraus folgt, dass es innerhalb des Netzwerkes verstärkt zu einer Auf trennung in kleinere Netzwerke kommt. Diese Analyse wird durch die leicht erhöhte *Degree-Centrality* im Vergleich zum Thema 13 unterstützt. Die Anzahl der Verbindungen steigt leicht, die Anzahl der Accounts mit ebenfalls hoher Anzahl an Verbindungen in der Umgebung sinkt hingegen. Das Verhalten wurde zuvor bei der Beschreibung des Netzwerkes ebenfalls festgestellt.

Werden die zehn Accounts mit der höchsten *Betweenness-Centrality* betrachtet, dann

zeigt sich, dass die einzelnen Cluster des Netzwerkes größtenteils über private Accounts miteinander verbunden sind. Diese Accounts kommunizieren beispielsweise sowohl mit dem Cluster um die AfDimBundestag und Malte Kaufmann als auch mit dem BMG und Paul Ziemiak. Der Account der CDU weist die höchste *Betweenness-Centrality* auf. Die *Betweenness-Centrality* lässt sich nur bedingt über verschiedene Netzwerke vergleichen, da eine Abhängigkeit von der Größe der Netzwerke besteht.

Im Netzwerk zeigen drei der ausgewählten Regierungsaccounts Einfluss. Das BMG und das BMF befinden sich innerhalb des Netzwerkes, wobei das BMG die höchste *Eigenvector-Centrality* hat. Der Account von Olaf Scholz weist eine geringere *Eigenvector-Centrality* als der Account des BMGs auf, aber eine höhere *Eigenvector-Centrality* als der Account des BMFs. Dies bestätigt die Erkenntnis der visuellen Analyse der besseren Integration des BMGs und des BMFs im Netzwerk im Vergleich zu dem Account von Olaf Scholz, wobei zu berücksichtigen ist, dass das BMF insgesamt weniger Verbindungen im Netzwerk aufweist. Der im Verhältnis zu den anderen Regierungsaccounts hohe Einfluss des BMGs im Kommunikationsnetzwerk zu Hilfspaketen und Konjunkturpaketen ist auffällig, da dieses Thema nicht in den Aufgabenbereich des BMGs fällt. Die Tweets des BMGs innerhalb des Themas 38 wurden manuell überprüft. Die vom BMG veröffentlichten Tweets setzten sich mit den in einem Hilfspaket integrierten finanziellen Unterstützungen für Krankenhäuser auseinander.

Tabelle 7: Die Zehn Accounts im Kommunikationsnetzwerk des Themas 38 nach Metrik

Account Eigenktor	Eigenvector centrality	Account Degree	Degree centrality	Account Betweenes	Betweeness centrality
tagesschau	1.000000	tagesschau	579	CDU	110
ZDFheute	0.564743	ZDFheute	326	Privatperson	25
sigi_maurer	0.341193	Markus_Soeder	197	Privatperson	15
Markus_Soeder	0.340431	sigi_maurer	196	Privatperosn	8
BMG_Bund	0.302326	AfDimBundestag	153	Privatperson	8
CDU	0.273022	heutejournal	139	Paritaet	6
AfDimBundestag	0.267491	BMG_Bund	125	Privatperson	5
heutejournal	0.240377	CDU	114	Privatperson	4
OlafScholz	0.196658	OlafScholz	114	Privatperson	3
MalteKaufmann	0.186900	MalteKaufmann	108	Privatperson	3

Werden die Kommunikationsnetzwerke um die einzelnen Regierungsaccounts des BMGs, des BMFs und des Accounts von Olaf Scholz auf alle Verbindungen bis zum dritten Grad reduziert, dann zeigen sich geringe Unterschiede in der Struktur des Netzwerkes (vgl. Anhang A.4 - A.4). Jede der drei Netzwerkstrukturen offenbart eine hohe Übereinstimmung mit der Struktur des Gesamtnetzwerkes. Das BMG weist keine

Verbindung dritten Grades zu dem Account von Olaf Scholz auf, jedoch zu dem Account des BMFs. Entsprechend besteht die Verbindung auch in die entgegengesetzte Richtung nicht. In keinem der drei Netzwerke existiert eine Verbindung zu dem Cluster um Sigrid Maurer. Wie oben beschrieben, bestehen Verbindungen zwischen dem BMG und dem Cluster um die AfD. Über Verbindungen zum Account von Beatrix von Storch ist der Account von Olaf Scholz ebenfalls mit diesem Cluster verbunden. Das BMF weist diese Verbindung nicht auf.

Abschließend wird das Kommunikationsnetzwerk des Themas 28 analysiert (vgl. Abbildung 15). Dieses Thema ist wie in Kapitel 5.1.1 beschrieben nicht durch die Nähe zu Suchwörtern, sondern aufgrund der allgemeinen Ablehnung von Maßnahmen innerhalb dieses Netzwerkes ausgewählt worden. Innerhalb des Themas 28 wurden drei Tweets von dem Account des Bundesministerium der Justiz und für Verbraucherschutz (BMjV) erfasst (vgl. Tabelle 5). Aufgrund dieser geringen Menge an Tweets ist eine Integration des Accounts des BMjV nicht gegeben. Insgesamt wurden 11.078 Tweets und 14.247 Antworten zu diesem Thema gefunden. Anders als bei den oben behandelten Themen übertrifft die Anzahl an Antworten die Anzahl der Tweets. Die Motivation und/oder der Anreiz für die Beteiligung an der Diskussion ist höher als bei den bisher betrachteten Themen. Mit denselben Parametern zur Bestimmung des Netzwerkes wie zuvor wurde Abbildung 15 erstellt. Die Größe der Knoten ist durch die *Eigenvector-Centrality* bestimmt. Eine weitere Auffälligkeit ist der deutlich höhere Einfluss von Privatpersonen im Netzwerk.³¹

Im Vergleich zu den oben analysierten Themen kommt es zu einer verstärkten Clusterbildung. Anders als in Thema 38 unterteilt sich das Netzwerk nicht in ein großes Pressecluster und ein politisches Cluster, die Bedeutung des Pressclusters nimmt stattdessen ab und es kommt zu mehreren kleinen Clustern. Das Pressecluster befindet sich im unteren Teil des Netzwerkes. In der räumlichen Nähe dieses Clusters liegt der Account von Christian Lindner, welcher in diesem Netzwerk den höchsten Einfluss auf die Kommunikation aufweist. Oberhalb des Accounts von Christian Lindner trennt sich das Netzwerk in verschiedene politische Cluster auf. Im rechten oberen Bereich des Netzwerkes befindet sich ein Cluster von AfD Politikern wie Alice Weidel³² (Alice_Weidel), Joerg Meuthen³³ (Joerg_Meuthen) und Malte Kaufmann. Dieses

³¹Die Deklarierung eines Accounts als Privatperson wurde nach bestem Wissen und Gewissen vorgenommen. In diesem Kommunikationsnetzwerk zeigten jedoch verschiedene Accounts eine nach außen gerichtete Kommunikation bei gleichzeitigem privaten Charakter. Um hier eine bessere Differenzierung vornehmen zu können, wurden Accounts nicht als privat gekennzeichnet, wenn ein angestrebtes Verwertungsinteresse der eigenen Person festgestellt wurde.

³²Mitglied des Bundestages (AfD) (vgl. Deutscher-Bundestag, 2017).

³³Bundessprecher der AfD (vgl. Alternative-für-Deutschland, 2020).

Cluster der AfD ist über den Account von Niko Härtling³⁴ (nhaerting) mit dem Pressecluster und dem Account von Christian Lindner verbunden. Links oberhalb des AfD-Clusters befindet sich ein weiteres Cluster mit Politikern. Innerhalb dieses Clusters haben die Accounts von Cem Özdemir³⁵ (cem_oezdemir), Renate Künast³⁶ (RenateKuenast) und Konstantin Kuhle³⁷ (KonstantinKuhle) den höchsten Einfluss.

