Universität Leipzig
Institut für Informatik
Studiengang B.Sc. Digital Humanities

Methoden des Resamplings von Twitter-Daten zur Approximation von Meinungsdynamiken in der Gesamtbevölkerung

Bachelorarbeit

Mathias Halbauer
geb. am: 18.05.1988 in Wolfen

Matrikelnummer 3737017

1. Gutachter: Prof. Dr. Martin Potthast
2. Gutachter: Dr. Jens Kersten

Datum der Abgabe: 5. September 2022

Erklärung

Hiermit versichere ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Leipzig, 5. September 2022

. . . . . .
Mathias Halbauer

Zusammenfassung

In dieser Arbeit wurde eine verfeinerte Methodik zur Approximation von Umfragedaten entwickelt. Dazu wurden 4 verschiedene Resamplingverfahren angewandt. Bei dem ersten Verfahren (Tweet-basiert) wurden allein die geocodierten Tweets eines Gebiets verwendet. Im zweiten Verfahren (Stratifiziert Tweet-basiert) wurde eine stratifizierte Auswahl dieser Tweets vorgenommen. Im dritten Verfahren (Populations-basiert) wurden alle Tweets (auch nicht geocodierte) der User eines bestimmten Gebiets verwendet. Im vierten Verfahren (Populations-geocodiert) wurden allein die Geo-Tweets der User eines bestimmten Gebiets verwendet. Dazu wurde in 2 Experimenten diese Methodik anhand von Großereignissen angewandt. In diesen Experimenten wurden die Twitter-Daten mit zusätzlichen Umfragedaten kombiniert, um die Resamplingverfahren evaluieren zu können. Hierbei wurde die Anpassung der Twitterzeitreihen aus dem Resampling an die Umfragezeitreihen gemessen. Im ersten Experiment wurden Twitter-Daten zum Thema Corona mit Daten einer Umfrage zu psychischen Belastungen in Folge von Corona verglichen. Im zweiten Experiment wurde die Methodik, am Beispiel von Tweets zur Bundestagswahl 2021, angewandt. Dabei wurde geprüft, ob die Bewertungen von Parteien durch Twitteruser mit Wahlabsichten aus einer Umfrage übereinstimmen. Für das erste Experiment zeigte sich ein kleiner Gewinn der Güte der Annäherung an die Umfragedaten. Jedoch konnte bei dem zweiten Experiment keine Übereinstimmung zwischen den geresampelten Twitter-Daten und den Wahlumfragedaten festgestellt werden.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Einleitung

Die Gewinnung von reliablen Daten zur Erfassung von Meinungstrends ist häufig sehr aufwendig. Jedoch benötigen Politik und zivilgesellschaftliche Organisation (NGOs) diese, um den Meinungsstand in der Bevölkerung für verschiedenste Themen, wie Corona oder Bewertungen politischer Parteien, zu erfassen. Klassischerweise werden dazu Umfragen durchgeführt. Aufgrund von Forschung in diesem Bereich existieren Methodiken, um die Verallgemeinerbarkeit der aus den Daten gewonnenen Ergebnisse zu gewährleisten. Dies beinhaltet u.a. systematische Ansätze zur Stichprobenziehung und Planung der Interviews, um unterschiedliche Auswahlwahrscheinlichkeiten für Personen in der Stichprobe, auszugleichen. Der Nachteil dabei ist der zeitliche Aufwand und die finanziellen Kosten. Es kann eine Analyse nur durchgeführt werden, wenn die vorgesehenen Interviews abgearbeitet wurden. Zudem ergeben sich Kosten für Interviewer und technische Infrastruktur, wie Telefonlabore. Diese Problematik verstärkt sich, wenn viele Erhebungszeitpunkte gewünscht werden, um mittel- und langfristige Trends zu erkennen. Dies wird versucht mit Online-Umfragen zu umgehen. Allerdings ist hier die Gefahr der Selbstselektion der Interviewten sehr groß. Außerdem wurden aufgrund des Corona-Virus Umfragen mit persönlichen Interviews vor Ort in das Jahr 2021 verschoben. Davon war zuletzt die Allgemeine Bevölkerungsumfrage (ALLBUS) betroffen GESIS [2022c].

Aufgrund der genannten Einschränkungen sind andere Datenquellen als Alternativen zu Umfragen interessant. Die APIs von Social Media-Plattformen, wie Twitter und Reddit, ermöglichen den Zugang zu Real Time-Daten ohne selber aufwändige Erhebungen durchzuführen. Dies verkürzt den Zeitraum von dem Beginn einer Untersuchung bis zur Veröffentlichung der Ergebnisse. Zudem können Forscher auf die Daten ohne weitere Kosten zugreifen. Zusätzlich lassen sich große Mengen an Daten (relativ) einfach mit verhältnismäßig wenig Aufwand gewinnen. Dies erleichtert die Bereitstellung zeitkontinuierlicher

Daten zur Extrapolation von Trends.

Allerdings gibt es in diesem Fall keine Kontrolle über die Auswahl von Befragten, hier die Social Media-User. Zum einen existiert ein Selbstselektionsbias, weil nur Personen mit einem Konto auf einer Social Media-Plattform, in die Stichprobe eingehen können. Zudem unterscheidet sich die Nutzung von Social Media nach verschiedenen sozio-demografischen Variablen wie Alter, Bildung oder Geschlecht. Studien haben dies für verschiedene Länder für die Plattform Twitter gezeigt. Hierbei für die USA Wojcik and und Shawnee Cohn [2019], für das Vereinigte Königreich Baghal et al. [2021] und für Deutschland Beisch and Schäfer [2020]. Für Umfragen wird dagegen geplant welche Personen aus welchen Haushalten in die Befragung eingehen. Zudem gibt es bei Social Media anders als bei Umfragen keine Kontrolle über die Generierung der Daten, da allein die User entscheiden, ob sie Datenmaterial bereitstellen. Es existiert damit ein Trade-Off zwischen bezüglich der Datenverfügbarkeit einfacheren Social Media-Analysen, deren Verallgemeinerbarkeit jedoch häufig unklar ist und besser generalisierenden Umfragestudien, die jedoch sehr zeit- und kostenaufwendig sind.

Die vorliegende Arbeit untersucht, ob eine verfeinerte Resampling-Methodik es ermöglicht Ergebnisse von (Meinungs)Umfragen mit Twitter-Daten zu reproduzieren. Dazu werden neue Stichproben aus den bestehenden Twitter-Daten gezogen, um den Effekt von Verzerrungen in den Twitter-Ausgangsdaten zu korrigieren. Hierbei werden Techniken aus der Umfrageforschung (Stratifikation, Panelbildung) angewandt, um Trends besser erfassen zu können. Der Vorteil einer verbesserten Methodik ist die Möglichkeit der Anwendung auch für Social Media-Analysen ohne thematisch ähnliche Umfragedaten. Diese sind nicht zu jeder Frage verfügbar, die mit Social Media analysiert werden soll.

Zu Beginn der Arbeit wird der Forschungsstand zu Studien, die versuchen soziale und politische Trends anhand von Social Media-Daten zu messen, referiert. Zudem wird auf methodische Herausforderungen der Kombination von Social Media- und Umfragedaten hingewiesen. Um notwendiges Hintergrundwissen beim Leser zu schaffen, wird im folgenden Kapitel auf Grundlagen der Umfrageforschung eingegangen. Anschließend wird die verwendete Resampling-Methodik dieser Arbeit erläutert. Es werden die verwendeten Resampling-Verfahren und ihre Motivation dargestellt. Dabei wird auf die Verfahren auf Tweet-Ebene (Tweet-basiert, Stratifiziert Tweet-Basiert) und die User-orientierten Verfahren (Populations-basiert, Populations-geocodiert) eingegangen. Diese Methodik wird dann in 2 Experimenten angewandt. Im ersten Experiment wird die Methodik auf das Thema Corona in einer regionalen Studie in Thüringen übertragen. Dazu werden mit der Methodik von Twitter gewonnene Daten mit Umfragedaten verglichen und ausgewertet. Die Frage ist dabei, welcher Grad der Annäherung der Twitter-Daten an die Umfragedaten

erreicht wird. Das zweite Experiment hat Wahldaten zum Gegenstand. Hier wird die Methodik auf Tweets zu Parteien anlässlich der Bundestagswahl angewandt. Es wird untersucht, ob ein Zusammenhang zwischen den Bewertungen der Parteien auf Twitter und ihren Ergebnisse in einer Wahlumfrage erkennbar ist. Zum Schluss werden die Erkenntnisse zusammengefasst und diskutiert. Dabei werden Probleme bei der Interpretation der Ergebnisse aufgezeigt.

Kapitel 2

Forschungsstand

In diesem Kapitel wird der für die Arbeit relevante Forschungsstand bei der Messung von Meinungen in der allgemeinen Bevölkerung auf Social Media dargestellt. Zuerst werden im ersten Unterkapitel relevante Studien zu Meinungsdynamiken, die allein mit Social Media-Daten arbeiten, dargestellt. Im darauffolgenden Kapitel werden die Möglichkeiten der Erweiterung der Datenbasis mit Umfragedaten dargestellt und auf die dabei bestehenden Herausforderungen eingegangen.

2.1 Abschätzung von Meinungsdynamiken mit Social Media

In dem Abschnitt werden Studien, die sich mit der Messung von Meinungsdynamiken auf Social Media beschäftigen, dargestellt. Diese Studien können als Ergänzung oder Ersatz für Umfragestudien verstanden werden, wobei dieses Ziel teils explizit erwähnt wird, teils sich aus den Arbeiten implizit ergibt. Da die Corona-Epidemie ein besonders einschneidendes Ereignis der letzten Jahre darstellt, beschäftigen sich sich viele solcher Studien mit der Messung von Sentiment in Folge von Corona. Schwerpunkt dabei ist der Gesichtspunkt, anhand welcher Attribute (u.a. Sprache, Geo-Tag) der Tweetcorpus generiert wurde. Twitter speichert neben dem Textinhalt eine Vielzahl von Metadaten zu den Tweets, wie u.a. Sprache (Attribut: “lang”) oder Ortsinformationen (Attribut: “geo”). Relevant sind dabei hauptsächlich Längsschnittstudien.

Kruspe et al. [2020] führten dazu eine vergleichende Studie verschiedener Länder (DEU, IT, FR, UK, ES) durch. Dabei verglichen sie das Sentiment corona-bezogener Tweets mit dem allgemeinen Sentiment auf Twitter als Baseline. Damit konnten sie corona-spezifische Effekte von der allgemeinen Entwicklung des Sentiments unterscheiden. Die Auswahl der Tweets erfolgte

anhand des Geo-Tags der Tweets, die den Ländern zugeordnet werden konnten. Mit dieser Methodik konnten schon explorativ Ergebnisse produziert werden und stellt damit einen guten Startpunkt dar.

In Ridhwan and Hargreaves [2021] wurde eine Analyse der Reaktionen auf Corona im Stadtstaat Singapur durchgeführt. Dazu wurden Tweets in der Anfangsphase von Corona, die für Singapur geocodiert, und auf Englisch verfasst wurden, ausgewählt. Die Autoren betrachteten den zeitlichen Verlauf der Anzahl positiver, negativer und neutraler Tweets. Dabei stellen sie fest, dass unerwarteterweise positive Tweets dominieren. Ein Nachteil ihrer Sampling-Methodik ist allerdings, dass sie eine Überabdeckung bestimmter Gruppen (englischsprachige Nutzer) und eine Unterabdeckung anderer Gruppen (nicht englisch-sprachige Twitter-Nutzer in Singapur) bewirkt. Twitter-User aus Singapur, die in anderen Sprachen als Englisch tweeten (u.a. Chinesisch, Malaisch) fallen hierbei aus der Analyse heraus. Auch das Kriterium, dass die Tweets einen Geo-Tag haben müssen, ermöglicht keine eindeutige Zuordnung der User. Im hypothetischen Fall, dass ein Tourist auf der Durchreise einen englischsprachigen Tweet über Corona postet, geht dieser auch in das Korpus ein. Sofern es darum geht die Meinungen von Bürgern Singapurs über das Thema Corona und eventuelle Maßnahmen in Singapur zu erfassen, ist dieser Tweet aber nicht relevant.

Weiterhin gibt es auch Studien, die die Tweets allein mit dem Sprachkriterium filtern, ohne Anwendung weiterer Kriterien vgl. dazu Priyadarshini et al. [2021]. Bei einer solchen Erhebung ist es kaum möglich die verwendeten Tweets auf einen konkreten geographischen Ort zurückzuführen. Gerade englischsprachige Tweets sind sicher nicht nur auf englischsprachige Länder beschränkt, sondern nehmen einen Anteil der Tweets in nicht englischsprachigen Ländern ein. Somit sind solche Studien für diese Arbeit weniger relevant. Damit ist eine Erfassung des Topics (Topic-Coverage) in Social Media möglich, aber keine Erfassung einer dahinterliegenden Population (Population Coverage) möglich. Zu diesen Begriffen siehe Hsieh and Murphy [2017]. Techische Fragen, wie verwendete Attribute für die Datenfilterung, sind damit immer von der Forschungsfrage abhängig.

Um Meinungsdynamiken mit Social Media anstelle von Umfragen zu erfassen, wurde zudem eine Studie für die Niederlande durchgeführt. Wang et al. [2020] erfassten dazu Sentiment-Trends in den Niederlanden bezüglich Corona-Maßnahmen. Datenbasis war neben Twitter zudem Reddit und die Diskussionsplattform nu.nl. Die Tweets wurden mithilfe des Sprachtags selektiert. Aus diesen Daten wurden wiederum Tweets mit Keywords zu bestimmten Maßnahmen ausgewählt und dabei das durchschnittliche Sentiment als Graph dargestellt. Ähnlich wie in Kruspe et al. [2020] wurde zudem das Sentiment aller Tweets mit den keyword-gefilterten Tweets verglichen.

Abgesehen von der Bedeutung des verwendeten Attributs in der Stichprobenziehung, stellt sich die Frage der Art und Weise der Stichprobenziehung. Die Autoren in Xu et al. [2015] unterscheiden dabei 3 verschiedene Sampling-Strategien: Random Sampling, Stratified Sampling, Community Sampling. Random Sampling stellt eine einfache Zufallsstichprobe dar. Bei dem Stratified Sampling werden die User nach bestimmten Eigenschaften aufgeteilt und anschließend innerhalb dieser Strata zufallsbasiert gesampelt. Dies soll verhindern, dass in den Daten kleine Gruppen gar nicht vertreten sind. Community Sampling sampelt dagegen alle User, die miteinander interagieren. Dazu werden zuerst alle User eines bestimmten Gebietes ausgewählt und daraus die Twitter-Interaktions-Graphen konstruiert. Bezüglich der Auswirkung der Verfahren stellen die Autoren dabei einen Einfluss der Topic Häufigkeit auf die Effektivität von Random Sampling fest. Dieses erfasst Topics mit vielen Daten und längere Zeitspannen besser, als wenn wenig Daten vorhanden sind. Sie sprechen sich in dem Paper für Community Sampling eines Netzwerkes von Usern aus, da dies am besten den Retweet-Graphen erhält.

Eine Herausforderung dieser Studien besteht darin die gewonnen Ergebnisse zu validieren. Denn diese sind stark abhängig vom ausgewählten Untersuchungsdesign. Jedoch lässt sich allein mit den Social Media-Daten nicht quantifizieren, wie gut die Offline-Trends erfasst wurden.

2.2 Kombinierung von Social Media und Umfragedaten

In diesem Teil soll der Stand der Forschung zur Kombinierung von Social Media-Daten mit Umfragedaten behandelt werden. Der Vergleich von Social Media Daten mit Umfragedaten ermöglicht die gewonnen Erkenntnisse aus ersterem zu validieren. Somit lässt sich besser evaluieren, ob mit Social Media reale Offline-Trends erfasst werden oder nur Trends, die allein auf Social Media stattfinden. Dazu wird zuerst auf grundlegende Herausforderungen eingegangen und anschließend werden Anwendungsstudien in verschiedenen Feldern, wie Wahlen und Gesundheit, referiert.

Beuthner et al. [2021] gehen in einem Überblicksartikel auf die Möglichkeiten und Probleme Umfragedaten mit zusätzlichem Datenmaterial (Social Media, Sensordaten, Geodaten) zu erweitern. Dies behandeln sie aus Perspektive der Umfrageforschung. Für diese Arbeit ist hauptsächlich die Möglichkeit der Ergänzung mit Social Media relevant, wobei hierbei auch Geo-Daten vorhanden sein können, wie Geo-Tags bei Twitter. Die Autoren unterscheiden nach dem Level der Erhebung zwischen Individual- und Aggregatdaten. Erstere haben Personen als Untersuchungseinheit, wogegen Aggregatdaten sich auf

z.B. auf Gruppen von Personen beziehen können und Forscher Informationen über diese Gruppen, wie Populationsmittelwerte oder -verteilungen extrahieren. Diese sind selber aber nicht mehr auf die Personen zurückführbar. Individualdaten liegen dann vor, wenn der User-Handle bei einer Umfrage abgefragt wird. Bei Aggregat-Daten muss dagegen nur ein Kriterium vorliegen, wonach die Populationen aus den Datenquellen zusammengeführt werden können. Dies kann die Zugehörigkeit zu einer bestimmten geographischen Region oder sozialen Gruppe sein. Ein Faktor, der bei der Arbeit mit Geo-Daten zu beachten ist, dass gemäß den Autoren, granulare Geo-Daten (u.a. Adressen und Geo-Koordinaten) aus Datenschutzgründen nicht zusammen mit den Umfragedaten gespeichert werden dürfen. In diesem Fall muss die auswertende Organisation eine datenschutzkonforme Speicherung gewährleisten, indem z.B. die Genauigkeit von Geo-Daten verringert wird Beuthner et al. [2021]. Damit ergeben sich Einschränkungen im Zugang zu den Daten, sodass ein freier Zugang leider häufig nicht möglich ist. Die Verfügbarkeit von Umfragedaten ist damit sehr eingeschränkt, was den Zugang zu Datensätzen und die Nutzung datenschutzrelevanter Attribute (z.B. Wohnort der befragten Person) angeht.

In Baghal et al. [2021] wurden in einer Studie Twitter- und Umfragedaten ex ante bei der Datenerhebung kombiniert. Dazu wurde bei der Datenerhebung bereits ein verlinkter Datensatz mit Antworten aus Umfragen und den zugehörigen Twitter-Profilen der Befragten erstellt. Die Autoren untersuchten Zusammenhänge zwischen den Eigenschaften der Twitter-Profile (z.B. Anzahl der Follower) und Eigenschaften (z.B. Lebenszufriedenheit), die in der Umfrage erhoben wurden. Die Ergebnisse einer logistischen Regression zwischen dem Sentiment der Tweets und der abgefragten Lebenszufriedenheit in der Umfrage zeigen unerwarteterweise einen positiven Zusammenhang zwischen der Anzahl negativer Tweets und Lebenszufriedenheit. Weiterhin stellen sie fest, dass dieser Effekt durch die Drittvariable Quantität der Tweets beeinflusst wird. Zudem stellen sie fest, dass die Häufigkeit negativ gelabelter Tweets sich nach Geschlecht der User unterscheidet. Frauen posten hierbei mehr negative Tweets als Männer. Sie weisen daraufhin, dass die Messung des Sentiments von Usern mit verschiedenen weiteren Variablen korreliert ist, welche das Messinstrument Sentimentanalyse von Tweets „verzerren“. Allerdings ist bei der Bewertung zu beachten, dass die Datenbasis recht gering ist. Bei annähernd 2000 Befragten im Innovation Panel, haben von den 171 Befragten, die einer Zusammenführung der Daten zugestimmt haben, nur 127 einen öffentlichen Twitter-Account mit mindestens einem Tweet.

O’Connor et al. [2010] gehen in ihrer Studie auf die Probleme im Vergleich von Textklassifikationen mit Umfragedaten ein. Dazu vergleichen sie das Sentiment auf Twitter mit verschiedenen Umfragen zu Konsumentenvertrauen und Wahlen in den USA. Hierbei ist es notwendig eine Glättung der Twitterzeit-

reihen durchzuführen, um das Hintergrundrauschen zu reduzieren. Statt einer Schätzung für jeden Datenpunkt verwenden sie einen Moving-Average für k Tage. Damit wird der Effekt von kurzfristigen Schwankungen in den Daten abgeschwächt und längerfristige Tendenzen betont. Jedoch gehen damit allerdings auch feinere Unterschiede verloren. Weiterhin schlagen sie einen Lag-Parameter L vor, um Sentiment-Daten mit Umfragedaten an späteren Tagen zu vergleichen. Mithilfe dieses Verfahren werden Korrelationen der Zeitreihen für unterschiedliche Werte von L berechnet und somit die beste zeitliche Übereinstimmung gesucht. Auf diese Weise lässt sich feststellen, welcher der beiden Zeitreihen zeitlich führend ist und welche folgt. Ohne diese Vorverarbeitung ist es möglich, dass bestehende Zusammenhänge in Zeitreihen übersehen werden, da aufgrund von Faktoren bei der Erhebung, z.B. unterschiedliche Zuordnung zu Erhebungszeitpunkten, die Zeitreihen zeitlich auseinander fallen.

Ein anderer Ansatz ist es statistische Daten aus Umfragen zu nutzen, um den Sampling Bias von Social Media-Daten zu reduzieren. Cui and He [2021] führten dies im Kontext von Daten zu Reisen in Kalifornien durch. Dazu werden Informationen über die Häufigkeiten bestimmter sozio-demografischer Attribute, wie Geschlecht und Alter, für das Resampling von diesen Personengruppen in den Twitter-Daten genutzt. Es wurden zuerst die Häufigkeiten bestimmter sozio-demografischer Attribute in den Survey-Daten bestimmt. Die Häufigkeiten der sozio-demografischen Attribute in den Twitter-Daten wurde mithilfe von gelinkten Facebook-Profilen, die diese Informationen enthalten oder wurden durch verschiedene Klassifikationsmodelle (u.a. Support-Vektor-Maschinen, Neuronale Netze) bestimmt. Zur Evaluierung des Verfahrens werden die geresampelten Twitter-Daten mit den Umfragedaten verglichen. Dazu werden die Wahrscheinlichkeitsverteilungen für Reisedauer und Ziel der Reise (Schule, Erholung etc.) untersucht. Die Autoren kommen zu dem Schluss, dass das durchgeführte Resampling eine bessere Repräsentativität des Social Media-Datensatzes ermöglicht.

2.3 Methodik von Umfragen

In diesem Kapitel geht es darum Grundlagen der Umfrageforschung zu erläutern, um die Unterschiede zur Erhebung von Social Media-Daten deutlicher zu machen. Weiterhin wird auf einige Probleme und Einschränkungen bei der Erstellung von Umfragen und der Interpretation der Ergebnisse eingegangen. Dabei konzentriere ich mich auf Aspekte, die der Umfrageforschung und der Social Media-Forschung gemein sind. Der erste Aspekt ist die Frage, wie Daten in Stichproben ausgewählt werden, um Verzerrungen zu vermeiden. Der zweite Aspekt ist die Verfügbarkeit von Daten im Zeitverlauf. Diese Ausfüh-

rungen sollen den Transfer von der Methodik der empirischen Sozialforschung zur Social Media-Forschung deutlich machen.

Ein generelles Problem, wenn keine Vollerhebung möglich ist, besteht in der Auswahl der zu befragenden Personen. Ein zentrale Problematik bei der Durchführung von Umfragen ist demzufolge die Stichprobenziehung. Eine Stichprobe bezieht sich immer auf die zugrundeliegende Grundgesamtheit. Diese ist wiederum abhängig von der Fragestellung. Daraus soll ein verkleinertes Abbild, die Stichprobe, gezogen werden [Schumann, 2012, S.84]. Zum Beispiel für Wahlumfragen zur Bundestagswahl besteht diese aus allen deutschen Staatsbürgern, die mindestens 18 Jahre alt sind. Eine vollständige Erhebung ist in den seltensten Fällen praktikabel, es sei denn die interessierende Grundgesamtheit ist sehr klein [Schnell et al., 2013, S.258]. Systematische Unterschiede zwischen der Grundgesamtheit/Auswahlpopulation und der Stichprobe werden als Under- oder Overcoverage bezeichnet. Undercoverage liegt vor, wenn interessierende Personen aus der Grundgesamheit nicht in der Stichprobe vorhanden sind. Overcoverage dagegen bedeutet, dass Personen in der Stichprobe enthalten sind, die aber nicht in der zugehörigen Grundgesamtheit enthalten sind [Diekmann, 2018, S. 377] Ein Beispiel: Personen, die keine deutsche Staatsbürgerschaft haben, und in Wahlumfragen befragt werden. Dabei ist es einfacher den Overcoverage von Gruppen zu identifizieren als den Undercoverage [Diekmann, 2018, S. 378].

Dabei existieren verschiedene Verfahren, um Stichproben zu ziehen. Erstens ist es möglich mittels Wahrscheinlichkeitsauswahl eine Zufallsstichprobe zu ziehen. Zweitens gibt es auch bewusste Auswahlen von Untersuchungseinheiten, wenn nach Quoten für bestimmte Merkmale (z.B. Geschlecht, Altersgruppen) ausgewählt wird. Als drittes Verfahren existiert zudem die willkürliche Auswahl, wobei die Auswahl der Untersuchungseinheiten nicht kontrolliert wird.

Eine Wahrscheinlichkeitsauswahl bedeutet, dass jedes Element der Grundgesamtheit die gleiche Wahrscheinlichkeit hat in die Stichprobe einzugehen [Schumann, 2012, S. 86]. Grundsätzlich sind Wahrscheinlichkeitsauswahlen zu bevorzugen, da die Abschätzung des Stichprobenfehlers und von Konfidenzintervallen auf einer Zufallsauswahl basiert. Die Bestimmung von Maßen zum Schluss von der Stichprobe auf die Grundgesamtheit (Interferenzstatistik) ist allein mit diesem Stichprobenverfahren möglich [Schnell et al., 2013, S. 258]. Jedoch ist dieses Auswahlverfahren auch recht aufwendig, da sichergestellt werden muss, dass die Auswahl von Orten, Haushalten und Haushaltsmitglieder kontrolliert durchgeführt wird, um die gleichen Auswahlwahrscheinlichkeiten sicherzustellen. Dazu werden Orte in Stimmbezirke eingeteilt und proportional zu ihrer Größe gesampelt. Die Adressen werden mittels Random Route ausgewählt und Haushaltsmitglieder werden zufallsbasiert ausgewählt mittels Geburtstagsverfahren oder Schwedenschlüssel [Diekmann, 2018, S. 384].

Bei den dabei entstehenden Zufallsstichproben ergibt sich in bestimmten Fällen das Problem, dass relevante Gruppen gering oder gar nicht in der Stichprobe vertreten sind. Ist die Verteilung der Gruppen in der Stichprobe bekannt, ist es möglich die Stichprobe zu stratifizieren. Dazu werden die Gruppen in der bereits erhobenen Stichprobe getrennt gesampelt. Für jede Gruppe wird dann unabhängig eine Zufallsstichprobe gezogen [Schumann, 2012, S. 93]. Hierbei wird zwischen proportional und disproportional geschichteten Stichproben unterschieden. Bei einer proportional geschichteten Stichprobe ist die Anzahl der Befragten in der Stichprobe proportional zu ihrem Anteil in der Grundgesamtheit. In disproportional geschichteten Stichproben wird dagegen immer die gleiche Anzahl an Befragten pro Schicht erhoben, unabhängig von ihrem Anteil an der Grundgesamtheit. Eine korrekte Schätzung von Populationsparametern ohne zusätzliche Nachgewichtung ist nur im ersten Fall möglich [Schumann, 2012, S. 94]. Die Nachgewichtung erfolgt mit der inversen Auswahlwahrscheinlichkeit gemäß der Formel 2.1.

$$\text{Inverse Auswahlwahrscheinlichkeit} = \frac{\text{Größe der Schicht in der Grundgesamtheit}}{\text{Stichprobengröße der Schicht}} \tag{2.1}$$

Anstelle der Korrektur des Undercoverages in Zufallsstichproben werden stattdessen auch häufig Auswahlen nach Quoten durchgeführt. Jedoch ist hier das Problem, dass nicht nach allen potentiell relevanten sozio-demografischen Variablen quotiert werden kann [Schumann, 2012, S. 98]. Stellt sich nachträglich heraus, dass z.B. für die Ergebnisse der Bildungsstatus relevant ist, aber nicht danach quotiert wurde, liegt hier eine Verzerrung des Samples vor. Umfragen mit willkürlicher Auswahl sind nicht weiter relevant, da sie keine methodisch kontrollierte Abbildung einer Grundgesamtheit ermöglichen [Schumann, 2012, S. 97].

Weiterhin werden Erhebungen nach Art des Designs unterschieden. Erhebungsdesigns bestimmen für eine Erhebung welche Untersuchungseinheiten (hier: Personen) zu welchen Zeitpunkt erhoben werden. Dabei werden entweder Querschnitt-, Trend- oder Paneldesigns verwendet. Querschnitt-Designs stellen einmalige Erhebungen zu einem Zeitpunkt dar [Diekmann, 2018, S. 304]. Davon werden Trenddesigns unterschieden, bei denen zu mehreren Zeitpunkten wieder neue Stichproben gezogen werden. Werden bei Studien immer wieder die gleiche Stichprobe an Personen verwendet, spricht man von einem Paneldesign [Schumann, 2012, S. 111]. Der Vorteil dieses Erhebungsdesigns ist, dass Änderungen auf einer individuellen Ebene z.B. von Einstellungen oder des Erwerbsstatus, erkennbar sind, wogegen bei einem Trenddesign nur Informationen über die interviewte Population als Ganzes vorliegen. Jedoch ist der Nachteil, dass die Pflege von Panels recht aufwendig ist, wenn bisher Interviewte für zukünftige Befragungen nicht mehr verfügbar sind (Panelmor-

talität). Die Gründe dafür können Wegzug, Tod oder Nicht-Antworten sein [Schumann, 2012, S. 112]. Dadurch entstehen höhere Kosten bei der Erhebung mit einem Paneldesign. Die Kosten der Erhebung des sozio-ökonomischen Panel (SOEP), ein in der Soziologie sehr häufig verwendeter Datensatz, liegen im zweistelligen Millionenbereich [Diekmann, 2018, S. 311]. Bei den Querschnittsstudien (Trend- oder Paneldesign) liegen oftmals größere Zeiträume, je nach Studie kann dies mehrere Wochen bis zu 2 Jahre sein, zwischen den Erhebungszeitpunkten. Dies ist zu beachten ist, wenn ein Vergleich mit zeitkontinuierlichen Social Media-Daten angestrebt wird. Bei letzterem liegen idealerweise Daten für jeden einzelnen Tag vor. Der Informationsgehalt einer Umfrage unterscheidet sich je nach dem Design. Paneldesigns ermöglichen Rückschlüsse auf Individuen und ihre Entwicklungen. Trenddesigns ermöglichen nicht diese Art von Rückschlüssen, wobei aber Entwicklungen von Populationen als Ganzes untersucht werden können. Dagegen ermöglichen Querschnittserhebungen nur die Untersuchung von Populationen zu einem gegebenen Zeitpunkt [Diekmann, 2018, S. 306].