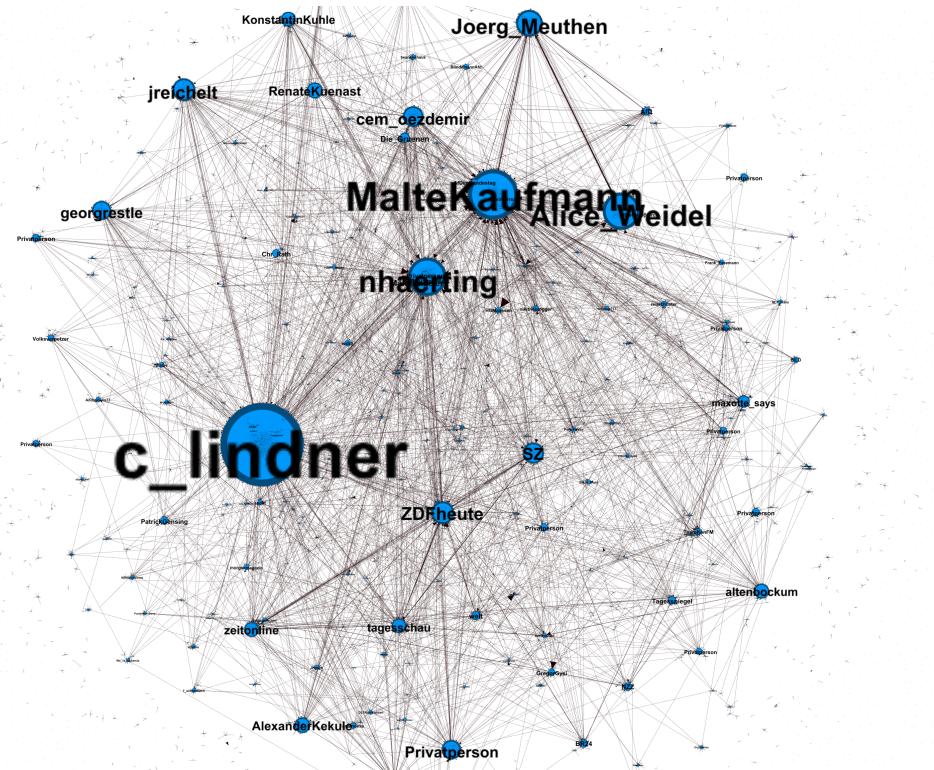


Abbildung 15: Kommunikationsnetzwerk des Themas 28 nach Eigenvector-Centrality (eigene Darstellung)

Neben dem Account von Christian Lindner zeigen im Netzwerk die Accounts der AfD Politiker Alice Weidel, Jörg Meuthen und Malte Kaufmann einen hohen Einfluss (vgl. Tabelle: 8). Die *Eigenvector-Centrality* ist ähnlich gestreut wie im Thema 38 und ergänzt die Einschätzung der verstärkten Clusterbildung in der visuellen Darstellung. Die Accounts der Presse zeigen im Vergleich zu den bisher analysierten Themen eine geringere *Eigenvector-Centrality* und somit einen geringeren Einfluss im Netzwerk auf. Bei der Betrachtung der *Degree-Centrality* kommt es zu geringen Verschiebungen in der Rangfolge. Die auffälligste Verschiebung ist die deutlich niedrigere *Degree-Centrality* im Vergleich zur *Eigenvector-Centrality* von Niko Härtling. Dieses Verhältnis zeigt eine überdurchschnittliche Verbindung von Niko Härtling zu anderen Accounts mit einem hohen Einfluss. Dieses und die in der visuellen Darstellung fest-

³⁴ Rechtsanwalt der Klagen gegen verschiedene Maßnahmen eingereicht hat (vgl. Spallek, 2020).

³⁵ Politiker der Partei Bündnis 90/Die Grünen (vgl. Britannica, 2020).

³⁶ Mitglied des Bundestages (Bündnis 90/Die Grünen) (vgl. Deutscher-Bundestag, 2020c).

³⁷ Mitglied des Bundestages (FDP) (vgl. Deutscher-Bundestag, 2020b).

gestellte Eigenschaft als Verbindungsknoten zwischen dem Pressecluster und dem AfD-Cluster wird durch die hohe *Betweenness-Centrality* bestätigt.

Tabelle 8: Die Zehn Accounts im Kommunikationsnetzwerk des Thema 28 nach Metrik

Account Eigenktor	Eigenvector centrality	Account Degree	Degree centrality	Account Betweenes	Betweeness centrality
c_lindner	1.000000	c_lindner	840	nhaerting	3274.5
MalteKaufmann	0.602057	MalteKaufmann	503	altenbockum	1235.0
nhaerting	0.459258	Alice_Weidel	377	MalteEngeler	487.0
Alice_Weidel	0.444710	Joerg_Meuthen	267	Privatperson	484.0
Joerg_Meuthen	0.317235	nhaerting	262	KonstantinKuhle	300.0
jreichelt	0.269017	ZDFheute	221	polenz_r	272.5
ZDFheute	0.266867	jreichelt	217	DieRoteFahne	241.5
SZ	0.247239	SZ	205	JakobSchlandt	197.5
cem_oezdemir	0.243436	cem_oezdemir	204	Privatperson	132.0
Privatperson	0.228839	Privatperson	187	houelle_beck	86.0

5.2.3 Zusammenfassung der Untersuchung der Regierungskommunikation

Die Kommunikationsnetzwerke der oben beschriebenen Themen zeigen unterschiedliche Strukturen auf. Während im Thema 13 die Kommunikation zentrierter in der Mitte des Netzwerkes stattfindet und es zu einer geringen Clusterbildung kommt, zeigen die Themen 28 und 38 eine deutliche Clusterbildung in der Diskussion. Im Thema 38 führt die Clusterbildung zu einer Unterteilung in ein Pressecluster und in ein politischen Cluster mit den jeweiligen weiteren Unterteilungen. Dagegen kommt es in Thema 28 zu keiner großteiligen Aufteilung des Netzwerkes, sondern zu einer Differenzierung in kleinere Cluster.

Die Analyse der einflussreichen Accounts innerhalb der einzelnen Themen führte in Thema 13 und 38 zu der Feststellung, dass die Accounts der Tagesschau, von ZDF heute und heute journal einen hohen Einfluss in der Kommunikation der Netzwerke ausüben. Im Thema 28 hingegen ist der Einfluss der Presse geringer und nimmt im Vergleich zu den anderen zwei Themen keinen zentralen Platz in der Kommunikation ein. Insgesamt zeigt sich im Thema 28 ein hoher Einfluss durch Accounts von Politikern die der FDP, den Grünen oder in der Mehrzahl der Fälle der AfD zugeordnet werden können.

Von den ausgewählten Regierungsaccounts zeigten die Accounts des BMFs, des BMWIs und der KfW die höchste Anzahl an Tweets in den ausgewählten Accounts. Während der Netzwerkanalyse erwies sich zudem, dass sowohl das BMG als auch der Account von Olaf Scholz im Thema 38 einflussreich im Kommunikationsnetzwerk waren. Der Account der KfW hingegen wies trotz der im Verhältnis zu den

anderen Regierungsaccounts hohen Anzahl an Tweets nur einen geringen Einfluss in Thema 13 auf. Ebenfalls hatten die Accounts der Landesbanken in Thema 13 einen hohen Einfluss. Werden die Kommunikationsnetzwerke um die Regierungsaccounts auf die Verbindungen dritten Grades beschränkt, dann wird deutlich, dass die Regierungsaccounts im Thema 13 und 38 keine abweichende Struktur zum Gesamtnetzwerk aufweisen. Mit Ausnahme des Accounts von Olaf Scholz liegen die Regierungsaccounts in der Nähe der Pressecluster und im Fall des BMFs mit einer geringen Distanz zu dem Account der Tagesschau. Bei der Analyse der Regierungskommunikation zeigt das Thema 28 ein deutlich abweichendes Verhalten zu den Themen 13 und 38 auf. Innerhalb dieses Themas sind keine ausgewählten Regierungsaccounts mit Einfluss festzustellen. Auch andere staatliche Institute nehmen an der Kommunikation in diesem Netzwerk nicht teil.

6 Diskussion

Im Folgenden werden die Ergebnisse aus dem Kapitel 5.2 erläutert und in die bisherige Literatur eingeordnet. Anschließend werden die sich aus der Begrenzung der Methodik ergebenden Problematiken für die Interpretation der Ergebnisse beschrieben. Aus diesen werden Aspekte abgeleitet, mit denen die Untersuchung der Regierungskommunikation während der COVID-19-Pandemie in weiterer Forschung erweitert und verbessert werden kann.

Innerhalb des untersuchten Zeitraums kommunizierten die Regierungsaccounts zu 391 der 804 extrahierten Themen (vgl. Kapitel 5.2.1). Dabei erfolgte eine Schwerpunktsetzung der Accounts auf die entsprechenden Aufgabengebiete der Ministerien, wobei das BMG eine starke Streuung über die Mehrzahl der Themen hinweg aufweist. Dies ist äquivalent zu der in Kapitel 2.2.2 vorgestellten Bedeutung des CDCs in der Krisenkommunikation während der COVID-19-Pandemie in den USA (vgl. Wang et al., 2021). Die Konzentration der Ministerien und Minister auf die entsprechenden Fachthemen unterstützt die bisherigen Erkenntnisse zur Krisenkommunikation von Regierungen. Die im Kapitel 2.2.1 vorgestellten Arbeiten zeigten eine informative und an Aufklärung orientierte Kommunikation von Regierungen (vgl. Crook et al., 2016; Wang et al., 2021).

Im Besonderen wurde in dieser Arbeit der Einfluss der Regierung in Themen zu den wirtschaftspolitischen Maßnahmen untersucht. Die erste Analyse zeigte, dass die Regierung, außer in den Themen 13 und 38, nur eine geringe Anzahl an Tweets

veröffentlicht hat. Diese zwei Themen behandeln die finanzielle Unterstützung der Wirtschaft und die Finanzierung der Hilfspakete. Themen, in denen die durch die Pandemie notwendigen Beschränkungen vom freien wirtschaftlichen Handeln wie beispielsweise die Schließung der Gastronomie diskutiert wurden, zeigten nur eine geringe oder keine Beteiligung von Regierungsaccounts auf. Gerade in diesen Bereichen ist jedoch eine umfangreiche Kommunikation über die Gründe der Einschränkungen und über die weitere Entwicklung notwendig, um bei den Betroffenen Verständnis erzeugen zu können und die Unsicherheit zu senken.

In den Netzwerkanalysen weisen besonders Accounts der Presse einen hohen Einfluss auf die Kommunikation auf. Regierungsaccounts hingegen sind zwar in die Netzwerke integriert, stellten sich jedoch nicht als außerordentlich einflussreich heraus. Zu einer vergleichbaren Erkenntnis gelangten Yun et al. (2016b) in der Analyse des Kommunikationsnetzwerkes während der Grippe saison 2013/2014. Die Erkenntnisse von Yun et al. (2016b) über fachbezogene Accounts mit einem hohen Einfluss innerhalb des Netzwerks konnten jedoch nicht vollständig bestätigt werden. Zwar wiesen die Investitionsbanken der Länder Berlin und Niedersachsen innerhalb des Themas 13 einen hohen Einfluss auf, im Thema 38 konnte jedoch kein Einfluss von nicht politischen Accounts mit Fachbezug festgestellt werden. Im Thema 28 wiederum hatten Personen mit einem entsprechenden Hintergrund einen Einfluss im Netzwerk.

Pourebrahim et al. (2019) konnten zeigen, dass während des Hurrikans Sandy in den USA vor allem persönliche Accounts von Politikern einen hohen Einfluss im Netzwerk hatten. In dieser Arbeit konnten Hinweise dafür festgestellt werden, dass der Einfluss von persönlichen Accounts von Politikern steigt, wenn die Thematik unspezifischer wird. In den spezifischen Themen der Soforthilfe und der Finanzierung der Hilfspakete ist der Einfluss durch die Accounts von Politikern gering. Im Thema 28, in dem die unspezifische Ablehnung von Maßnahmen betrachtet wird, ist der Einfluss durch die Accounts von Politikern hoch. Politiker mit einer Regierungsverantwortung sind innerhalb dieses Netzwerkes jedoch nicht zu finden.

Insgesamt ist der Einfluss der Regierungsaccounts im Bereich der wirtschaftspolitischen Maßnahmen als gering zu bewerten. In den konkreten Themen zu den Maßnahmen, welche zu Einschränkungen führen, ist die Kommunikation der Regierungsaccounts nicht vorhanden. In Themen der Unterstützung von Unternehmen kann eine Integration ins Netzwerk festgestellt werden, der Einfluss ist jedoch gering. Insbesondere im Thema 28, der unspezifischen Ablehnung von Maßnahmen, ist die Regierungskommunikation nicht integriert. Der gleichzeitige geringere Einfluss der Presse in diesem

Netzwerk führt zu einer Verringerung von Gestaltungsmöglichkeiten in der Diskussion und die Positionen der Regierung sind dadurch schlechter darstellbar.

Die Analyse der Regierungskommunikation kann durch mehrere verschiedene Problematiken aus der Methodik oder der Datenerhebung verzerrt sein. Die Tweets des COVID-19-Pandemie-Datensatzes wurden nicht parallel zu den Veröffentlichungen, sondern gesammelt Ende Oktober 2020 heruntergeladen (vgl. Kapitel 4.5.1). Der Zeitraum zwischen der Veröffentlichung der Tweets und dem Herunterladen der Daten kann zu Veränderungen in der Gesamtzahl der Tweets führen. Tweets können zum einen durch die Nutzer und zum anderen von Twitter gelöscht werden, wenn diese gegen die Twitter Richtlinien verstößen. Aufgrund der hohen Anzahl an Tweets und der Länge des Zeitraums ist davon auszugehen, dass der COVID-19-Pandemie-Datensatz unvollständig ist. Dies kann beim Bilden von Themen oder auch in der Struktur der Kommunikationsnetzwerke zu Verzerrungen führen. In Anbetracht der hohen Anzahl an Tweets und Antworten zu den untersuchten Themen führt das Fehlen von einzelnen Tweets jedoch nur zu einer geringen Verzerrung.

In der Datenbeschaffung liegt ein weiterer Grund für eine systematische Verzerrung der Analyse. Untersucht wird in dieser Arbeit die an die eigene Bevölkerung gerichtete Regierungskommunikation der Bundesrepublik Deutschland. Der Datensatz umfasst jedoch alle deutschsprachigen Tweets und beinhaltet somit auch Accounts aus der Schweiz oder aus Österreich beziehungsweise aus der gesamten Welt. Wie in Kapitel 2.2.1 beschrieben fanden Lu & Brelsford (2014) Hinweise dafür, dass diese deutschsprachige Kommunikation im Datensatz zwar vorhanden ist, durch die Clusterbildung jedoch eine Verzerrung der Analyse vermieden wird. Dieser Effekt der Clusterbildung konnte in der Analyse der Regierungskommunikation beobachtet werden. Die Kommunikation im selben Thema in Bezug auf ein anderes Land führt somit nicht zu einer Verzerrung der Analyse. Antworten Nutzer aus einem anderen Land auf Tweets innerhalb eines Clusters mit Bezug auf die Bundesrepublik Deutschland, führt dies jedoch zu einer Verzerrung. Dieser Effekt wird durch die hohe Anzahl an Tweets und Antworten ausgeglichen.

In der Auswahl der zu analysierenden Regierungsaccounts folgte eine Fokussierung auf die Accounts des Bundes. Die Netzwerkanalyse des Themas 13 zeigte einen hohen Einfluss der Investitionsbanken der Bundesländer Berlin und Niedersachsens und einen hohen Einfluss des Wirtschaftsministeriums Nordrhein-Westfalens. Die Schlussfolgerung hieraus ist, dass die verschiedene Ausgestaltung von Maßnahmen auf Bundesländerebene dazu führte, dass die Bedeutung der Kommunikation der Lan-

desregierungen zunimmt. Das Berücksichtigen von Accounts der Landesregierungen kann somit zu einer präziseren Analyse der Regierungskommunikation beitragen.

Die in dieser Arbeit verwendeten Modelle zum Extrahieren der Themen aus dem COVID-19-Datensatz deklarierten eine hohe Zahl der Tweets als Rauschen. Gleichzeitig wurde ein Silhouettenkoeffizient um null festgestellt. Insgesamt deutet diese Kombination darauf hin, dass in der Bestimmung der Themen Optimierungsbedarf besteht. Jedoch werden die Themen und auch der Silhouettenkoeffizient auf Basis des euklidischen Abstands der Tweets im Raum berechnet. Zu einer Verbesserung der Bestimmung der Themen sollten deshalb nicht ausschließlich weitere Modellbeziehungsweise Parameterkombinationen zur Bestimmung der Themen untersucht werden, sondern es sollte auch analysiert werden, welchen Einfluss andere Modelle zur Bestimmung des semantischen Raums auf die Themencluster haben. Vortrainierte Modelle wie *Bidirectional Encoder Representations from Transformers* (BERT) (vgl. Devlin et al., 2019) oder *FastText* (vgl. Joulin et al., 2016) würden sich hier zur Überprüfung anbieten.

Um ein Modell zum Schätzen der Antworten auf die Tweets erhalten zu können, wurden im Kapitel 5.1.2 verschiedene neuronale Netze trainiert. Bei der Schätzung eines dem Modell unbekannten Testdatensatzes zeigten die neuronalen Netze jedoch eine geringe Genauigkeit, weshalb die Analyse des Sentiments der Antworten in dieser Arbeit verworfen wurde. Der Mangel einer Sentiment-Schätzung innerhalb der Netzwerkanalyse führt dazu, dass nicht festgestellt werden kann, auf welcher Basis zwei Accounts im Netzwerk nah beieinanderstehen. Die Nähe zueinander bedeutet, dass eine hohe Anzahl an Accounts auf beide Accounts geantwortet hat. Ob die Antwortgeber zu beiden Accounts positiv, zu beiden Accounts negativ oder zu einem Account positiv und zum anderen negativ stehen, ist jedoch unbekannt. Die Berücksichtigung des Sentiments im Netzwerk kann also die Analyse der Integration in ein Netzwerk verbessern. Darüber hinaus kann durch das Sentiment der Antworten bestimmt werden, wie in den verschiedenen Themen die Reaktionen auf die Regierungskommunikation sind. Für das Entwickeln eines Sentiment-Modells bieten sich verschiedene Verfahren an, die eine hohe Genauigkeit aufweisen könnten. Neben der Optimierung der Architektur von den in dieser Arbeit verwendeten neuronalen Netzen weisen Sentiment-Modelle, die auf Finetuning von großen vortrainierten Modellen wie BERT (vgl. Devlin et al., 2019) basieren, eine hohe Erfolgswahrscheinlichkeit auf. Besonders vielversprechend wäre hier ein auf den gesamten Twitter-Datensatz während der COVID-19-Pandemie angepasstes Modell wie Covid-Twitter BERT (vgl. Müller et al., 2020). Covid-Twitter

BERT ist bisher jedoch lediglich in englischer Sprache verfügbar.