Bei der Stichprobenziehung sind Zufallsstichproben vorzuziehen, da nur so die interferenzstatistischen Maße, wie Konfidenzintervalle und Standardfehler, berechnet werden können. Sofern Gruppen in einer Umfrage unterrepräsentiert sind und die Anteile dieser Gruppen in der Grundgesamtheit und Stichprobe bekannt sind, ist stratifiziertes Sampling hilfreich, um unverzerrte Schätzungen durchzuführen. Für die Erforschung von Entwicklungen im Zeitverlauf sind Paneldaten sehr nützlich, da sie die Analyse individueller Entwicklungen ermöglichen. Sofern diese Daten nicht vorhanden sind, sollte auf Trenddaten ausgewichen werden. Diese stellen eine einfachere Alternative zur Erfassung von Trends mittels Umfragen dar, lassen aber hierbei keinen Schluss auf veränderte individuelle Einstellungen zu. Damit sind auch die datenschutzrechtlichen Hürden für den Zugang niedriger, da anders als bei Paneldaten der Rückschluss auf die Personen in der Studie schwieriger ist. Dementsprechend ist der Zugang zu den Datensätzen zwecks Durchführung eigener Analyse einfacher.

Kapitel 3

Methodik

3.1 Resamplingverfahren

In diesem Kapitel werden die verschiedenen Resampling-Verfahren, die in dieser Arbeit angewandt wurden, dargestellt. Dazu werden die Verfahren beschrieben und die Motivation ihrer Verwendung anhand des Forschungsstandes erläutert. Für jedes Verfahren werden Hypothesen über den Zusammenhang zwischen Eigenschaften der Twitter-Daten und ihrer Repräsentanz formuliert. Die Verfahren variieren zum einen danach, welche Art von Tweets (geocodiert oder nicht geocodiert) in die Daten eingehen. Zum anderen wird unterschieden von welchen Usern die Tweets für das Resampling verwendet werden.

Verfahren 1: Tweet-basiert

Dieses Verfahren verwendet allein geocodierte Tweets. Um die Tweets der User einer bestimmten Gegend zu erfassen, werden die Tweets anhand des Geo-Labels gefiltert. Es werden nur Tweets verwendet, deren Geo-Label mit der untersuchten Region übereinstimmt. Die Zuordnung erfolgt hierbei auf Tweet-Ebene. Es wird damit keine User-Population konstruiert. Für einen gegebenen User werden nur die geocodierten Tweets verwendet, die in dem Gebiet gepostet wurden. Andere Tweets der User werden nicht verwendet. Ein Beispiel: Postet ein User ingseamt 100 Tweets, davon 3 Tweets, die Thüringen zugeordnet werden können und die restlichen 97 Tweets anderen Bundesländern. Damit werden nur die 3 erstgenannten Tweets verwendet, da anzunehmen ist, dass nur sie Aussagen über Thüringen enthalten. Das Verfahren nimmt im Anschluss keine weitere Reduktionen der Tweet-Menge vor. Dieses Verfahren wird in der Forschung häufig angewandt (vgl. Ridhwan and Hargreaves [2021]). Aufgrund der „Einfachheit“ lässt sich dieses Verfahren als Baseline betrachten. Die Hypothese ist, dass die anderen Verfahren bei der formellen Evaluation

mit den verwendeten Metriken eine bessere Annäherung an die Umfragedaten ermöglichen.

Verfahren 2: Stratifiziert Tweet-basiert

Im ersten Verfahren wurden noch alle Tweets, die in einem Bereich gepostet wurden, verwendet. Dabei entsteht das Problem, dass regionale Unterschiede in den Häufigkeiten der Tweets je Gebiet bestehen, sodass nicht alle Landkreise oder Bundesländer, gleichermaßen in der Stichprobe vertreten sind. Ergebnisse der Forschung in den USA zeigen, dass Personen in ländlichen Gebieten seltener auf Twitter vertreten sind Auxier and Anderson [2021, S. 7]. Anknüpfend an Kapitel 2.3 werden Verfahren aus der Umfrageforschung für diese Korrekturen genutzt. Dazu wird hier ein stratifiziertes Sampling genutzt. Die Häufigkeitsverteilung für die Bevölkerung in der Grundgesamtheit (z.B. nach Landkreisen) ist aus der amtlichen Statistik bekannt und mittels der Geo-Tweets ist die Häufigkeitsverteilung der Tweets in der Twitter-Stichprobe bekannt. Somit kann hier geschichtet gesampelt werden. Dabei fungieren die Landkreise/kreisfreie Städte als Schichten in diesem Resamplingverfahren. Dazu wird die Formel 2.1 in veränderter Form angewandt. Dabei fungiert die Einwohnerzahl je Landkreis als Größe der Schicht in der Grundgesamtheit. Die Größe der Schicht in der Stichprobe wird durch den Samplingparameter n (Anzahl der Tweets pro Schicht) angegeben. Dazu werden die Tweets nach der Herkunft getrennt gesampelt. In Bezug auf die verwendeten Tweetmengen von Verfahren 1, wird hierbei eine Teilmenge der geo-codierten Tweets der verwendet.

Verfahren 3: Populations-basiert

Die dahinterstehende Überlegung bei diesem Verfahren ist die Bewohner einer Region zielgerichtet einzufangen. Eine Problematik bei der Interferenz von Wohnorten der Twitter-User ist die Tatsache, dass eine standardisierte Geocodierung in den Twitter-Daten nur auf der Ebene von Tweets und nicht von Usern vorliegt. Für User existiert ein Feld „location“ in den User-Daten. Dies ist aber ein Freitextfeld, wo die User selbst einen Text eintragen können. Dagegen enthalten die geocodierten Tweets GPS-Koordinaten des Geräts, mit dem ein Tweet gepostet wurde. Diese GPS-Koordinaten wurden vorab dem jeweiligen Landkreis/kreisfreie Stadt bzw. Bundesland zugeordnet. Dazu wurde die Übereinstimmmung mit der Bounding Box der Landkreise bzw. Bundesländer bestimmt. Damit liegt eine Liste von Standorten für die User vor. Die Annahme bei dem Populations-basierten Verfahren ist, dass User tendenziell eher über Ereignisse an ihrem Wohnort tweeten. Um einen eindeutigen Standort zu erhalten, wird der häufigste Standort (statistischer Modus der Häufigkeitsver-

teilung) als User-Standort für das Resampling verwendet. Dies verändert die räumlichen Verteilungen der Tweets, indem Tweets neu den Landkreisen/Bundesländern der User mit den inferierten Standorten zugeordnet werden. Bei Studien, die allein Tweet-basiert den Standort verwenden, ergeben sich spezielle Probleme. User, die sich im Urlaub befinden oder auf der Durchreise sind, werden trotzdem den Gebieten zugeordnet. Mitchell et al. [2013] fanden bei ihrer Studie zum Sentiment in amerikanischen Bundesstaaten ein hohen Sentimentwert für die Gesamtheit der Tweets in Hawaii, dabei waren aber Tweets von Urlaubern überrepräsentiert.

Für dieses Verfahren wurden vorab mittels der Abfrage der User-Timelines in der Twitter-API zusätzlich nicht geocodierte Tweets heruntergeladen. Das Verwenden der gesamten User-Timelines, die auch nicht-geocodierte Tweets enthalten, ermöglicht eine höhere Tweetmenge und damit (idealerweise) auch eine bessere zeitliche Abdeckung. Mit diesen Daten lässt sich ein Panel (siehe hierzu Kapitel 2.3) von Twitter-Usern erstellen, die wiederholt tweeten. Damit sollten sich Veränderungen auf der Ebene einzelner User fassen lassen. Dies können mögliche Veränderungen des Sentiments in Folge von langfristigen Entwicklungen z.B. Stimmungsveränderungen im Verlauf der Corona-Epidemie, sein. Besonderer Vorteil gegenüber Umfragen ist, dass Twitter-Daten bei (vielen) Usern viele Messungen zu vielen Zeitpunkten ermöglichen. Dagegen sind bei Panelumfragen aus Gründen der Praktikabilität nicht immer dermaßen viele Erhebungen möglich. Eine Ausnahme stellen allein online durchgeführte Panels dar. Das Problem der Panelmortalität stellt sich hier in anderer Form dar. Da die Daten der User bereits vorliegen, entfällt das Problem der Non-Response, jedoch ist es möglich, dass User ihre Accounts löschen, auf privat umstellen oder nicht mehr tweeten.

Verfahren 4: Populations-geocodiert

Dieses Verfahren bedient sich der gleichen User-Zuordnung, wie das vorherige Verfahren. Jedoch besteht der Unterschied darin, dass hierbei allein die geocodierten Tweets, statt der vollständigen User-Timelines zusätzlich mit den nicht-geocodierten Tweets, verwendet werden. Quantitativ betrachtet ist die Tweetmenge bei diesem Verfahren für eine gegebene Userpopulation eine Teilmenge der Tweets von Verfahren 3 für diese Population. Der Vergleich der Ergebnisse mit Verfahren 3 ermöglicht es verschiedene Effekte voneinander zu unterscheiden. Dabei lassen sich folgende Hypothesen unterscheiden:
1) Sollte das Panel eine bessere Abbildung der Stimmungen der User ermöglichen, wird Verfahren 3 verglichen mit Verfahren 4 einen besseren Fit der Daten bewirken.
2) Sollte der Effekt der Bestimmung der User-Orte größer sein, sollten Verfah-

ren 3 und Verfahren 4 ähnliche Ergebnisse erzielen.

3.2 Sentimentklassifikation und Datenaggregation

Das verwendete Sentiment-Modell Guhr et al. [2020] wurde von den Entwicklern mit verschiedenen Datenquellen trainiert. Dazu gehören u.a. Tweets, Film- und Hotelbewertungen, Wikipedia-Einträge. Die Entwickler geben für das Modell einen F1-Score von 0.9744 an. Das German-Sentiment-Modell wurde in 2 Varianten entwickelt: Ein Modell basierend auf **FastText** und ein Modell basierend auf **BERT**. Das öffentlich verfügbare Modell als Python-Package basiert auf der letzteren Variante. Die Verwendung des Modells ist unkompliziert möglich. Ein Preprocessor ist bereits enthalten, sodass die Tweets als Liste von Strings dem Modell übergeben werden können. Dabei werden Tweets als entweder negativ, positiv oder neutral gelabelt.

Aufbauend auf den Sentiment-Klassifikationen der einzelnen Tweets wurde ein Mittelwert des Sentiment in dem gewählten Zeitintervall berechnet. Die Motivation dabei ist die Minimierung des Rauschens in den Daten aufgrund tagesabhängiger Fluktuationen. Dies ermöglichte zudem eine bessere Exploration der Graphen der Sentimentzeitreihen in den Plots, da die Bewegungen der Graphen eindeutiger sind. Dabei werden die Anteile positiv und negativ gelabelter Tweets verrechnet, um das durchschnittliche Sentiment in einer Zahl ausdrücken zu können. Die Berechnung erfolgte anhand dieser Formel:

$$\text{Average-Sentiment} = \frac{\text{Anzahl positiver Tweets}}{\text{Anzahl aller Tweets}} - \frac{\text{Anzahl negativer Tweets}}{\text{Anzahl aller Tweets}} \tag{3.1}$$

Der Wertebereich der Variablen wurde zudem normalisiert, um die Vergleichbarkeit von den Werten aus der Umfrage mit den Sentiment-Daten zu ermöglichen. Damit wurde der Wertebereich der aggregierten Scores auf 0 - 1 eingegrenzt.

$$z_i = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} \tag{3.2}$$

3.3 Metriken

Für die formale Evaluierung der Resampling-Verfahren wurden verschiedene Metriken verwendet. Diese werden folgend beschrieben und die Motivation ihrer Verwendung erläutert. Dabei wird auch auf die Interpretation der Werte und ihre Wertebereiche eingegangen.

Zur Berechnung der Distanz der Datenpunkte aus verschiedenen Zeitreihen wurde der Mean-Squared-Error verwendet. Dieser berechnet den Durchschnitt der quadrierten Distanz der Werte y von dem geschätzten Wert $\hat{y}$ und ist damit unabhängig vom Vorzeichen der Differenz.

$$\mathrm{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=1}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2 \tag{3.3}$$

Daneben wurde die Pearson-Korrelation verwendet, um die Richtung der Graphen zu erfassen. Der Mean-Squared-Error bewertet nur den Abstand zweier Graphen, aber nicht die Richtung der Entwicklung. Jedoch sollte ein gutes Resampling-Verfahren auch die zeitliche Entwicklung des Stimmungsbildes in der Umfrage erfassen. Um dies zu messen wird die Pearson-Korrelation verwendet. Der Wertebereich des Pearson-Korrelationskoeffizienten verläuft zwischen -1 und +1. Positive Werte für r zeigen einen gleichläufigen Zusammenhang an, wogegen negative Werte einen gegenläufigen Zusammenhang anzeigen. Dazu wird für die Werte x und y die Differenz zu dem Mittelwert $\bar{x}$ bzw. Mittelwert $\bar{y}$ berechnet.

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}} \tag{3.4}$$

Zusätzlich wurde die Granger-Kausalität berechnet, um unterschiedliche zeitliche Verschiebungen feststellen zu können. Dabei wird gemessen, ob eine gegebene Zeitreihe eine andere Zeitreihe vorhersagen kann. Die Twitter-Daten haben einen Timestamp, der sekundengenau das Datum angibt. Jedoch wird bei den Werten zur Umfrage zwar ein Datum angegeben, aber es ist unwahrscheinlich, dass die Umfrage genau und an diesem Datum durchgeführt wurde. Es ist realistischerweise davon auszugehen, dass dies in einem Zeitfenster in Umgebung des Datums geschah. In O'Connor et al. [2010] wurde für Wahldaten gezeigt, dass manchmal eine zeitliche Verzögerung in der Erhebung stattfindet und diese Verzögerung in die Analyse einbezogen werden muss, um eine Übereinstimmung der Datenreihen festzustellen.

$$\mathbb{P}[Y(t+1) \in A \mid \mathcal{I}(t)] \neq \mathbb{P}[Y(t+1) \in A \mid \mathcal{I}_{-X}(t)] \tag{3.5}$$

Hierbei wird auf eine Kausalität von X auf Y getestet. Es wird geprüft, ob es einen Unterschied zwischen der Wahrscheinlichkeit für Y zum Zeitpunkt t+1 Teil der beliebigen Menge A zu sein bei gegebenen Informationen (inklusive X), von der Wahrscheinlichkeit von Y zum Zeitpunkt t+1 in A enthalten zu sein bei gegebenen Informationen (ohne X) unterscheidet. Für die Interpretation ist der p-Wert der verwendeten Teststatistik relevant. Unterschreitet der p-Wert des gewählten Tests den Wert des gewählten Signifikanzniveaus (p= 0,05 oder

p= 0,01) kann die Nullhypothese der fehlenden Kausalität für die gegebene zeitliche Verschiebung verworfen werden.