Neben der Verbesserung der Modelle durch die Berücksichtigung des Sentiments kann die Regierungskommunikation auch über den Zeitverlauf der Pandemie hinweg analysiert und dadurch Veränderungen in der Regierungskommunikation untersucht werden können. Da in einer zu kleinteiligen Betrachtung die genutzten Modelle zu einer zu hohen Schwankung führen, kann durch den Vergleich der nun erhebbaren Daten der zweiten Welle die Veränderung der Regierungskommunikation zwischen den Wellen analysiert werden.

7 Fazit

In Zeiten einer weltumspannenden Pandemie besteht eine Unsicherheit über die wirtschaftliche Entwicklung. Diese Unsicherheit wird verschärft, wenn Unklarheit über die politische Reaktion auf die Pandemie herrscht. Eine umfangreiche und präzise Regierungskommunikation kann die Unsicherheit über die wirtschaftspolitischen Maßnahmen reduzieren und ein Verständnis für pandemiebedingte Einschränkungen schaffen. Überzeugende Argumente und eine gute Darstellung der Informationen sind bei der Verbreitung von Informationen erfolglos, wenn die Kommunikation die Bevölkerung nicht erreichen kann. In dieser Arbeit wurde die Regierungskommunikation spezifisch auf der Plattform Twitter untersucht. Im ersten Schritt der Analyse wurde analysiert, zu welchen Themen die Regierung kommuniziert. Anschließend wurde explizit der Einfluss der Regierungskommunikation bei wirtschaftspolitischen Themen analysiert.

Um das Untersuchen der Regierungskommunikation zu ermöglichen, wurde der erhobene Twitter-Datensatz mit Techniken des *Deep-Learning* und des NLP bearbeitet. So konnte eine Darstellung der Tweets im mehrdimensionalen Raum erreicht und es konnten durch die Anwendung von hierarchisch dichtebasierten Clustern die Themen während der COVID-19-Pandemie bestimmt werden. Aus allen extrahierten Themen wurden mithilfe einer abstandsbasierten Suche Themen ausgewählt, die eine geringe Distanz zu wirtschaftspolitischen Themen aufwiesen. Auf Basis der gesamten Themen wurden die Kommunikationsschwerpunkte der Regierung analysiert und es wurde untersucht, zu welchen wirtschaftspolitischen Themen die Regierung kommuniziert hatte. Die Regierungskommunikation in den zwei wirtschaftspolitischen Themen mit der höchsten Anzahl an Tweets wurde durch die Anwendung von Netzwerkanalysetechniken untersucht.

Die Untersuchung der Regierungskommunikation zeigte, dass eine Vielzahl unterschiedlicher Themen durch die Regierung angesprochen wurde. Jedoch nur zu einer geringen Anzahl der Themen wurde eine hohe Anzahl an Tweets veröffentlicht. Insbesondere das BMG zeigte auf Twitter eine hohe Aktivität über die Themen gestreut bei gleichzeitiger geringer Konzentration auf einzelne Themen. Als Folge der weiten Streuung der Kommunikation und der geringen Gesamtanzahl an Tweets konnten aus den 16 extrahierten Themen im Bereich der Wirtschaft nur zwei auf den Einfluss im Netzwerk untersucht werden. Diese zwei Themen betreffen die Unterstützung von Unternehmen und die Finanzierung von Hilfspaketen. In Themen, die Einschränkungen durch wirtschaftspolitische Maßnahmen diskutieren, wurde eine geringe Anzahl an Tweets von der Regierung veröffentlicht. Die ausgewählten Regierungsaccounts waren in die beiden wirtschaftspolitischen Netzwerke integriert. Presseaccounts und fachspezifische Accounts wie Landesbanken wiesen jedoch einen höheren Einfluss im Netzwerk auf.

Eines der extrahierten Themen zeigte eine unspezifische Ablehnung der Maßnahmen der Regierung auf. Für ein besseres Verständnis der Kommunikation zur Ablehnung der Maßnahmen wurde dieses Netzwerk ebenfalls analysiert. Im Vergleich zu den beiden wirtschaftspolitischen Themen war im Netzwerk der Einfluss durch Presseaccounts deutlich reduziert. Anders als bei den wirtschaftspolitischen Themen wiesen vor allem Politiker der Opposition, insbesondere der AfD, einen hohen Einfluss im Netzwerk auf.

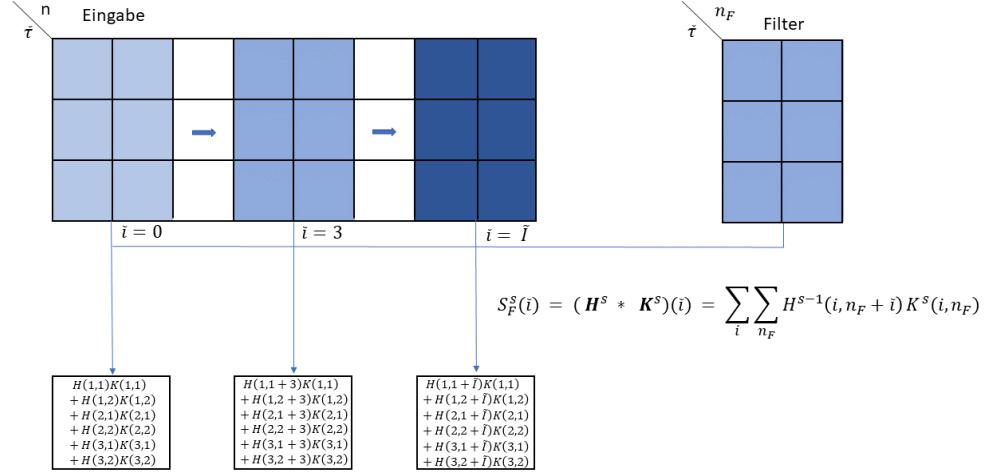
In dieser Arbeit wurden erstmals nicht nur die Themen der Regierungskommunikation während der COVID-19-Pandemie, sondern auch der Einfluss der Regierungskommunikation im Netzwerk untersucht. Dabei stellt diese Arbeit die einzige bekannte Arbeit dar, welche die wirtschaftspolitische Kommunikation der Regierung auf Twitter während eines Krankheitsausbruches betrachtet. Darüber hinaus sind bisher keine Analysen zur Krisenkommunikation im deutschsprachigen Twitter während der COVID-19-Pandemie bekannt. Die Ergebnisse dieser Arbeit bestätigen jedoch eine Reihe von Erkenntnissen der bisherigen Forschung zur Krisenkommunikation. Sowohl die Bedeutung von Presseaccounts in den Kommunikationsnetzwerken als auch die fachspezifische Konzentration der Regierungskommunikation wurden bereits in verschiedenen Arbeiten gezeigt.

Die in dieser Arbeit gewonnenen Erkenntnisse können für eine Anpassung und Optimierung der Regierungskommunikation während der COVID-19-Pandemie genutzt werden, um dadurch die Unsicherheit der Bevölkerung reduzieren zu können. Die

Regierungsaccounts stellen dabei eine ergänzende Plattform dar, über die interessierte Bürger mit spezifischen Informationen erreicht werden können. Um in der Informationsverteilung jedoch eine breite Masse ansprechen zu können, sollte bewusst die Kooperation mit der Presse gesucht werden. Die Analyse zeigte in der Kommunikation zu Einschränkungen der wirtschaftspolitischen Handlungsfreiheit Lücken auf. Durch das Schließen dieser Lücke könnten eine höhere Akzeptanz der Maßnahmen sowie eine Verringerung von Unsicherheit erreicht werden.