Kapitel 4

Experiment 1: Messung von Corona-bezogenen Belastungen

Folgend soll das erste Experiment zum Thema Corona vorgestellt werden. Dabei werden Umfragedaten zu psychischen Belastungen, in Folge von Corona, mit Twitter-Sentiment-Daten verglichen, die beide in Thüringen erhoben wurden. Diese Fallstudie hat das Ziel, die vorab beschriebene Methodik in einen kleineren Rahmen (eines Bundeslandes) zu testen. Der Vorteil des Themas Corona ist, dass es seit dem Beginn der Pandemie sehr präsent für die Bevölkerung ist und kein Nischenthema darstellt. Auch die Frage der damit einhergehenden Belastungen ist für die Mehrheit der Bevölkerung relevant. Zuerst werden in diesem Kapitel die jeweiligen Twitterdaten und ihre Beschaffung dargestellt. Im Anschluss wird auf die damit verglichenen Umfrage-Daten eingegangen und ihre Entstehung erläutert. Für das Experiment spezifische Vorverarbeitungsschritte werden in einem eigenen Unterkapitel beschrieben. Weiterhin wird die Filterung der Tweets durch Corona-spezifische Keywords beschrieben. Diese gewährleisten einen inhaltlichen Bezug der Twitterdaten zur Umfragestudie. Am Ende wird dann auf die angewandten Verfahren und die dabei entstandenen Ergebnisse eingegangen.

4.1 Daten

4.1.1 Beschreibung der Twitter-Daten

In diesem Unterkapitel sollen die verwendeten Twitterdaten dargestellt werden. Dabei wird auf die Beschaffung der Daten (Twitter-API) und das Datenformat eingegangen.

Der mir bereitgestellte Datensatz wurde folgendermaßen erhoben: Zuerst wurden live Daten mithilfe der Twitter Stream-API mit einer weltweiten Bounding Box erhoben. Dies geschah im Zeitraum 01.12.2019 bis 07.05.2020. Dabei wurden ~426 000 Tweets aus Deutschland erhoben. Anschließend wurden mit einem Research-Account, mit Zugriff auf das vollständige Archiv an Tweets, alle geocodierten Tweets aus Deutschland heruntergeladen.

```
{
 „id": xxxxxxx,
 „text": „Tweet-Text",
 „created_at": „YYYY-MM-DDTHH:MM:SS",
 „lang": „Sprachlabel des Tweets",

 „user": {
  „username": „Auf Twitter angezeigter Name",
  „location": „Freitextbeschreibung des Ortes des Users",
  „id": „User-ID"
 },

 „geo": {
 „country_code": „ISO-Code Staat",
 „county": „Name des Landkreises",
 „state": „Name des Bundeslandes"
 }
}
```

Die Tweets liegen im typischen Twitter-Datenformat vor. Ein Tweet-Objekt ist hierarchisch als JSON-Objekt strukturiert (siehe dazu 4.1.1). Auf der obersten Ebene liegen allgemeine Informationen zu dem Tweet vor, wie die Id(tweet_id), Sprache(lang), Timestamp(created_at), Nachrichtentext(text). Für die Userinformationen liegt ein weitereres Key-Value-Tupel „user" mit User-ID (user_id), Accountname(screen_name), angegebener Ort des Users als Freitext (user_location). Die Geo-Daten befinden sich in der Unterebene „geo", worin die GPS-Koordinaten des verwendeten Geräts gespeichert sind. Die Geo-Tweets wurden

mit dem jeweiligen Landkreis/kreisfreier Stadt der Bounding Box der Koordinaten gelabelt. Dazu wurde die Übereinstimmung der Fläche der Bounding Box des Tweets mit der Bounding Box des Landkreises/kreisfreier Stadt gemessen und bei einer Übereinstimmung mit mindestens 51% des Kreises/kreisfreier Stadt zugeordnet. Dann wurden im Anschluss die Attribute „`county`" für Landkreis/kreisfreie Stadt, sowie „`state`" für das dazugehörige Bundesland in den Key-Value-Tupel für Geo-Daten eingefügt. Für die Analyse ergab dies 86 662 Tweets, deren Geo-Tag dem Bundesland Thüringen zugeordnet werden konnte (vgl. Tabelle 4.1).

Tabelle 4.1: Tweetmengen nach Kategorie

Anzahl der Tweets nach Kategorie	
Kategorie	Tweetanzahl
geocodierte Tweets	86 662
Tweets aus User-Timelines	414 023

Für das Resampling wurden diese Tweets verwendet. Bei den Usern liegt für jeden Tweet ein Geo-Label mit Landkreis/kreisfreier Stadt und Bundesland vor. Die User wurden anschließend dem Landkreis zugeordnet, der unter den Geo-Labels seiner Tweets am häufigsten vertreten war. Für die User, deren Geo-Label (Landkreis/kreisfreie Stadt) innerhalb Thüringens lag, wurden die vollständigen User-Timelines bezogen. Dazu wurden mithilfe des API-Endpoints für die User-Timelines (**GET /2/users/:id/tweets**) die Tweets von den Usern (inklusive Retweets) heruntergeladen. Für die Identifizierung der User, deren Tweets heruntergeladen werden sollen, wird die User-ID verwendet. Bei dem Herunterladen der Tweets existiert eine Beschränkung auf die letzten 3200 Tweets pro User. Diese wurden mit ihren Timestamp und Sprachattribut heruntergeladen. Dies ergab eine Gesamtheit von 414 023 Tweets, die hierbei für ganz Thüringen in dem Zeitraum vom 31.03.2020 bis 24.08.2021 heruntergeladen wurden. Allerdings ergab sich dabei das Problem, dass nicht für alle User, die in den geocodierten Tweets Thüringen zugeordnet wurden, die User-Timelines heruntergeladen werden konnten. Teilweise waren die Accounts bereits gelöscht oder die Sichtbarkeit auf privat gesetzt.

Eine Problematik bei der Verwendung von Twitter-Daten ist die ungleiche Verteilung der Tweetanzahl für die User. In Grafik 4.1 ist dies in einem Histogramm dargestellt. Dieses stellt die Anzahl an Tweets (geocodiert und nicht geocodiert) für User dar, deren Tweets für Experiment 1 heruntergeladen wurden. Diese Menge an Tweets stellte die Ausgangsmenge für die Analyse mit den Resampling-Verfahren, sowie die spätere Keywordfilterung Corona-bezogener Tweets.

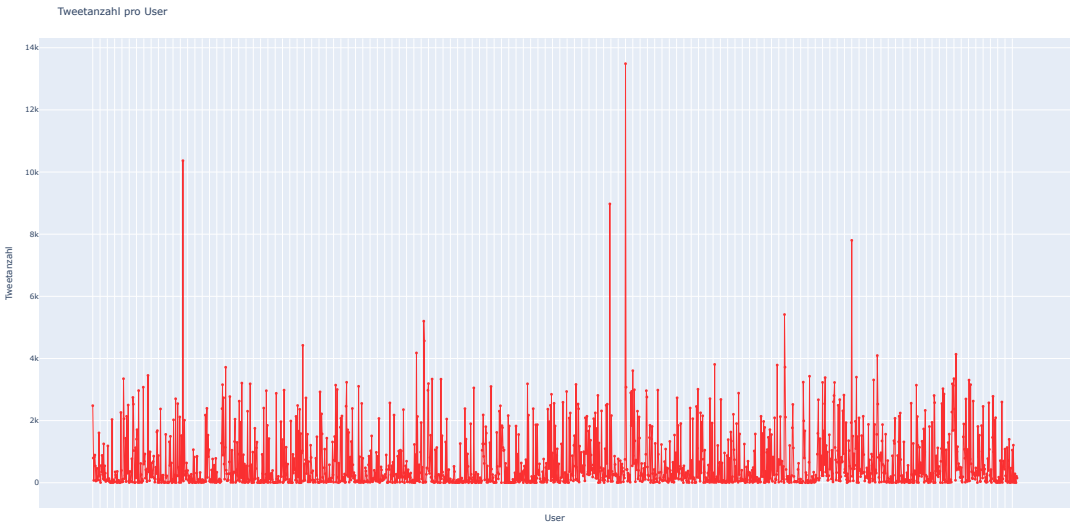

Abbildung 4.1: Verteilung der Tweets nach Usern

4.1.2 Beschreibung der Umfrage-Daten

Nachdem im vorherigen Unterkapitel die Twitterdaten dargestellt wurden, wird in diesem Kapitel auf die verwendeten Umfragedaten eingegangen. Für das Experiment zu Corona wurden Umfragedaten der **Covid 19 Snapshot Monitoring-Studie (Cosmo)** verwendet Betsch et al. [2020]. Diese wird von der Universität Erfurt in Kooperation mit dem Bernhard Nocht-Institut für Tropenmedizin, Robert-Koch-Institut, der Bundeszentrale für gesundheitliche Bildung, dem Leibniz Institut für Psychologie und dem Science Media-Center erstellt. Die Studie beschäftigt sich aus einer psychologischen Perspektive mit dem Phänomen der Wahrnehmung von Corona. Dabei wurden u.a. Risikowahrnehmung, Sorgen/Ängste, psychische Belastungen, Einstellungen zu Impfungen usw. untersucht. Die Daten sind frei verfügbar auf der Webseite herunterzuladen, was bei den meisten Studien aus Datenschutzgründen nicht möglich ist. In diesem Kapitel werden die dabei verwendten Umfragedaten hinsichtlich der Methodik ihrer Entstehung, Datenformat und Inhalten vorgestellt.

Der Cosmo-Studie liegt eine quotierte Stichprobe zugrunde, die offenbar nach Alter (in Gruppen), Geschlecht und regionaler Verteilung quotiert wurde [Betsch et al., 2020]. Die relevante Grundgesamtheit ist die deutschsprachige Bevölkerung zwischen dem Alter 18 und 74 mit einem Internetanschluss. Dabei wurden die Personen in der Auswahl durch einen Dienstleister, der die Umfrage durchführte, kontaktiert und online befragt. Dementsprechend ergibt sich für Menschen, die jünger als 18 und älter als 74 sind, sowie Menschen in medizinischen Einrichtungen, die keinen Internetanschluss haben, ein Un-

dercoverage. Wenn es das Ziel ist die Bevölkerung der BRD abzubilden, wäre diese Auslassung relevant. Jedoch sind in den Vergleichsdaten von Twitter auch keine Personen ohne Internetanschluss enthalten, wobei allerdings die Altersgruppe unter 18 vermutlich stark auf Twitter verbreitet ist. In der Studie Beisch and Schäfer [2020] wurde festgestellt, dass unter den 14-29 Jährigen 8% Twitter nutzen. Auch wenn hierbei der Anteil von unter 18-Jährigen nicht explizit angegeben ist, dürfte dieser Anteil doch recht groß sein. Als Design wurde vorrangig ein Trenddesign verwendet. Parallel wurde auch ein Panel erhoben, jedoch mit allein 2 Erhebungszeitpunkten ist ein Vergleich mit den zeitkontinuierlichen Twitter-Daten nicht sinnvoll.

Es fanden 64 Erhebungen (Wellen) zwischen dem 03.03.2020 und dem 07.06.2022 (Stand: 11.07.2022) statt. Dabei wurden ca. 1000 Personen pro Welle befragt. In dieser Arbeit wurden die Daten zwischen dem 31.03.2020 und dem 24.08.2021 verwendet. In diesem Zeitraum sind aber nicht alle relevanten Variablen immer abgefragt wurden. Meistens liegen die relevanten Daten im zweiwöchentlichen Abstand vor, wobei auch teilweise entweder eine Woche dazwischen liegt oder manchmal auch ein Monat.

Die Daten liegen in tabellarischer Form mit Angabe des Datums des Erhebungsintervalls, des Mittelwerts der Variable, dem dazugehörigen Konfidenzintervall, Anzahl der befragten Personen vor. Dabei sind durch die Daten keine Rückschlüsse auf befragte Individuen möglich. Dementsprechend sind keine Gruppierungen der Daten pro Zeiteinheit notwendig. Allerdings sind somit auch keine weiteren Korrekturen der Stichprobe möglich und die Repräsentativität der Umfrage ist dabei erst einmal als gegeben anzunehmen. Das verwendete Design mit der Ziehung jeweils neuer Stichproben ermöglicht aber keine Rückschlüsse über veränderte Einstellungen der befragten Personen, sondern gibt allgemeine Trends wieder.

Inhaltlich betrachtet wurden Aspekte wie Umgang mit dem Risiko einer eigenen Infektion, Belastung (Situative Belastung, Pandemiemüdigkeit), Sorgen und Ängste (wirtschaftlicher Art, eigene Erkrankung) abgefragt. Für die Analyse in der Arbeit wurden die Variable Situative Belastung verwendet. Hierbei lagen 31 Datenpunkte über einen längeren Zeitraum vor, wobei Daten zwischen den 31.03.2020 und 24.08.2021 verwendet wurden. Zudem war die zeitliche Verteilung der Datenpunkte gleichmäßig. Dies ermöglichte eine Analyse über einen längeren Zeitraum.

4.2 Vorprozessierung der Daten

Zusätzlich zu der in Kapitel 3.1 erläuterten Methodik waren bei den Experimenten weitere Verarbeitungsschritte notwendig. Dies war Folge der spezifi-

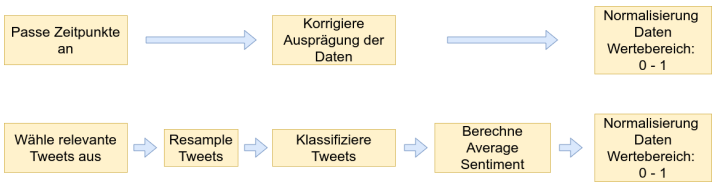

Abbildung 4.2: Verarbeitungsschritte der Twitter- und Umfragedaten in dem Experiment zu Corona

schen Struktur und Format des verwendeten Umfragedatensatzes. Dies betrifft die ersten beiden Punkte des oberen Workflows, sowie die Auswahl der relevanten Tweets.

Workflow der Twitterdaten

Bei der Auswahl der relevanten Tweets (Schritt 1 unten in Grafik 4.2) wurde eine Bereinigung durchgeführt. Dabei wurden User-Mentions und Links entfernt. Zudem wurden nicht-alphanumerische Zeichenketten und Stoppwörter entfernt. Dazu wurde das Python-Package **Spacy** verwendet. Jedoch waren die Berechnungen sehr zeitintensiv, da infolge eines Memory Leaks in der Implementierung von **Spacy** sehr große Mengen Arbeitsspeicher verbraucht wurden. Für die Auswahl der Corona-bezogenen Tweets in der Stichprobe wurde das **Online-Wortschatz-Informations-System Deutsch** verwendet [Leibniz-Institut für deutsche Sprache, 2022]. Die hohe Anzahl der Keywords (2069 Keywords) ermöglicht es Tweets zu erfassen, die allein mit einzelnen Keywords oder Hashtags, wie „Corona“, „Covid-19“ nicht erfasst werden könnten. Dies ermöglicht eine höhere Abdeckung von Tweets. Zudem sollen damit auch umgangssprachliche Redeweisen, die auf Twitter häufig zu finden sind, eingefangen werden. Ein weiterer positiver Aspekt ist, dass die Sammlung von Sprachwissenschaftlern systematisch erstellt wurde, sodass die Wahrscheinlichkeit, dass alle relevanten Begriffe enthalten sind, höher ist als bei ad hoc erstellten Wortlisten. Bei einer Evaluierung eines Samples von 375 Tweets wurden dabei nur 26 falsch-positive Tweets gefunden. Die Mehrzahl der Tweets wurden mit den Keywords korrekt gefiltert (335 Tweets, davon 167 richtig positiv und 168 richtig negativ).