A Anhang

A.1 Visuelle Ergänzung zur Durchführung einer 1D Faltung (eigene Darstellung)



A.2 Visuelle Darstellung des Hyperparametertraining

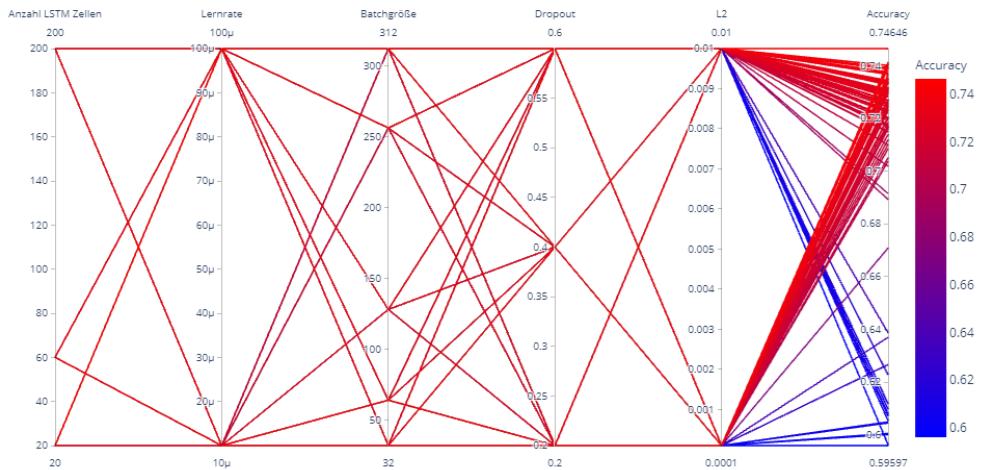


Abbildung A1: Hyperparametertraining LSTM (eigene Darstellung)

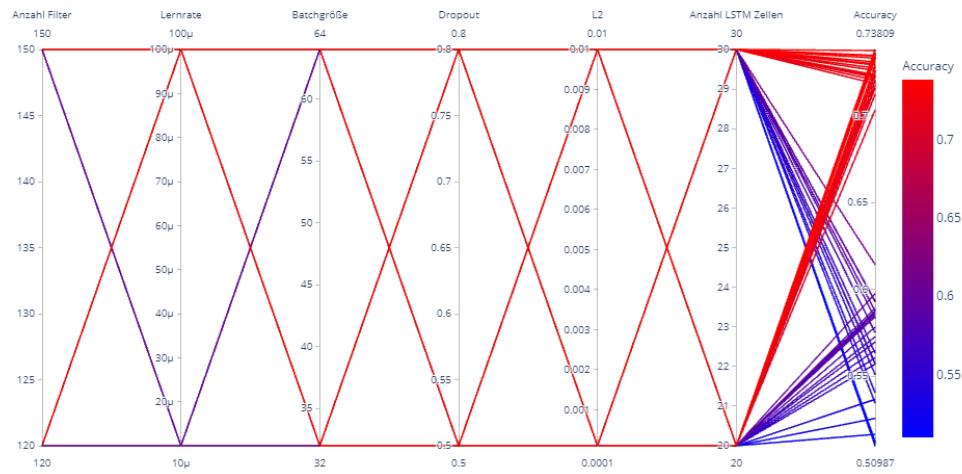


Abbildung A2: Hyperparametertraining CNN-LSTM (eigene Darstellung)

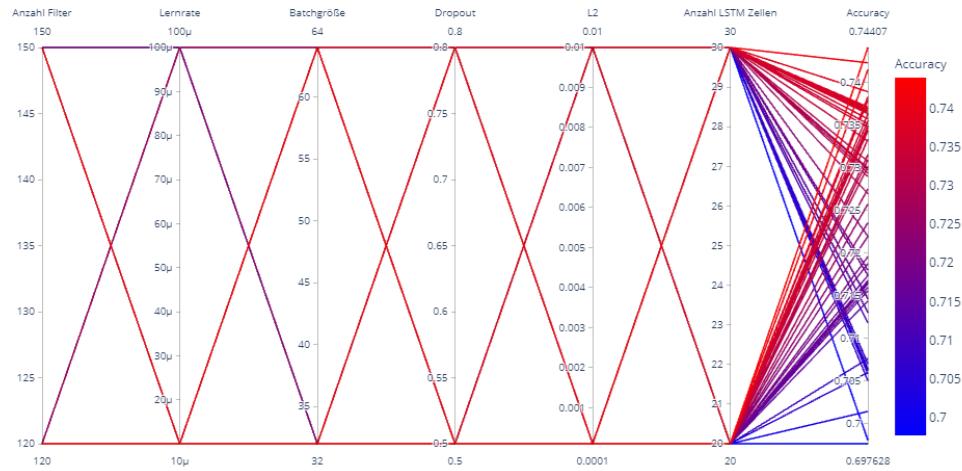


Abbildung A3: Hyperparametertraining LSTM-CNN (eigene Darstellung)

A.3 Visuelle Darstellung der durch die Abstände im semantischen Raum gewonnenen Themen



Abbildung A4: Die fünf Themen mit den geringsten Abstand im semantischen Raum zum Wort "Bchliessungen" (eigene Darstellung)



Abbildung A5: Die fünf Themen mit den geringsten Abstand im semantischen Raum zum Wort "hilfen" (eigene Darstellung)

A.4 Kommunikationsnetzwerk des BMGs, des BMFs und des Accounts von Olaf Scholz im Thema 38 (eigene Darstellung)

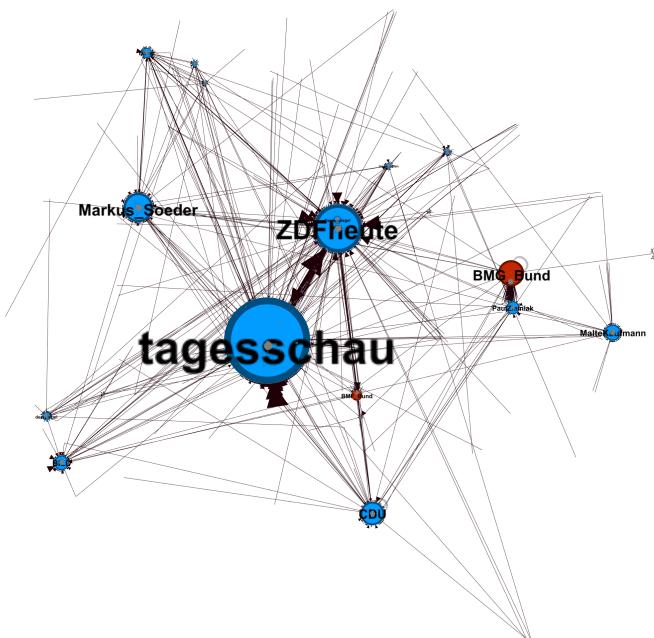


Abbildung A6: Kommunikationsnetzwerk des BMGs in Thema 38 (eigene Darstellung)

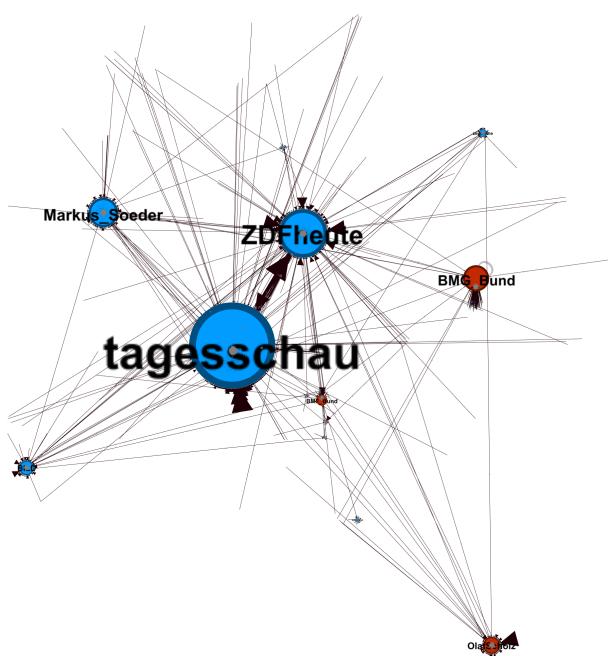


Abbildung A7: Kommunikationsnetzwerk des BMFs in Thema 38 (eigene Darstellung)

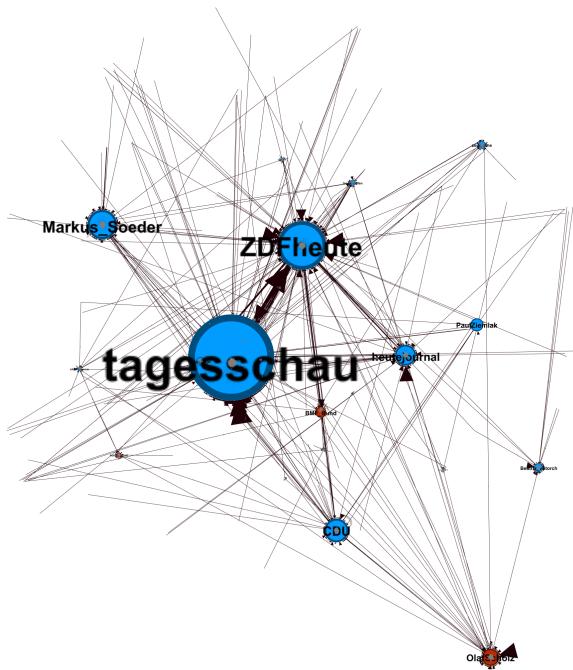


Abbildung A8: Kommunikationsnetzwerk des Accounts von Olaf Scholz in Thema 38 (eigene Darstellung)

A.5 Untersuchte Regierungsaccounts

Tabelle A1: Auflistung der untersuchten Regierungsaccounts

Name	Amt	Twittername	Twittername Ministerium/Institut
Steffen Seibert Ulrike Demmer	Sprecher der Bundesregierung Stellvertretende Sprecherin der Bundesregierung	RegSprecher UlrikeDemmer	NA NA
Helge Braun Olaf Scholz Heiko Maas	Bundesminister für besondere Aufgaben/Bundeskanzleramt Bundesminister der Finanzen/Vizekanzler Bundesminister des Auswärtigen Amts	HBrAun OlafScholz HeikoMaas	NA BMF_Bund AuswaertigesAmt
Peter Altmaier Jens Spahn Hubertus Heil Andreas Scheuer Julia Klöckner	Bundesminister für Wirtschaft und Energie Bundesminister für Gesundheit Bundesminister für Arbeit und Soziales Bundesminister für Verkehr und digitale Infrastruktur Bundesminister für Ernährung und Landwirtschaft	peteraltmaier jensspahn hubertus_heil AndiScheuer JuliaKloeckner	BMWi_Bund BMG_Bund BMAS_Bund BMVI_bnel
Annette Kramp-Karrenbauer Svenja Schulze Anja Karliczek Gerd Müller Franziska Giffey Horst Seehofer Christine Lambrecht NA NA NA	Bundesministerin der Verteidigung Bundesministerin für Umwelt, Naturschutz und nukleare Sicherheit Bundesministerin für Bildung und Forschung Bundesminister für wirtschaftliche Zusammenarbeit und Entwicklung Bundesministerin für Familie, Senioren, Frauen und Jugend Bundesminister des Innern, für Bau und Heimat Bundesministerin der Justiz und für Verbraucherschutz Robert Koch-Institut Paul-Ehrlich-Institut Kreditanstalt für Wiederaufbau	akk SvenjaSchulze68 AnjaKarliczek Kein Twitter Account vorhanden Kein Twitter Account vorhanden Kein Twitter Account vorhanden Kein Twitter Account vorhanden NA NA NA	BMVg_Bundeswehr bmu BMBF_Bund BMZ_Bund BMFSFJ BML_Bund BMJV_Bund rki_de PEI_Germany KW

Literaturverzeichnis

- Alternative-für-Deutschland. (2020). *Vorstellung des Bundesvorstand. Alternative für Deutschland*. <https://www.afd.de/partei/bundesvorstand/>
- Angelov, D. (2020). Top2Vec: Distributed representations of topics. *arXiv:2008.09470 [Cs, Stat]*. <http://arxiv.org/abs/2008.09470>
- Bayerische-Staatskanzlei. (2020). *Ministerpräsident – bayerisches landesportal*. <https://www.bayern.de/staatsregierung/ministerpraesident/>
- Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., & Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, 37(1), 38–44. <https://doi.org/10.1038/nbt.4314>
- Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A neural probabilistic language model. *The Journal of Machine Learning Research*, 3(null), 1137–1155.
- Berlin, B. (2017). *About us – leadership and organization*. Bundesland Berlin. <https://www.berlin.de/sen/kultur/en/about-us/>
- Berlin, B. (2019, September 2). *Ramona pop - senator for economics, energy and public enterprises*. <https://www.berlin.de/sen/web/ueber-uns/leitung-und-organisation/senatorin-ramona-pop/artikel.580013.en.php>
- Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1–8. <https://doi.org/10.1016/j.jocs.2010.12.007>
- Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *The Journal of Mathematical Sociology*, 2(1), 113–120. <https://doi.org/10.1080/0022250X.1972.9989806>
- Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network analysis in the social sciences. *Science*, 323(5916), 892–895. <https://doi.org/10.1126/science.1165821>
- Bovet, A., & Makse, H. A. (2019). Influence of fake news in twitter during the 2016 US presidential election. *Nature Communications*, 10(1), 7. <https://doi.org/10.1038/s41467-018-07761-2>
- Bovet, A., Morone, F., & Makse, H. A. (2018). Validation of twitter opinion trends

- with national polling aggregates: Hillary Clinton vs Donald Trump. *Scientific Reports*, 8(1), 8673. <https://doi.org/10.1038/s41598-018-26951-y>
- Britannica. (2020). *Cem Özdemir german politician* britannica. <https://www.britannica.com/biography/Cem-Ozdemir>
- Brynielsson, J., Johansson, F., Jonsson, C., & Westling, A. (2014). Emotion classification of social media posts for estimating people's reactions to communicated alert messages during crises. *Security Informatics*, 3(1), 7. <https://doi.org/10.1186/s13388-014-0007-3>
- BASIG - gesetz über das bundesinstitut für impfstoffe und biomedizinische arzneimittel, (1972). <https://www.gesetze-im-internet.de/basig/BJNR011630972.html>
- Bundesministerium-für-Wirtschaft-und-Energie. (2020). *Pressemitteilung - Zusätzliches KfW-Sonderprogramm 2020 für die Wirtschaft startet heute*. Bundesministerium für Wirtschaft und Energie. <https://www.bmwi.de/Redaktion/DE/Pressemitteilungen/2020/20200323-zusaetzliches-kfw-sonderprogramm-2020-fuer-die-wirtschaft-startet-heute.html>
- Bundespressamt. (2017a). *Vorstellung des Bundesminister des Innern, für Bau und Heimat. Bundesregierung*. <https://www.bundesregierung.de/breg-de/bundesregierung/bundeskabinett/horst-seehofer>
- Bundespressamt. (2017b). *Vorstellung des Bundesminister für wirtschaftliche Zusammenarbeit und Entwicklung. Bundesregierung*. <https://www.bundesregierung.de/breg-de/bundesregierung/bundeskabinett/gerd-mueller>
- Bundespressamt. (2019a). *Vorstellung der Bundesministerin der Justiz und für Verbraucherschutz. Bundesregierung*. <https://www.bundesregierung.de/breg-de/bundesregierung/bundeskabinett/christine-lambrecht>
- Bundespressamt. (2019b). *Vorstellung der Bundesministerin für Familie, Senioren, Frauen und Jugend. Bundesregierung*. <https://www.bundesregierung.de/breg-de/bundesregierung/bundeskabinett/franziska-giffey>
- Bundespressamt. (2019c). *Vorstellung des Bundesminister der Finanzen. Bundesregierung*. <https://www.bundesregierung.de/breg-de/bundesregierung/bundeskabinett/olaf-scholz>
- Bundespressamt. (2020a). *Besprechung der Bundeskanzlerin mit den Regierungschefinnen und Regierungschefs der Länder vom 22.03.2020*. <https://www.bundesregierung.de/breg-de/themen/coronavirus/besprechung-der-bundeskanzlerin-mit-den-regierungschefinnen-und-regierungschefs-der-laender-vom-22-03-2020>

- Bundespressamt. (2020b). *Vorstellung des Bundesminister für Verkehr und digitale Infrastruktur*. Bundesregierung. <https://www.bundesregierung.de/breg-de/bundesregierung/bundeskabinett/andreas-scheuer>
- Bundeswirtschaftsministerium. (2020). *Soforthilfe für Solo-Selbstständige und Kleinbetriebe*. <https://www.bmwi.de/Redaktion/DE/Artikel/Wirtschaft/Corona-Virus/unterstuetzungsmassnahmen-faq-04.html>
- Burkov, A. (2019). Grundlegende Techniken. In *Machine Learning kompakt*. mitp Verlag. <https://learning.oreilly.com/library/view/machine-learning-kompakt/9783958459977/Text/buchmlch5.xhtml>
- Caldarelli, G., De Nicola, R., Del Vigna, F., Petrocchi, M., & Saracco, F. (2020). The role of bot squads in the political propaganda on twitter. *Communications Physics*, 3(1), 1–15. <https://doi.org/10.1038/s42005-020-0340-4>
- Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (Eds.), *Advances in knowledge discovery and data mining* (pp. 160–172). Springer. https://doi.org/10.1007/978-3-642-37456-2_14
- Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K. (2010). Measuring user influence in twitter: The million follower fallacy. *Proceedings of the International AAAI Conference on Web and Social Media*, 4(1). <https://ojs.aaai.org/index.php/ICWSM/article/view/14033>
- Cherven, K. (2015). Mastering gephi network visualization. In *A network graph analysis primer*. <https://learning.oreilly.com/library/view/mastering-gephi-network/9781783987344/ch01s02.html>
- Christlich-Demokratische-Union-Deutschlands. (2018). *Vorstellung von Paul Ziemiak*. Christlich Demokratische Union Deutschlands. <https://www.cdu.de/vorstand/paul-ziemiak>
- Cieliebak, M., Deriu, J. M., Egger, D., & Uzdilli, F. (2017). A twitter corpus and benchmark resources for german sentiment analysis. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 45–51. <https://doi.org/10.18653/v1/W17-1106>
- Crook, B., Glowacki, E. M., Suran, M., Harris, J. K., & Bernhardt, J. M. (2016). Content analysis of a live CDC twitter chat during the 2014 ebola outbreak. *Commu-*

nication Research Reports, 33(4), 349–355. <https://doi.org/10.1080/08824096.2016.1224171>