Anschließend wurden die in Kapitel 3.1 benannten Resampling-Verfahren angewandt. Dazu wurde anhand von Eigenschaften der Daten auf Tweetebene, hier der Geo-Tag selektiert. Dies wurde in Verfahren 1 (Tweet-basiert) und Verfahren 2 (Stratifiziert Tweet-basiert) durchgeführt. In Verfahren 2 wird das stratifizierte Sampling mit verschiedenen Werten für die Anzahl an Tweets, die für jede Schicht (hier: Landkreise) gezogen werden, durchgeführt. Dabei wurde ein Parameter n für die Anzahl an Tweets, die pro Landkreis gesampelt wurden,

verwendet. Die verwendeten Werte waren $n \in 250, 500, 1000$. Dieser Wert ist für alle Landkreise gleich (disproportionales Sampling). Die Korrektur erfolgt mit den Designgewichten. Diese wurden mit der Formel 4.1 in Anlehnung an die Umfrageforschung (siehe hierzu Kapitel 2.3) berechnet.

$$\text{Inverse Auswahlwahrscheinlichkeit} = \frac{\text{Größe der Schicht in der Grundgesamtheit}}{\text{Anzahl der Tweets der Schicht}} \quad (4.1)$$

Dagegen wird bei den Verfahren 3 (Populations-basiert) und Verfahren 4 (Populations-geocodiert) anhand von Informationen auf User-Ebene selektiert. Dazu wird der Wohnort der User anhand der Verteilung der Geo-Tags (Landkreis/kreisfreie Stadt) bestimmt. Hierbei wird der häufigste Ort als Wohnort der User (hier: Landkreis/kreisfreie Stadt) gesetzt. Bei einer Kontrolle von 40 Usern zeigte sich eine Anomalie. Eine fehlende Übereinstimmung fand durchgehend bei den Usern (30% der kontrollierten User) statt, die fremdsprachige Tweets gepostet haben. Bei diesen Usern wurde durchgehend in der Userbeschreibung ein anderer Ort oder anderes Land angegeben. Angesichts der Übereinstimmung zwischen dem Sprachlabel der Tweets und dem Land in der Userbeschreibung ist davon auszugehen, dass in der Userbeschreibung das Heimatland der User gemeint ist. Jedoch gehen Tweets dieser User nicht in die Sentimentklassifikation ein, da hierfür allein deutschsprachige Tweets verwendet werden. Somit können diese fehlerhaften Ortsbestimmungen nicht in den Datensatz eingehen. Bei dem Populations-basiertem Verfahren wurden zusätzlich nicht-geocodierte Tweets der User verwendet, wogegen bei dem Populationsgeocodierten Verfahren allein die geocodierten Tweets verwendet werden.

Tabelle 4.2: Evaluation des Ortsbestimmungsverfahren der User

Übereinstimmung User-Location mit automatischer Geo-Labelung	
Grad der Übereinstimmung	Anzahl Tweets
exakt	18
fast exakt	3
negativ (fremdsprachige Tweets)	12
negativ (deutschsprachige Tweets)	1
unklar	6
insgesamt	40

In Bezug auf die Verarbeitung wurden nur deutschsprachige Tweets verwendet. Dazu wurden nur Tweets mit dem Attribut „`lang`", die den Wert „`de`" hatten verwendet. Da die Keywords zur Filterung auf einem Wörterbuch der deutschen Sprache aufbauen und das verwendete Sentimentmodell mit deutschsprachigen Tweets trainiert wurde, kann nicht sinnvoll mit nichtdeutschsprachigen Daten gearbeitet werden. In Tabelle 4.3 werden beispielhaft

Fehlklassifikationen des Modells dargestellt. Die Namen oder Usernamen in den Tweets wurden dabei aus Datenschutzgründen entfernt. Bei einer Sichtung eines Samples von 100 Tweets wurden 70 Tweets korrekt klassifiziert und 30 falsch klassifiziert. Dabei waren Muster in den Fehlklassifikationen erkennbar. Es wurden Tweets, die die Zeichenkette „Guten Morgen“ enthalten, entweder neutral oder negativ gelabelt, aber nie positiv. Weiterhin wurden Tweets mit „Hat gerade ein Foto gepostet...”durchgehend negativ gelabelt.

Tabelle 4.3: Fehlklassifikationen des Sentiments

Tweets mit vergebenen Sentimentlabel	
Tweet	Sentimentlabel
„Guten Morgen TL, wünsche euch einen schönen Tag, guten Start in die neue Woche und gutgelaunte Mitmenschen”	neutral
„Hat gerade ein Foto gepostet...”	negativ
„Liebe Frau ___ du bist ein Sonnenschein und eine wundervolle und liebenswerte Frau”	neutral

Workflow der Umfragedaten

Bei der Verarbeitung der Cosmo-Daten zeigte sich, dass eine Korrektur der Daten notwendig ist. In den meisten Fällen liegt das Datum der Erhebungszeitpunkte 14 Tage auseinander, aber auch teilweise 7 Tage oder nur einmal pro Monat. Das lässt ggf. vermuten, dass teilweise bei den Daten im Intervall eventuell zuwenig Antworten der Befragten erhalten wurden oder dass die Auftraggeber der Studie sich entschieden zu dem Zeitpunkt keine Umfragen durchzuführen. Tendenziell lagen die meisten Datenpunkte aber im 14-Tagesintervall vor. Bei den Twitter-Daten wurde für den Vergleich die Daten für ein Intervall, dass mit einem Cosmo-Erhebungszeitpunkt endet, zusammengefügt (Schritt 1 oben in Grafik 4.2). Dabei wurde ein Mittelwert für das jeweilige Intervall gebildet. Die Berechnung erfolgte anhand der Formel 3.1.

Die Motivation dahinter ist zum einen das Rauschen in den Daten durch tagesabhängige Fluktuationen zu minimieren. Dies ermöglicht auch eine bessere Exploration der Graphen in den Plots, da die Bewegungen der Graphen eindeutiger sind. Zum anderen ist es wie oben beschrieben notwendig die Vergleichszeiträume zwischen den beiden Zeitreihen anzugleichen, um nicht z.B. tagesaktuelle Daten mit Daten im 2-Wochenrhythmus zu vergleichen. Zudem wurde die Ausprägung der Werte mit den Sentiment-Daten harmonisiert. Da bei den Sentiment-Daten ein höherer Wert ein positiveres Sentiment auf Twitter darstellt, wurden die Ausprägungen der Werte aus der Umfrage bei der

Variable `situative Belastung` umgekehrt. Damit bedeutet ein höherer Wert eine geringere Belastung. Dementsprechend ist die inhaltliche Bedeutung beider Graphen gleichgerichtet.

Weiterhin wurde der Wertebereich der Variablen normalisiert, um die Vergleichbarkeit von den Werten aus der Umfrage mit den Sentiment-Daten zu ermöglichen (vgl. dazu Formel 3.2).

4.3 Ergebnisse

In diesem Kapitel wird zum einen auf die Evaluierung der Resampling-Ansätze mit den verwendeten Metriken (Mean-Squared Error, Pearson r, Granger-Kausalität) eingegangen. Dabei werden die Ergebnisse dargestellt und es wird auf die vorhandene Datenbasis eingegangen.

Tabelle 4.4: Übersicht über Anzahl der verwendeten Tweets in den Resampling-Verfahren.

Anzahl der Tweets		
	alle Tweets	Tweets mit Keywords
Tweet-basiert	86 626	2329
Stratifiziert Tweet-basiert	5 500	-
Populations-basiert	501 503	25 335
Populations-geocodiert	87 483	2313

In Bezug auf die Metrik Mean-Squared-Error zeigen sich einige Unterschiede in den Daten. Die höchste Abweichung der beiden Zeitreihen (MSE) ist bei Verfahren 3 (Populations-basiert) mit einem Wert von 0,186 zu finden, wobei der niedrigste Abstand auftritt, wenn bei diesem Verfahren nur Keyword-gefilterte Tweets zu Corona zur Messung verwendet werden (siehe Tabelle 4.5). In diesem Fall beträgt der MSE 0,07. Generell ist das Spektrum der Werte des MSE für die verschiedenen Verfahren gering, sodass hier bei der Interpretation Vorsicht geboten ist. Bei der Metrik Pearson-Korrelation schwanken die Ergebnisse stärker. Die schwächste Korrelation findet sich bei dem Verfahren, dass auch den höchsten MSE hat. Dies spricht dafür, dass die Metriken ähnlich gut die Effektivität der Verfahren messen. Für Verfahren 3 mit allen Tweets beträgt die Korrelation -0,113, was sogar auf einen gegenläufigen Zusammenhang hindeutet. Die höchste Korrelation wurde bei Verfahren 4 (Population-geocodiert) mit allen Tweets festgestellt. Hier beträgt der Wert 0,362, was auf einen schwachen Zusammenhang hindeutet. In der Reihenfolge der Korrelationen folgt dann die Stratifikation (Verfahren 2, n=250, ohne Oversampling) mit

einer Korrelation von 0,262. Generell zeigt sich, dass das verwendete Baseline-Verfahren 1, mit den tweetbasierten Geo-Daten, nicht besonders gut performt. Die Korrelation ist hier nah bei 0 und mit einem negativen Vorzeichen (Wert von -0,016). Wie bereits in Kapitel 2.1 erwähnt stellt dieses Verfahren die Grundlage für viele Studien, die mit einer Lokalisierung von Nutzern arbeiten (vgl. Kruspe et al. [2020]. Diese Ergebnisse deuten daraufhin, dass solche Studien ihre Messung von Sentimententwicklungen verfeinern können, sofern sie die räumliche Verteilung der User genauer erfassen. Dies kann durch die Bestimmung des Wohnortes (Verfahren 3 und 4) oder Stratifikation teilweise verbessert werden. Jedoch scheint die Auswirkung der Verfahren dabei auch nicht allzu groß zu sein, da die Unterschiede im MSE recht klein sind und die Korrelationen für Verfahren 3 sich, je nachdem ob mit Keywords gefiltert wurden, doch sehr unterscheiden (siehe hierzu Zeile 3 in Tabelle 4.5).

Tabelle 4.5: Ergebnisse der verwendeten Verfahren (MSE, Pearson-Korrelation)

Ergebnisse der Resampling-Verfahren				
	alle Tweets MSE	alle Tweets r	Keywordtweets MSE	Keywordtweets r
Tweet-basiert	0.105	-0.016	0.106	0.076
Stratifiziert Tweet-basiert	0.069	0.262	-	-
Populations-basiert	0.186	-0.113	0.070	0.171
Populations-geocodiert	0.074	0.362	0.099	0.116

Bei der stratifizierten Auswahl veränderten sich die Ergebnisse je nach verwendeter Tweetzahl für die Landkreise. Es wurden verschiedene Werte [100, 250, 500, 1000] für die Samplegröße n pro Landkreis verwendet. Bei einer Tweetzahl $n > 500$ muss mit einem Oversampling der Tweets aus Landkreisen mit weniger Tweets gearbeitet werden, da hier nicht genügend Tweets vorhanden sind. Dies bedeutet, dass bei einem disproportionalen Sampling mit Gewichten für gewisse Landkreise die Tweetzahl sich eher einer Vollerhebung annähert, wogegen für tweetreiche Landkreise eine (kleinere) Stichprobe erhoben wird. Für die Darstellung der Metrik wurde der Wert ausgewählt, der bei den Metriken MSE und Pearson-Korrelation die höchsten Werte erreicht. Dies war der Fall bei 250 Tweets pro Landkreise ohne Oversampling. Dabei betrug der MSE 0,069 und die Korrelation 0,262 (siehe Tabelle 4.5). Ähnliche Ergebnisse wurden auch mit dem Wert 500 für die Tweets pro Landkreise erreicht. Auffällig ist, dass bei einer höheren Anzahl an Tweets (1000 Tweets pro Landkreis) keine Verbesserung erreicht wird, sondern die Korrelation recht stark fällt. Sie beträgt hierbei nur noch 0,147. Eine höhere Menge von Daten scheint hierbei keine bessere Anpassung der Tweet-Zeitreihen an die Umfragedaten zu gewährleisten. Bei den Keyword-gefilterten Tweets ließ sich die Stratifikation nicht anwenden, da bei dieser Tweetmenge nicht genügend Tweets für

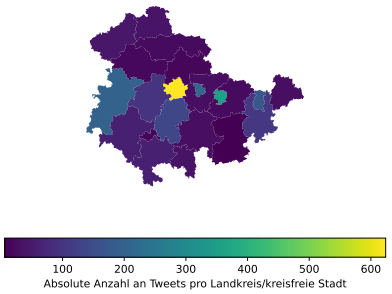

Abbildung 4.3: Anzahl der Tweets pro Landkreis

einige Landkreise verfügbar sind. Da die räumliche Verteilung sehr ungleichmäßig ist, ist es schwierig ein regional balanciertes Sampling durchzuführen. Bei 8 Landkreisen konnten in diesem Datensatz nur weniger als 30 Tweets mit Corona-Keywords gefunden werden. Bei diesen Orten ist statistisch gesehen pro Datenpunkt weniger als 1 Tweet vorhanden ist (siehe hierzu Abbildung 4.3).

In der Grafik 4.4 sind die 20 häufigsten Wörter aus dem Datensatz Corona-bezogener Tweets dargestellt. Wie zu erwarten sind Begriffe, wie „Corona“, „Neuinfektionen“ und „Infizierte“ häufig in dem Datensatz vertreten, der mit dem Corona-bezogenen Wörterbuch gefiltert wurde. In dem Datensatz finden sich viele Retweets von Mitteilungen des Thüringer Gesundheitsministeriums. Dabei wurde in hoher Frequenz die aktuellen Corona-Fallzahlen des Landes Thüringen gepostet. Dadurch ergaben sich bereits 130 Tweets, die das Thema Corona behandelten.