- Curiskis, S. A., Drake, B., Osborn, T. R., & Kennedy, P. J. (2020). An evaluation of document clustering and topic modelling in two online social networks: Twitter and reddit. *Information Processing & Management*, 57(2), 102034. <https://doi.org/10.1016/j.ipm.2019.04.002>
- Das, K., Samanta, S., & Pal, M. (2018). Study on centrality measures in social networks: A survey. *Social Network Analysis and Mining*, 8(1), 13. <https://doi.org/10.1007/s13278-018-0493-2>
- Davidson, G. S., Wylie, B. N., & Boyack, K. W. (2001). Cluster stability and the use of noise in interpretation of clustering. *IEEE Symposium on Information Visualization, 2001. INFOVIS 2001.*, 23–30. <https://doi.org/10.1109/INFVIS.2001.963275>
- Deb, A., Luceri, L., Badaway, A., & Ferrara, E. (2019). Perils and challenges of social media and election manipulation analysis: The 2018 US midterms. *Companion Proceedings of the 2019 World Wide Web Conference*, 237–247. <https://doi.org/10.1145/3308560.3316486>
- Deutscher-Bundestag. (2017). *Deutscher Bundestag - Alice Weidel*. Deutscher Bundestag. <https://www.bundestag.de/mdb>
- Deutscher-Bundestag. (2018). *Deutscher Bundestag - Christian Lindner*. Deutscher Bundestag. <https://www.bundestag.de/abgeordnete/biografien/L/521640-521640>
- Deutscher-Bundestag. (2020a). *Deutscher Bundestag - Beatrix von Storch*. Deutscher Bundestag. <https://www.bundestag.de/mdb>
- Deutscher-Bundestag. (2020b). *Deutscher Bundestag - Konstantin Kuhle*. Deutscher Bundestag. <https://www.bundestag.de/mdb>
- Deutscher-Bundestag. (2020c). *Deutscher Bundestag - Renate Künast*. Deutscher Bundestag. <https://www.bundestag.de/mdb>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805 [Cs]*. <http://arxiv.org/abs/1810.04805>
- Effrosynidis, D., Symeonidis, S., & Arampatzis, A. (2017). A comparison of pre-processing techniques for twitter sentiment analysis. In J. Kamps, G. Tsakonas, Y. Manolopoulos, L. Iliadis, & I. Karydis (Eds.), *Research and advanced technology*

- for digital libraries* (pp. 394–406). Springer International Publishing. https://doi.org/10.1007/978-3-319-67008-9_31
- Feld, L., Grimm, V., Schnitzer, M., Truger, A., & Wieland, V. (2020). *Corona-Krise gemeinsam bewältigen, Resilienz und Wachstum stärken* (p. 556). Sachverständigenrat zur Begutachtung der gesamtwirtschaftlichen Entwicklung.
- Ferrara, E. (2020). What types of COVID-19 conspiracies are populated by twitter bots? *First Monday*. <https://doi.org/10.5210/fm.v25i6.10633>
- Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 40(1), 35–41. <https://doi.org/10.2307/3033543>
- Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, 1(3), 215–239. [https://doi.org/10.1016/0378-8733\(78\)90021-7](https://doi.org/10.1016/0378-8733(78)90021-7)
- GmbH, G. K. K. (2020). *Schutzschirmverfahren*. <https://www.galeria.de/Schutzschirmverfahren.html>
- Goldberg, Y. (2017). *Neural network methods for natural language processing*. Morgan & Claypool Publishers.
- Goldberg, Y., & Levy, O. (2014). Word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv:1402.3722 [Cs, Stat]*. <http://arxiv.org/abs/1402.3722>
- Goodfellow, I., Bengio, Y., & Courville, A. (2018). *Deep learning : Das umfassende handbuch : Grundlagen, aktuelle verfahren und algorithmen, neue forschungsansätze* (1. Auflage). mitp. http://digitale-objekte.hbz-nrw.de/storage2/2019/05/03/file_122/8434717.pdf
- Harris, Z. S. (1954). Distributional structure. *WORD*, 10(2), 146–162. <https://doi.org/10.1080/00437956.1954.11659520>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer New York. <https://doi.org/10.1007/978-1-4614-7138-7>
- Jianqiang, Z., & Xiaolin, G. (2017). Comparison research on text pre-processing methods on twitter sentiment analysis. *IEEE Access*, 5, 2870–2879. <https://doi.org/10.1109/ACCESS.2017.2672677>
- Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, 32(3), 241–254. <https://doi.org/10.1007/BF02289588>

- Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv:1607.01759 [Cs]*. <http://arxiv.org/abs/1607.01759>
- JustAnotherArchivist. (2020). *JustAnotherArchivist/snsrape*. <https://github.com/JustAnotherArchivist/snsrape>
- Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv:1408.5882 [Cs]*. <http://arxiv.org/abs/1408.5882>
- Kobak, D., & Linderman, G. C. (2019). UMAP does not preserve global structure any better than t-SNE when using the same initialization. *bioRxiv*, 2019.12.19.877522. <https://doi.org/10.1101/2019.12.19.877522>
- Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is twitter, a social network or a news media? *Proceedings of the 19th International Conference on World Wide Web*, 591–600. <https://doi.org/10.1145/1772690.1772751>
- Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents. *arXiv:1405.4053 [Cs]*. <http://arxiv.org/abs/1405.4053>
- LeCun, Y. (1989). Generalization and network design strategies. *Connectionism in Perspective*. <https://nyuscholars.nyu.edu/en/publications/generalization-and-network-design-strategies>
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. <https://doi.org/10.1038/nature14539>
- Lee, J. M. (2011). *Introduction to topological manifolds* (Vol. 202). Springer New York. <https://doi.org/10.1007/978-1-4419-7940-7>
- Lu, X., & Brelsford, C. (2014). Network structure and community evolution on twitter: Human behavior change in response to the 2011 japanese earthquake and tsunami. *Scientific Reports*, 4(1), 6773. <https://doi.org/10.1038/srep06773>
- Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4), 309–317. <https://doi.org/10.1147/rd.14.0309>
- Manning, C. D., Raghavan, P., & Schütze, H. (2009). *Introduction to information retrieval*. 569.
- Martin, S., Brown, W. M., Klavans, R., & Boyack, K. W. (2011). OpenOrd: An open-source toolbox for large graph layout. *Visualization and Data Analysis 2011*, 7868, 786806. <https://doi.org/10.1117/12.871402>

- McInnes, L. (2018). *How UMAP works — umap 0.5 documentation*. https://umap-learn.readthedocs.io/en/latest/how_umap_works.html
- McInnes, L., & Healy, J. (2017). Accelerated hierarchical density based clustering. *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, 33–42. <https://doi.org/10.1109/ICDMW.2017.12>
- McInnes, L., Healy, J., & Melville, J. (2020). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv:1802.03426 [Cs, Stat]*. <http://arxiv.org/abs/1802.03426>
- McInnes, L., Healy, J., & Revision, S. R. (2016). *How HDBSCAN works — hdbSCAN 0.8.1 documentation*. https://hdbSCAN.readthedocs.io/en/latest/how_hdbSCAN_works.html
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv:1301.3781 [Cs]*. <http://arxiv.org/abs/1301.3781>
- Mozetič, I., Grčar, M., & Smailović, J. (2016). Multilingual twitter sentiment classification: The role of human annotators. *PLOS ONE*, 11(5), e0155036. <https://doi.org/10.1371/journal.pone.0155036>
- Müller, M., Salathé, M., & Kummervold, P. E. (2020). COVID-twitter-BERT: A natural language processing model to analyse COVID-19 content on twitter. *arXiv:2005.07503 [Cs]*. <http://arxiv.org/abs/2005.07503>
- New York, C. of. (2020). *Mayors of the city of new york*. <https://www1.nyc.gov/site/dcias/about/green-book-mayors-of-the-city-of-new-york.page>
- Olah, C. (2015, August 27). *Understanding LSTM networks – colah’s blog*. <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>
- Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, III–1310–III–1318.
- Pourebrahim, N., Sultana, S., Edwards, J., Gochanour, A., & Mohanty, S. (2019). Understanding communication dynamics on twitter during natural disasters: A case study of hurricane sandy. *International Journal of Disaster Risk Reduction*, 37, 101176. <https://doi.org/10.1016/j.ijdrr.2019.101176>
- Republik-Österreich. (2020). *Sigrid Maurer, BA, Biografie*. https://www.parlament.gv.at/WWER/PAD_83101/index.shtml

- Robert-Koch-Institut. (2020). *RKI - informationen*. https://www.rki.de/DE/Content/Institut/institut_node.html;jsessionid=6B1B06889681C52BF25EA84038A538A5.internet102
- Roesslein, J. (2009). *API reference — tweepy 3.5.0 documentation*. <http://docs.tweepy.org/en/v3.5.0/api.html#tweepy-api-twitter-api-wrapper>
- Rojas, R. (1996). The backpropagation algorithm. In *Neural networks* (pp. 149–182). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-61068-4_7
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
- Ruan, Y., Durresi, A., & Alfantoukh, L. (2018). Using twitter trust network for stock market analysis. *Knowledge-Based Systems*, 145, 207–218. <https://doi.org/10.1016/j.knosys.2018.01.016>
- Rufai, S. R., & Bunce, C. (2020). World leaders' usage of twitter in response to the COVID-19 pandemic: A content analysis. *Journal of Public Health*, 42(3), 510–516. <https://doi.org/10.1093/pubmed/fdaa049>
- Ruz, G. A., Henríquez, P. A., & Mascareño, A. (2020). Sentiment analysis of twitter data during critical events through bayesian networks classifiers. *Future Generation Computer Systems*, 106, 92–104. <https://doi.org/10.1016/j.future.2020.01.005>
- Schlitt, A.-L. (2020). Autogipfel: Wirtschaftsweise gegen Kaufprämie für Autos mit Verbrennungsmotor. *Die Zeit*. <https://www.zeit.de/mobilitaet/2020-09/autogipfel-monika-schnitzer-kaufpraemie-verbrennungsmotor>
- Schwenk, H. (2007). Continuous space language models. *Computer Speech & Language*, 21(3), 492–518. <https://doi.org/10.1016/j.csl.2006.09.003>
- Shaw, M. E. (1954). Group structure and the behavior of individuals in small groups. *The Journal of Psychology*, 38(1), 139–149. <https://doi.org/10.1080/00223980.1954.9712925>
- Singh, L., Bansal, S., Bode, L., Budak, C., Chi, G., Kawintiranon, K., Padden, C., Vanarsdall, R., Vraga, E., & Wang, Y. (2020). A first look at COVID-19 information and misinformation sharing on twitter. *arXiv:2003.13907 [Cs]*. <http://arxiv.org/abs/2003.13907>
- Singh, L., Bode, L., Budak, C., Kawintiranon, K., Padden, C., & Vraga, E. (2020).

- Understanding high- and low-quality URL sharing on COVID-19 twitter streams. *Journal of Computational Social Science*, 3(2), 343–366. <https://doi.org/10.1007/s42001-020-00093-6>
- Sluban, B., Smailović, J., Battiston, S., & Mozetič, I. (2015). Sentiment leaning of influential communities in social networks. *Computational Social Networks*, 2(1), 9. <https://doi.org/10.1186/s40649-015-0016-5>
- Smailović, J., Grčar, M., Lavrač, N., & Žnidaršič, M. (2014). Stream-based active learning for sentiment analysis in the financial domain. *Information Sciences*, 285, 181–203. <https://doi.org/10.1016/j.ins.2014.04.034>
- Sosa, P. M. (2017). *Twitter sentiment analysis using combined LSTM-CNN models*. 9.
- Spallek, S. (2020). Corona-Maßnahmen in Berlin: „Es gibt keine Begründung dafür, dass diese Sperrstunde etwas bringt“. *SPIEGEL*. <https://www.spiegel.de/panorama/justiz/corona-in-berlin-sperrstunde-aufgehoben-alkoholverbot-gilt-weiterhin-a-2c7c8aa4-1177-4e82-9aa5-997166e7c68a>
- Sparck, J. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), 11–21. <https://doi.org/10.1108/eb026526>
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1), 1929–1958.
- Südwestrundfunk, S. W. R. (2020). *Malte Kaufmann setzt Fokus auf Arbeitssplätze und Sicherheit*. *swr.online*. <https://www.swr.de/swraktuell/baden-wuerttemberg/stuttgart/stuttgart-oberbuergermeisterwahl-kandidat-afd-kaufmann-100.html>
- Symeonidis, S., Effrosynidis, D., & Arampatzis, A. (2018). A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. *Expert Systems with Applications*, 110, 298–310. <https://doi.org/10.1016/j.eswa.2018.06.022>
- tagesschau.de. (2020). SSatumäuf Reichstagsgebäude: Mit Falschmeldungen aufgetaucht. *tagesschau.de*. <https://www.tagesschau.de/faktenfinder/reichstag-berlin-sturm-fakenews-101.html>
- Tang, D., Qin, B., Feng, X., & Liu, T. (2016). Effective LSTMs for target-dependent

- sentiment classification. *arXiv:1512.01100 [Cs]*. <http://arxiv.org/abs/1512.01100>
- Taylor, D. B. (2021). A timeline of the coronavirus pandemic. *The New York Times*.
<https://www.nytimes.com/article/coronavirus-timeline.html>
- Twitter, I. (2017). *Giving you more characters to express yourself*. https://blog.twitter.com/en_us/topics/product/2017/Giving-you-more-characters-to-express-yourself.html
- Twitter, I. (2020a). *About replies and mentions*. <https://help.twitter.com/en/using-twitter/mentions-and-replies>
- Twitter, I. (2020b). *Developer agreement and policy – twitter developers*. <https://developer.twitter.com/en/developer-terms/agreement-and-policy>
- Uysal, A. K., & Murphay, Y. L. (2017). Sentiment classification: Feature selection based approaches versus deep learning. *2017 IEEE International Conference on Computer and Information Technology (CIT)*, 23–30. <https://doi.org/10.1109/CIT.2017.53>
- Wang, Y., Hao, H., & Platt, L. S. (2021). Examining risk and crisis communications of government agencies and stakeholders during early-stages of COVID-19 on twitter. *Computers in Human Behavior*, 114, 106568. <https://doi.org/10.1016/j.chb.2020.106568>
- Weltgesundheitsorganisation. (2020a). *Coronavirus press conference 11 february, 2020*.
- Weltgesundheitsorganisation. (2020b). *WHO erklärt COVID-19-Ausbruch zur Pandemie*. <https://www.euro.who.int/de/health-topics/health-emergencies/coronavirus-covid-19/news/news/2020/3/who-announces-covid-19-outbreak-a-pandemic>
- Witten, I. H., Frank, E., & Hall, M. (2011). *Data mining: Practical machine learning tools and techniques*. Elsevier. <https://doi.org/10.1016/C2009-0-19715-5>
- Wu, J. (2017). *Introduction to convolutional neural networks*.
- Yadav, A., & Vishwakarma, D. K. (2020). Sentiment analysis using deep learning architectures: A review. *Artificial Intelligence Review*, 53(6), 4335–4385. <https://doi.org/10.1007/s10462-019-09794-5>
- Yun, G. W., Morin, D., Park, S., Joa, C. Y., Labbe, B., Lim, J., Lee, S., & Hyun, D. (2016a). Social media and flu: Media twitter accounts as agenda setters.

International Journal of Medical Informatics, 91, 67–73. <https://doi.org/10.1016/j.ijmedinf.2016.04.009>

Yun, G. W., Morin, D., Park, S., Joa, C. Y., Labbe, B., Lim, J., Lee, S., & Hyun, D. (2016b). Social media and flu: Media twitter accounts as agenda setters.

International Journal of Medical Informatics, 91, 67–73. <https://doi.org/10.1016/j.ijmedinf.2016.04.009>

Zhang, X., Fuehres, H., & Gloor, P. A. (2011). Predicting stock market indicators through twitter “i hope it is not as bad as i fear”. *Procedia - Social and Behavioral Sciences*, 26, 55–62. <https://doi.org/10.1016/j.sbspro.2011.10.562>

Zheludev, I., Smith, R., & Aste, T. (2014). When can social media lead financial markets? *Scientific Reports*, 4(1), 4213. <https://doi.org/10.1038/srep04213>

Eidesstattliche Erklärung

Ich versichere an Eides statt durch meine Unterschrift, dass ich die vorstehende Arbeit selbständig und ohne fremde Hilfe angefertigt und alle Stellen, die ich wörtlich oder annähernd wörtlich aus Veröffentlichungen entnommen habe, als solche kenntlich gemacht habe, mich auch keiner anderen als der angegebenen Literatur oder sonstiger Hilfsmittel bedient habe. Die Arbeit hat in dieser oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen.

Düsseldorf, den 04.02.2021 Name: Paul Drecker