Abbildung 4.4: Top 20 der häufigsten Begriffe

Kapitel 5

Experiment 2: Prognostizierung von Wahlumfragen

In dem zweiten Experiment der vorliegenden Arbeit soll die dargestellte Methodik auf den Bereich Wahlen und Wahlprognosen angewandt werden. Ähnlich, wie in dem ersten Experiment werden Zeitreihendaten von Twitter mit Daten aus Umfragen verglichen. Hierbei werden Umfragedaten zu Wahlabsichten bei der Bundestagswahl 2021 verwendet. Diese war ein Großereignis, wobei aufgrund der Wichtigkeit der Wahl angenommen werden kann, dass dazu eine ausreichende Datengrundlage auf Twitter vorhanden ist. Bei regionalen oder kommunalen Wahlen stellt sich dies wohl schwieriger dar, da der Kreis der betroffenen Personen deutlich kleiner ist. Ein weiterer Vorteil des Themas ist, dass die Daten bereits quantitativ vorliegen und als Prozentwerte einen standardisierten Wertebereich haben. In dem Kapitel werden erst die verwendeten Twitter-Daten und in Folge die verwendeten Umfragedaten beschrieben. Zudem wird auf die Filterung von Tweets durch parteispezifische Keywords eingegangen. Der Experimentalablauf mit den verwendeten Vorverarbeitungsschritten wird dargestellt und abschließend werden die Ergebnisse der verwendeten Resamplingverfahren erläutert.

5.1 Daten

5.1.1 Beschreibung der Twitter-Daten

Vor der durchgeführten Analyse wurde mir ein Datensatz geocodierter Tweets freundlicher bereit gestellt. Dieser enthält geo-getaggte Tweets die in Deutschland gepostet wurden. Die Methodik der Erstellung ist die selbe, wie sie in 4.1.1 beschrieben wurde. Hierbei liegen Daten für den Zeitraum von 02.08.2021 bis 25.09.2021 vor. Der Zeitraum der verwendeten Twitter-Daten ergibt sich aus

Tabelle 5.1: Verwendete Tweetmengen nach Kategorie

Anzahl der Tweets nach Kategorie	
Tweetkategorie	Tweetanzahl
geocodierte Tweets	1 164 897
Tweets aus User-Timelines	2 061 080

dem verwendeten Umfragedatensatz. Der Zeitraum der verwendeten Twitter-Daten ist selber frei wählbar, allerdings gibt es nicht zu jedem Zeitpunkt Umfragedaten zur Wahlentscheidung. Dementsprechend musste sich hier an den Erhebungszeitpunkten der verwendeten Umfrage orientiert werden, damit ein zeitliches Matching möglich ist. Das Format der Tweets entspricht dem Format, wie es in Grafik 4.1.1 beschrieben wurde.

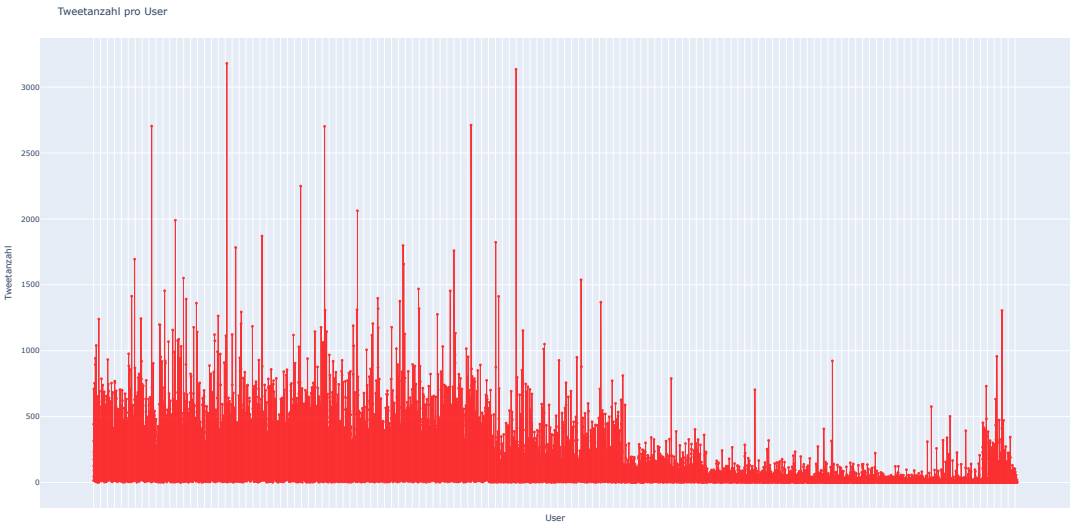

Abbildung 5.1: Verteilung der Tweets nach Usern

Eine Problematik bei der Verwendung von Twitter-Daten ist die ungleiche Verteilung der Tweetanzahl für die User. In Grafik 5.1 ist dies in einem Histogramm dargestellt. Dieses stellt die Anzahl an Tweets (geo-codiert und nicht geocodiert) für User dar, deren Tweets für Experiment 2 heruntergeladen wurden.

Es stellte sich später heraus, dass die Zuordnung der User zu ihrem Wohnort in manchen Fällen aufgrund der ungünstig gelegenen Grenzen von kleineren Bundesländern korrigiert werden musste. Die Bounding Box dieser Bundesländer umfasste weite Gebiete von angrenzenden Bundesländern. Dies betraf die Bundesländer Bremen, Hamburg, Saarland. Im Fall von Bremen nahm Niedersachsen einen sehr großen Anteil der Bounding Box ein. Da sie ohne ein

Geolabel nicht in die Verfahren eingehen können, wären somit Daten verloren. Deshalb wurden die Daten vorab mit dem Bundesland und nicht dem Landkreis gelabelt, anders als in dem ersten Experiment.

Dementsprechend wurde der Ort der User anhand der Angaben zu dem Bundesland der Tweets bestimmt. Für die User wurde der häufigste Wert des Geo-Tags des Bundeslandes verwendet, um auf das Bundesland des Wohnortes des Users zu schließen. Dazu wurden anschließend für diese User die User-Timelines heruntergeladen. Für die Beschreibung siehe Kapitel 4.1.1. Dazu wurden insgesamt 2 061 080 Tweets heruntergeladen.

5.1.2 Beschreibung der Umfrage-Daten

Inhalt dieses Kapitels ist die Darstellung der verwendeten Umfragedaten, die mit den heruntergeladenen Twitter-Daten verglichen werden. Dazu gehe ich auf die Erhebung, Design und das Format des Umfrage-Datensatzes ein.

Die **German Longitudinal Election Study (GLES)** wird seit 2009 regelmäßig im Umfeld von Wahlen durchgeführt. Beteiligte Institute sind GESIS – Leibniz Institut für Sozialwissenschaften und die Gesellschaft für Wahlforschung.

Im Rahmen der GLES werden verschiedene Teilstudien mit unterschiedlichen Designs durchgeführt. Dies sind zu einem der GLES-Querschnitt, die GLES-Rolling Cross Section, das GLES-Panel, die GLES-Kandidatenstudie und das GLES-Langfrist-Online-Tracking. Diese unterscheiden sich zudem auch nach der Anzahl und Abstand der Erhebungszeitpunkte. Teilweise sind auch die Befragungen langfristiger Natur und haben einen großen Abstand zwischen den Datenpunkten. Da die Twitter-Daten auf einer täglichen Basis vorliegen, ist ein ähnlicher Rhythmus der Erhebungen sinnvoll bei einem Vergleich (z.B. täglich oder wöchentlich). Dies ist bei der GLES-Rolling Cross Section gegeben, die tägliche Umfragedaten vom Zeitpunkt vor der Bundestagswahl bereitstellt. Ein weiterer Vorteil ist, dass diese Daten frei verfügbar sind, ohne eine Nutzungsvereinbarung zur datenschutzkonformen Verwendung der Daten zu unterzeichnen.

Methodisch ähnelt die Studie einem Trenddesign, da es mehrere Erhebungszeitpunkte mit verschiedenen Personen gibt. Jedoch wird beim Rolling Cross Section-Design die Stichprobe vorab als Querschnitt geplant und dann in verschiebene Scheiben aufgeteilt. Jede dieser Scheiben wird zu der geplanten Zeiteinheit (hier: ein Tag) erhoben und fungiert für diesen Tag als Zufallstichprobe. Das Design ermöglicht damit Veränderungen bei den Einstellungen der Wähler zeitlich genau zu erfassen. Die Befragung selber wurde telefonisch durchgeführt. Dazu wurden insgesamt 7000 Befragte interviewt, mit je 100 – 200 Befragten pro Tag. Ein Teil wurde als Vorwahlbefragung mit täglichen Befra-

gungen vom 02.02.2011 bis zum 25.09.2011 durchgeführt. Anschließend wurde eine Nachwahlbefragung als Panel durchgeführt GESIS [2022a]. In dieser Arbeit wurde die Analyse für das Vorfeld der Wahlen durchgeführt, da dazu das höchste Tweetaufkommen auf Twitter zu erwarten war.

Die Grundgesamtheit für die Studie sind die Wahlberechtigten zur Bundestagswahl 2021, d.h. allein deutsche Staatsbürger GESIS [2022a]. Damit liegt für unter 18-Jährige, sowie nicht-deutsche Staatsbürger, die in der BRD leben, ein Undercoverage vor. Zu beachten ist, dass diese Gruppen nicht von dem Twitter-Sample ausgeschlossen sind und auch vorab keine Daten über diese User vorliegen. Bei der telefonischen Erhebung wurden sowohl Personen mit Festnetzanschluss als auch mit Mobiltelefonen befragt. Dementsprechend liegt auch für Personen ohne Telefon jedweder Art (z.B. Personen in Pflegeheimen, Gefängnissen und andere Anstalten) ein Undercoverage vor.

Inhaltlicher Fokus der Studie sind Wahleinstellungen der Befragten, Wahrnehmung des Wahlkampfs. Insgesamt wurden dabei u.a. 355 Variablen inhaltlicher Art erhoben, sowie personenspezifische Daten (z.B. Bundesland, Geschlecht), Daten zum Interview (z.B. Dauer des Interviews, Anteil beantworteter Fragen) GESIS [2022a].

Für die durchgeführte Analyse sind folgende Variablen relevant: `Bundesland des Befragten`, `Wahlabsicht Zweitstimme Bundestagswahl 2021` (pre006ba) , `Timestamp` (timestamp_pre031). Für die Abschätzung der Wähleranteile wurde die Variable `Zweitstimme Bundestagswahl 2021`: „Welche Partei werden sie bei der Bundestagswahl wählen?“ verwendet. Zu beachten ist, dass aus Datenschutzgründen Variablen, wie z.B. Geschlecht, Postleitzahl, sexuelle Orientierung entfernt wurden [GESIS, 2022b, S.21-22].

5.2 Vorprozessierung der Daten

Nach der Vorstellung der verwendeten Daten von Twitter bzw. der Umfrage (GLES) in den vorherigen Kapiteln wird hier auf die benötigten Verarbeitungsschritte eingegangen, um die gegebenen (Roh)Daten für die Analyse der Wahldynamiken auf Twitter anzuwenden. Das Ziel ist es Zeitreihen von Twitterdaten zu schaffen, die mit den Zeitreihen aus Umfragedaten, verglichen werden können. Dazu werden die relevanten Tweets gefiltert, die Resampling-Verfahren angewendet, das Sentiment vermessen und die entstandenen Aggregate als Zeitreihe gespeichert.

Diese Verarbeitung erfolgt in 2 Strängen. Der Strang der Umfragedaten selber benötigt weniger Umformungsschritte als die Bearbeitung der Twitter-Daten 5.2.

Workflow der Twitter-Daten

In dem ersten Schritt wurde aus dem bestehenden Twitter-Korpus die relevanten Tweets mit Keywords gefiltert. Grundlage dazu war eine Liste von Keywords/Hashtags mit den Namen der Parteien, sowie den Namen der Spitzenkandidat(innen) der Parteien. Dies wurde für die 6 aktuell im Bundestag vertretenen Parteien durchgeführt (CDU, SPD, FDP, Grüne, Linke, AfD). Die Liste der verwendeten Keywords befindet sich in Anhang A.2.

Da für Deutschland ein Verhältniswahlrecht für die Bundestagswahl besteht und die Zweitstimme für die Parteien am wichtigsten ist, müssen die extrahierten Twitter-Daten nach Partei getrennt werden. Dies ermöglicht erst den Vergleich zwischen den Umfragedaten und den Twitter-Daten. Dabei ergab sich das Problem, dass sich nicht in jedem Fall ein Tweet eindeutig einer bestimmten Partei zuordnen ließ, da es Mehrfachnennungen von Parteien in den Tweets gab. Diese Tweets wurden dabei aus dem verwendete Korpus entfernt, da sie zur Messung der Wahrnehmungen der Parteien auf Twitter nicht eindeutig waren.

Die für diese Analyse verwendeten Tweets wurden mit den in Kapitel 3 beschriebenen Resampling-Verfahren, Verfahren 1 (Tweet-basiert), Verfahren 2 (Stratifiziert Tweet-basiert), Verfahren 3 (Populations-basiert) und Verfahren 4 (Populations-geocodiert) ausgewählt. Bei dem Tweet-basierten Verfahren wurden alle (relevanten) geocodierten Tweets für das Resampling verwendet. Die Stratifiziert Tweet-basierte Auswahl arbeitet mit den geocodierten Tweets, die auch in dem Tweet-basierten Verfahren verwendet werden. Hierbei wird gemäß dem Parameter n eine Anzahl von Tweets für jedes Bundesland ausgewählt. Die Werte betrugen hierbei $n \in 100, 250, 500$. Diese Tweets werden gesampelt und mit Gewichten versehen, um eine balanciertere räumliche Verteilung zu ermöglichen. Für die Verfahren Populations-basiert und Populations-geocodiert wurden die Tweets der User verwendet, die hier eindeutig einem Bundesland zugeordnet werden konnten (siehe Kapitel 5.1.1). Dabei wurden für das erste

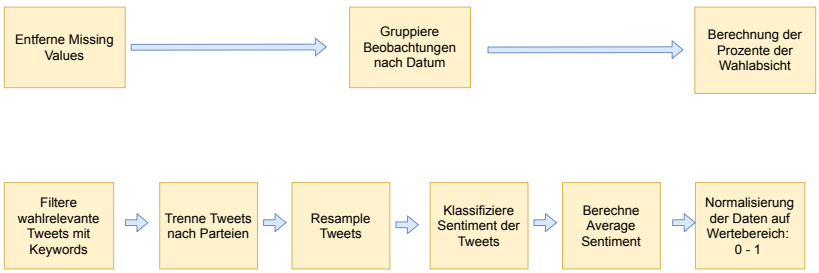

Abbildung 5.2: Verarbeitungsschritte der Twitter- und Umfragedaten in dem Experiment zur Prognostizierung von Wahlumfragen

Verfahren vollständige User-Timelines auch mit nicht-geocodierten Tweets verwendet, wogegen für letztes Verfahren allein die geocodierten Tweets verwendet wurden. Damit konnten für die verschiedenen Verfahren getrennte Datensätze geschaffen werden, um diese mithilfe der Umfragedaten zu evaluieren.

Tabelle 5.2: Beispiele für Fehlklassifikationen (Sentiment)

Tweetbeispiele	
Tweet	Sentimentlabel
"@_ Die CDU ist so #Todeslostinkompetentseineoma"	Neutral
"@ArminLaschet du Umwelts** https://t.co/2sCN9yjCLu"	Neutral
"#AnnalenaBaerbock rasiert die Dudes. #triell"	Neutral
"RT marcobuelow: #laschet abschieben nach Afghanistan...Oder verwechsle ich da was?"	Neutral

Die parteibezogenen Tweets wurden dann mit dem verwendeten Sentimentmodel (**Germansentiment**) als ‚positiv', ‚neutral' oder ‚negativ' gelabelt, um die Einstellungen der User gegenüber den Parteien anhand einzelner Tweets zu erfassen. Bei einer Sichtung von 100 zufällig ausgewählten Tweets zeigte sich, dass die Genauigkeit der Messung im mittleren Bereich lag. Dabei wurden 60 Tweets korrekt gelabelt und 40 falsch gelabelt. Tweets, die sehr umgangssprachlich oder in Jugendsprache formuliert wurden, wurden häufig falsch gelabelt. Dabei lässt sich eine Tendenz zur neutralen Bewertung durch das Modell erkennen. Beispieltweets mit durch das Modell vergebenen Sentimentlabel siehe Tabelle 5.2.

In einem weiteren Schritt wurden diese Sentiment-Bewertungen der Tweets für einen Zeitabschnitt aggregiert, um ein allgemeines Stimmungsbild gegenüber der jeweiligen Partei an einem Tag oder Woche zu berechnen. Dabei wird der Anteil der negativ gelabelten Tweets mit dem Anteil der positiv gelabelten Tweets verrechnet, siehe dazu Formel 3.1.

Zum Abschluss wurde der Wertebereich der daraus entstandenen Zeitreihe von Werten auf den Bereich 0 bis 1 normalisiert, um das Messspektrum der parteibezogenen Sentimentreihen, an das Messpektrum der Wahlanteile in den Umfragedaten anzugleichen, welches selber aus Prozenten besteht. Als End-Resultat liegt dann eine Liste von Beobachtungszeitpunkten mit den aus den Tweets aggregierten Sentiment-Scores vor. Diese sind durch die Anwendung der in diesem Kapitel beschriebenen Schritt dementsprechend vorprozessiert, sodass eine Vergleichbarkeit mit den Umfragedaten gegeben ist.

Workflow der Umfragedaten

Nach dem Einlesen des Umfragedatensatzes als Stata-Datei mithilfe der Python-Bibliothek **Pandas** wurden zuerst Einträge mit Missing Values entfernt. Bei einigen Befragungen liegen bei Variablen, die für die Analyse benötigt werden,

keine Werte vor. Dies betraf den Timestamp, sowie die abgefragte Wahlabsicht der Zweitstimme. Diese Beobachtungen wurden dann entfernt (Schritt 1 in der Grafik 5.2).

Die Daten in dem verwendeten File der GLES-Wahlstudie sind erst einmal tabellarisch geordnet, aber weisen noch keine zeitliche Struktur auf. Um eine geordnete Zeitreihe zu erstellen (Schritt 2), wurden die Einträge (Zeilen) mit den Interviews nach dem Timestamp als Datum gruppiert, um für jeden Tag die erhobenen Daten zusammenzufassen.

In Schritt 3 wurde anschließend eine Transformation der Datenausprägung für die Wahlabsicht durchgeführt. Da die Wahlabsicht als kategoriale Variable gegeben ist, kann vorab mit dieser ohne eine Bearbeitung nicht gerechnet werden. Dazu wurden pro Tag pro Partei die angegebenen Wahlabsichten gezählt und auf die Gesamtheit der Angaben für diesen Tag bezogen. Dabei wurden neben den 6 in der Analyse verwendeten Parteien auch die Ausprägung „Sonstige“ einbezogen. Damit entsteht eine Zeitreihe mit den Daten der Erhebung und den Prozenten der geäußerten Wahlabsicht für jede Partei. Aufgrund des Wertespektrums von Prozenten können diese dann direkt in der Analyse mit normalisierten Sentimentwerten verglichen werden.

Die dabei entstandenen Zeitreihendaten können auf verschiedene Weisen verglichen werden. Für explorative Zwecke wurden Plots der Zeitreihen visuell verglichen. Der Fokus der Arbeit liegt aber stattdessen auf der formalen Evaluierung des Verhältnisses der Zeitreihen. Dazu wurden die Metriken Mean-Squared Error, Pearson-Korrelation und Granger-Kausalität genutzt. Die Beschreibung der Metriken erfolgte in Kapitel 3.3.

5.3 Ergebnisse

Zuerst wird in diesem Kapitel die quantitative Datengrundlage mit der Anzahl an Tweets, die verwendet wurden, beschrieben. Dies wird dann weiter nach verwendeten Verfahren, sowie Partei aufgeschlüsselt. Anschließend werden die Ergebnisse der verwendeten Metriken (MSE, Pearson r, Granger-Kausalität) beschrieben. Da sich der Ablauf dieses Experiments von dem ersten Experiment unterscheidet, wird zudem auf die Unterschiede in der Datenbasis, als auch der Berechnung der Metriken eingegangen.

Für die Auswahl der relevanten Tweets mussten in dem zweiten Experiment zusätzliche Schritte durchgeführt werden, die in dem ersten Experiment nicht nötig waren. Deshalb muss bei der quantitativen Beschreibung der Datengrundlage zwischen der Gesamtanzahl an Tweets, die primär zur Bundestagswahl gepostet wurden, inklusive Tweets mit Mehrfachnennungen von

Parteien, und der Anzahl an Tweets, die mithilfe der Verarbeitungsschritte in Kapitel 5.2 den einzelnen Parteien eindeutig zugeordnet wurden. Ersteres wird dabei folgend als Bruttoanzahl von Tweets und letzteres als Nettoanzahl von Tweets bezeichnet. Für die Einordnung der verwendeten Datenmengen in den Verfahren sind allein die Anzahl der Tweets in der Nettogröße relevant.

Tabelle 5.3: Bei den Verfahren verwendete Tweetmengen

Datenbasis der Resampling-Verfahren	
Verfahren	Tweetanzahl
Tweet-basiert	41 488
Stratifiziert Tweet-basiert	8000 (n= 500 pro Bundesland)
Populations-basiert	90 417
Populations-geocodiert	41 089

Bei den verwendeten Verfahren unterscheiden sich die Tweetzahlen (Tabelle 5.3) weniger stark als bei der Fallstudie zu den Belastungen bei Corona (vgl. Tabelle 4.4). Alle Tweetzahlen beziehen sich auf Tweets, die inhaltlich im Zusammenhang mit der Bundestagswahl stehen und in den Resampling-Verfahren verwendet wurden. Durch die User-Timelines ergaben sich mehr Tweets, jedoch waren davon nur ein 5,4 % zu dem Thema Bundestagswahl. Bei der Stratifikation (Verfahren 2) wird von der Menge geocodierter Tweets zur Bundestagswahl ausgegangen.

Tabelle 5.4: Netto-Anzahl an Tweets pro Partei

Tweetanzahl pro Partei (Verfahren:Tweet-basiert)	
Partei	Tweetanzahl
SPD	7 325
CDU	18 066
Grüne	5 137
FDP	3 553
Linke	3 078
AfD	4 329
insgesamt	41 488

Unter den geocodierten Tweets ist der hohe Anteil an Tweets auffällig, die CDU-bezogen sind (Tabelle 5.4). Diese Partei ist mit 18 066 mehr als doppelt so häufig vertreten, wie die Partei mit der zweithöchsten Anzahl an Tweets (SPD). Die Unterschiede bei den kleinen Parteien (Grüne, FDP, Linke, AfD) sind weniger stark ausgeprägt, mit einer Anzahl von 3078 Tweets, die inhaltlich der Linken zugeordnet sind. Dieser Wert stellt den geringsten Wert bei der Tweetanzahl bezüglich der kleinen Parteien dar. Unter den kleinen Parteien haben die Grünen, die meisten zuordenbaren Tweets. Die Anzahl dieser Tweets beträgt 5 137.

Tabelle 5.5: Netto-Anzahl an Tweets pro Partei

Tweetanzahl pro Partei (Verfahren: Populations-basiert)	
Partei	Nettotweetanzahl
SPD	15 365
CDU	39 863
Grüne	13 077
FDP	7194
Linke	7151
AfD	7767
insgesamt	90 417

Trotz der deutlich größeren Datenbasis mit den User-Timelines mit sowohl geo- und nicht geocodierten Tweets ist erkennbar, dass davon nur ein Bruchteil in Zusammenhang mit der Bundestagswahl steht. Es wurden aus den User-Timelines 2 061 080 Tweets heruntergeladen, wovon 111 783 Tweets im Zusammenhang mit den Parteien zur Bundestagswahl stehen. Dies ergibt 5,4% der gesamten User-Timeline-Tweets. In Bezug auf die Gesamtheit der Tweets zu den Parteien bestehend aus geocodiert und nicht geocodierten Tweets ergibt sich eine ähnliche Verteilung, wie bei den rein geocodierten Tweets. Die Proportionen der Tweetzahlen zwischen den Parteien entsprechen von der Reihenfolge den Tweetzahlen der Parteien bei den rein geocodierten Tweets. Mit den User-Timelines hat sich die Gesamtanzahl parteibezogener Tweets von 41 000 auf 90 000 mehr als verdoppelt. Für SPD, CDU, Grüne, FDP und Linke übersteigt der Zugewinn an Tweets die 100%, wogegen die AfD weniger als 100% der Anzahl der geocodierten Tweets zugewinnt.

Tabelle 5.6: Evaluation der Verfahren

Ergebnisse der Verfahren		
	Durchschnittlicher MSE	Durchschnittliche Korrelation
Tweet-basiert	0.105	0.048
Stratifiziert Tweet-basiert	0.105	0.048
Populations-basiert	0.106	-0.019
Populations-geocodiert	0.105	0.038

Die Metriken MSE, Pearson-Korrelation und Granger-Kausalität wurden in diesem Experiment jeweils für die Übereinstimmung zwischen Sentiment- und Umfragedaten für jede Partei einzeln berechnet. Die Fragestellung hierbei ist aber die „Erklärungskraft“ der einzelnen Resampling-Verfahren. Zudem wäre eine Darstellung der Ergebnisse für alle Parteien für alle Verfahren zu unübersichtlich, da bei 4 Verfahren für 6 Parteien 24 Plots bzw. Ergebnisse vorliegen würden. Deshalb wurde der Durchschnitt der Ergebnisse der Parteien für das gewählte Verfahren berechnet. Auffällig ist bei der Metrik MSE die geringe Bandbreite der Ergebnisse in Tabelle 5.6. Bei dem durchschnittlichen

MSE der Verfahren (für alle Parteien kombiniert) sind nur geringe Unterschiede festzustellen. Die Werte der Parteien variieren jeweils für die Verfahren und sind in (geringem Maße) unterschiedlich. Auch bei den Korrelationen ergeben sich nur kleinere Unterschiede in der zweiten Nachkommastelle. Die Ausnahme ist hierbei das Populations-basierte Verfahren, wobei das Vorzeichen negativ ist. Jedoch ist auch hier mit einem r von -0,019 kein statistischer Einfluss erkennbar. Damit lässt sich für kein Verfahren ein statistischer Zusammenhang zwischen den Werten der Sentimentdaten und der Umfragedaten feststellen.

Tabelle 5.7: Ergebnisse für einzelne Parteien (Tweet-basiert)

Ergebnisse pro Partei (Verfahren: einfach geocodiert		
Partei	MSE	Korrelation
SPD	0.106	0.201
CDU	0.088	-0.003
Grüne	0.143	-0.077
FDP	0.106	-0.191
Linke	0.118	0.193
AfD	0.066	0.164

Um die Ergebnisse der Verfahren an einem Beispiel genauer zu verdeutlichen, werden in Tabelle 5.7 die Ergebnisse der 6 Parteien für das Tweet-basierte Verfahren dargestellt. Hierbei zeigen sich bei der Metrik MSE recht gering ausgeprägte Unterschiede. Es ergibt sich ein Spektrum von 0.143 als Maximum bei der Vorhersage für die Grünen und ein Minimum von 0.066 bei den Linken. Für den Vergleich der Korrelationen ergeben sich etwas größere Unterschiede, wobei die Werte zwischen 0,2 für die SPD und -0,19 für die FDP verlaufen.

Tabelle 5.8: Test auf zeitliche Kausalität der Resampling-Verfahren

Ergebnisse Granger-Kausalität	
Verfahren	Signifikante Kausalität (Ja/Nein)
Tweet-basiert	Nein
Stratifiziert Tweet-basiert (n= 500)	Ja (FDP, Grüne)
Populations-basiert	Nein
Populations-geokodiert	Nein

Bei der Frage der zeitverschobenen Kausalität zwischen den Sentiment- und den Umfragedaten ließ sich für die Mehrzahl der Datenpunkte keine statistischer Zusammenhang erkennen. Die Granger-Kausalität wurde dabei bei jedem der 4 Verfahren für alle 6 Parteien angewandt, wobei auf eine maximale Verschiebung von 5 Tagen geprüft wurde. Dabei wird die Teststatistik pro Zeitintervall von Tagen berechnet. Damit lässt sich das Ausmaß der zeitlichen Verschiebung der Daten prüfen.

Aus Gründen der Übersichtlichkeit werden in diesem Kapitel nicht die Ergebnisse für jede Untersuchungseinheit dargestellt, sondern nur für die Verfahren angegeben, ob und für welche Parteien eine zeitlich verschobene Kausalität in den Daten festgestellt werden konnte (Tabelle 5.8). Für die Verfahren Tweet-basiert, Populations-basiert und Populations-geocodiert konnte für keine Partei die Nullhypothese der Nichtbeeinflussung verworfen werden. Dies war bei beiden üblichen Signifikanzniveaus (p = 0,05 bzw. p = 0,01) der Fall. Die Ausnahme war der Fall des Verfahrens Stratifiziert Tweet-basiert (mit n = 500). Dort konnte für die Grünen bei einem Lag von 4 Tagen eine signifikante Kausalität bei einem Signifikanzniveau von p= 0,05 festgestellt werden. Ebenso konnte bei der FDP auf diesem Signifikanzniveau eine Kausalität festgestellt werden, wobei hier für einen Lag von 3 Tagen ein Einfluss der Sentimentreihen auf die Umfragereihen festgestellt werden konnte.

Aufgrund der parteibezogenen Methodik der Messung der Übereinstimmung der Sentiment- und Umfragedaten ergaben sich besondere Herausforderungen bei den durchgeführten Workflows. Ein Tweet musste hierbei eindeutig genau einer Partei zugeordnet werden, was die Datenmenge reduzierte. Zudem musste der Fit der Daten auch zuerst für einzelne Parteien berechnet werden. Im Anschluss wurde dann der Durchschnitt dieser Werte für die einzelnen Verfahren berechnet, was wiederum Ungenauigkeiten erzeugt.

Die Wörter „Union“, „Laschet“, „ArminLaschet“, „Scholz“, „Olafscholz“ befinden sich unter den Top 20 der häufigsten Wörter in dem Gesamtcorpus parteibezogener Tweets. Wie zu erwarten zeigt sich, dass die Terme mit denen die Tweets gefiltert wurden, in dem Corpus sehr häufig vertreten sind. Dementsprechend sind auch die Namen der im Bundestags vertretenen Parteien häufig zu finden. Die Ausnahme bildet dabei die Partei „Die Linke“, obwohl auch hierbei mit dem Namen der Partei entsprechende Tweets gefiltertert wurden.

Abbildung 5.3: Top 20 Begriffe

Die meisten Tweets im verwendeten Corpus haben einen inhaltlichen Bezug zur CDU. Bezüglich der Tweetanzahl zweithäufigste Partei ist die SPD.

Mit Abstand folgen die kleinen Parteien (FDP, Grüne, Linke, AfD). Der Zugewinn an parteibezogenen Tweets durch die User-Timelines mit zusätzlich nicht geocodierten Tweets beträgt nur ca. 100%. Trotz der hohen Tweetanzahl der durch die User-Timelines gewonnen wurde, ist nur eine geringe Anzahl an Tweets (5,4 %) relevant für die Bundestagswahl 2021. Die Ergebnisse der verschiedenen Verfahren unterscheiden sich nur gering. Das Baseline-Verfahren (Tweet-basiert) hat eine geringe Korrelation, wie auch die anderen Verfahren. Die Werte der Granger-Kausalität deuten nur für die Grünen und die FDP bei stratifizierter Auswahl auf eine Kausalität zwischen den Daten.

Kapitel 6

Diskussion

Im Folgenden werden die in den Experimenten gewonnen Ergebnisse diskutiert und mögliche Ursachen dazu behandelt. Diese werden auf den Forschungsstand bezogen und Ähnlichkeiten zu den Studien aus Kapitel 2 werden dabei referiert.

6.1 Diskussion der Ergebnisse aus dem Experiment zu Corona

In Bezug auf die allgemeine Evaluation ist es schwierig größere Trends zu erkennen. Die Unterschiede sind teilweise sehr spezifisch. Tendenziell ist der MSE, bei dem Vergleich des Durchschnitts von Keyword-basierten und nicht Keyword-basierten Verfahren (ohne Verfahren 2, da hier kein Resampling der Keyword-basierten Daten möglich ist), niedriger bei den Keywordbasierten Verfahren. Bei dem Vergleich des Pearson r gibt es keine großen Unterschiede. Allerdings ist hier zu beachten, dass die Korrelationskoeffizienten in diesem Fall Werte mit negativen und positiven Vorzeichen aufweisen. Bei der Berechnung des Durchschnitts der Werte addieren sich dementsprechend diese zu 0. Dies erschwert die Interpretation der Berechnung dieser Durchschnitte. Bei dem Vergleich des Mean-Squared-Errors schneidet Verfahren 4 im Durchschnitt von Keyword- und nicht-keywordgefilterten Tweets am besten ab.

In der Forschungsliteratur (siehe Kapitel 2.2) wurde auf den Fakt hingewiesen, dass bei dem Vergleich von Umfrage- und Zeitreihendaten besondere Vorverarbeitungsschritte benötigt werden. Häufig sind die Zeitreihen nicht von sich aus zeitlich synchron. Um zu prüfen, ob eine Verschiebung um einen variablen Lag-Parameter vorliegt, wurde die Granger-Kausalität berechnet. Für die vorliegenden Zeitreihendaten konnte auf eine Verschiebung von bis zu 8 Intervallen getestet werden mit der gegebenen Anzahl an Daten. In Anbetracht der Tatsache, dass ein Interval in den meisten Fällen 14 Tage umfasst, ist die An-

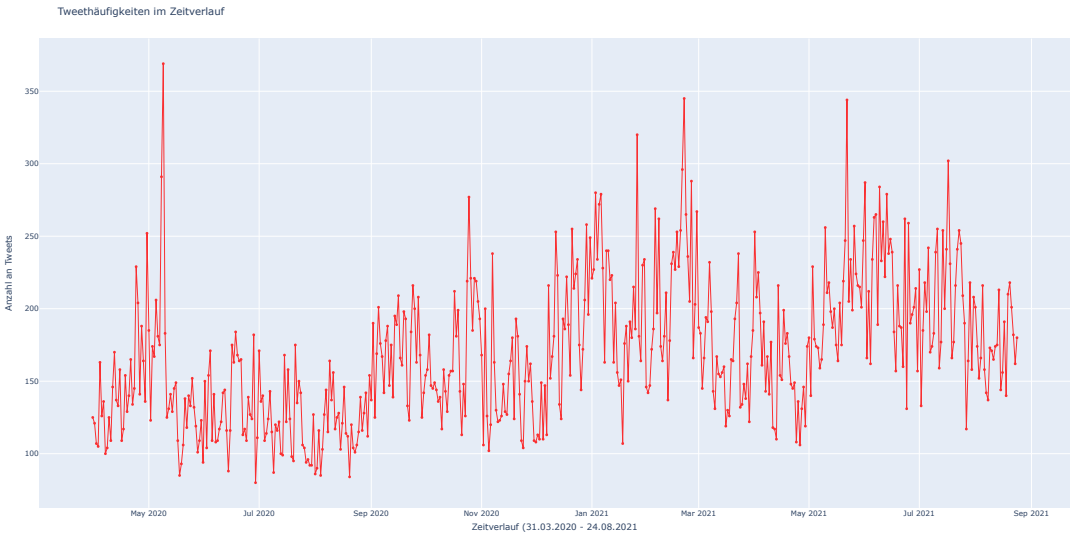

Abbildung 6.1: Tagesbasierte Anzahl der Tweets in dem Experiment zu Corona

zahl der möglichen Intervalle mehr als ausreichend für einen Vergleich. Hierbei wurde für die Verfahren mit den geringsten MSE und der höchsten Korrelation auf Beeinflussung getestet, jedoch zeigt sich in diesen Fällen, dass der p-Wert das gesetzte Signifikanzniveau (0,05 und 0,01) überschreitet, was bedeutet, dass die Nullhypothese der Nichtbeinflussung nicht verworfen werden kann.

Visuell die größte Übereinstimmung der beiden Graphen liegt bei Verfahren 3 mit Keywords vor (siehe Abbildung 4.3). Dies ist das Verfahren mit der visuell betrachtet geringsten Abweichung. Hier ist ab Ende Oktober 2020 eine gewisse Übereinstimmung der Richtung der Graphen festzustellen. Die Werte davor lassen nicht auf eine Übereinstimmung deuten. Der Verlauf der Kurven ist ähnlich, aber es sind weniger Daten vorhanden, da zwischen Mitte Juli 2020 und dem Oktober 2020 weniger Datenpunkte vorliegen, sodass der bestehende Graph verzerrt wird. Jedoch zeigt sich, dass das Verfahren mit der höchsten Korrelation visuell auf den ersten Blick einer weniger gute Übereinstimmung suggeriert als der Plot des Verfahren mit der höheren Korrelation. Jedoch sollte hier den Ergebnissen der Metriken den Vorrang gegeben werden, da sie eindeutiger sind.

Zusammenfassend lässt sich sagen, dass im Vergleich der Verfahren grobe Tendenzen in Bezug auf die Effektivität der Verfahren erkennbar sind. Dabei zeigt sich, dass manche Verfahren eine bessere Anpassung an die Umfragedaten ermöglichen als das Baseline-Verfahren (Verfahren 1). Dabei schneidet Verfahren 4 (mit allen Tweets) besonders gut ab. Ebenso ist Verfahren 2 (stratifiziert-tweetbasiert) in Bezug auf die Korrelation eine Verbesserung. Dies deutet daraufhin, dass die Anpassung der geographischen Verteilung der

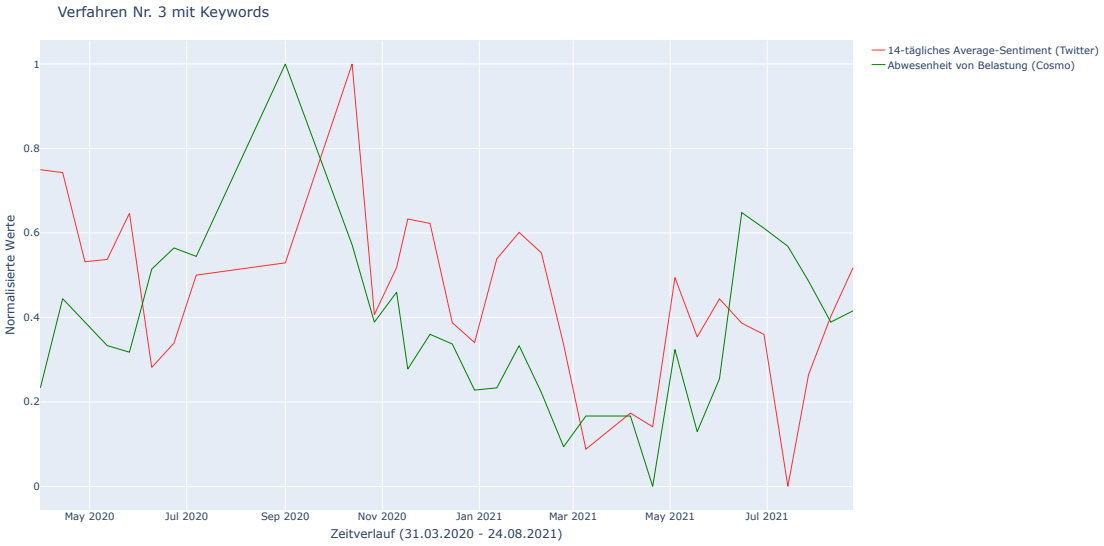

Abbildung 6.2: Plot der Ergebnisse für Verfahren 3 (Populations-basiert) mit Keywords

User in dem resampelten Datensatz eine Verbesserung ermöglicht. Die Bildung eines Panels mit den User-Timelines führt dagegen nur bei der Verwendung von keyword-bezogenen Tweets zur Verbesserung. Bei der Verwendung von allen Tweets gehen offenbar die bestehenden Tendenzen im Rauschen der großen Datenmenge unter, da hierbei absolut betrachtet die höchste Tweetanzahl vorliegt. Für weitere Arbeiten wäre es vielversprechend einen keyword-basierten Tweetdatensatz zu stratifizieren, um zu testen ob eine gleichmäßigere regionale Verteilung an Tweets zu einer besseren Abschätzung von Sentimentdynamiken führt. Bei den Störfaktoren kann jedoch die Hypothese einer zeitlichen Verschiebung von anderweitig matchenden Zeitreihen ausgeschlossen werden. Der Test auf Granger-Kausalität bestätigte hierbei keine zeitlich nach Intervall variierende Vorhersage der Umfragedaten durch die Sentimentdaten.

6.2 Diskussion der Ergebnisse aus dem Experiment zu Wahlen

Im Kapitel 2.2 wurde bereits auf das Problem der zeitlichen Synchronizität zwischen den Sentimentdaten aus Social Media und den verwendeten Umfragedaten verwiesen. Bei den verwendeten Verfahren zur Umfragenapproximation konnten wenig Belege für eine zeitliche Verschiebung in den Tweetdaten gefunden werden. Allein für das Verfahren Stratifiziert Tweet-basiert wurde ein zeitlicher Lag mithilfe der Granger-Kausalität festgestellt. Dieser wurde für die Parteien Grüne und FDP festgestellt, aber nicht für die anderen Parteien (CDU, SPD, Linke, AfD). Eine weitere Einschränkung ist, dass dabei die Ergebnisse auf dem weniger restriktiven Signifikanzniveau von 0,05 vorlagen, aber nicht auf auf dem Signifikanzniveau von 0,01. Generell spricht dies gegen einen zeitlichen Lag in der Übereinstimmung zwischen Twittersentiment- und Umfragedaten.

Zudem wurde untersucht, ob das Vorhandensein von Retweets Einfluss auf die aggregierten Sentimentscores von Usern hat. Forscher stellten bei der Verwendung von Retweets eine Tendenz zur Verstärkung von positiven und negativen Peaks des Sentiments und eine Abschwächung periodischer Effekte fest Rodríguez-Ibáñez et al. [2020]. Deshalb wurde in diesem Experiment eine getrennte Analyse für einen Datensatz ohne Retweets aus den User-Timelines durchgeführt. Allerdings konnten hierbei nur geringe Effekte bei der Auslassung von Retweets in der Analyse festgestellt werden. Der MSE für die User-Timelines ohne Retweets betrug 0,103. Damit ist der Wert geringfügig besser als bei einer Analyse mit Retweets. Hierbei befand sich der MSE in der Spanne von 0,105 bis 0,106. Für die Pearson-Korrelation konnte mit einem Wert von 0,028 eine minimale Verbesserung im Vergleich mit dem Verfahren 3 (Populations-basiert) festgestellt werden. Generell zeigen die Ergebnisse, dass der Einfluss von Retweets auf die Berechnungen der Sentiment-Scores gering ist.

Ein Problem bei der Messung des Sentiments von Parteien auf Twitter als Proxy der Wählerzustimmung zu diesen Parteien ist die Besonderheit der User-Accounts. Hierbei lassen sich häufig offizielle Accounts von politischen Parteien bzw. Parteimitgliedern auf Twitter finden. Diese lassen sich damit nicht in jedem Fall als Ersatz für Wählerbefragungen verwenden. Dies lässt sich bei der Recherche der Metadaten der User-Accounts feststellen. Häufig findet sich hierbei die Nennung der entsprechenden Partei im Feld „description“ oder „name“. Eine explorative Kontrolle in den Daten stellte dies verstärkt bei Usern fest, die immer zu genau einer Partei getweetet haben. Sind diese Accounts zu stark vertreten, weicht das Twitter-Sample stark von der existierenden Wählerstruktur ab. Letztlich soll die Messung der Tweets dazu dienen

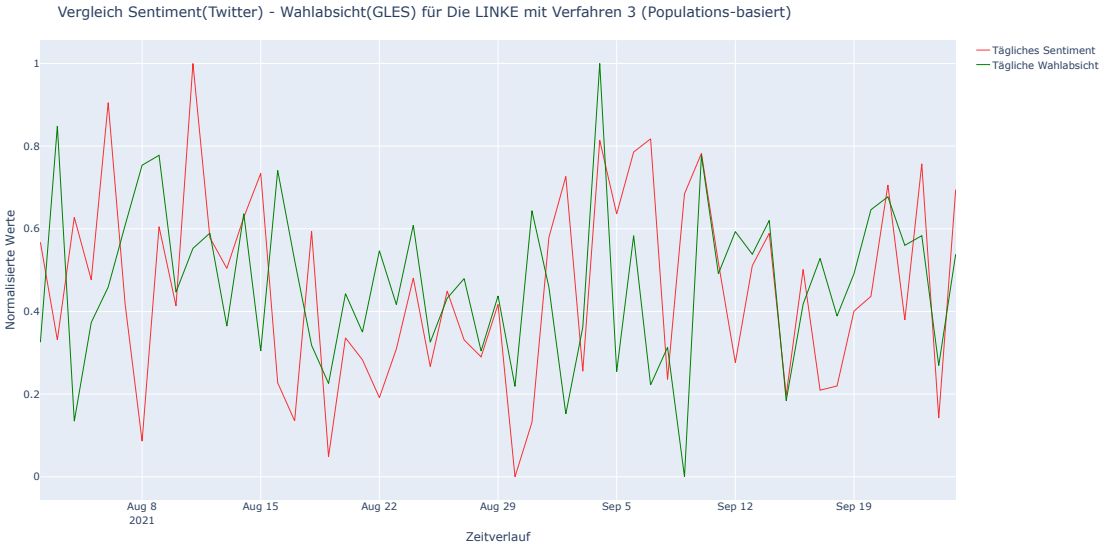

Abbildung 6.3: Plot der Ergebnisse für Die Linke

die Bewertung der Parteien durch potentielle Wähler zu erfassen. Dabei stellen diese Accounts einen möglichen Störfaktor dar, da diese User tendenziell positiver über die entsprechenden Parteien berichten dürften. Ebenso ist aufgrund der hohen Präsenz von Parteimitgliedern, hierbei von einer Überrepräsentation dieser Gruppe auszugehen. Der Anteil von SPD-Mitgliedern an der deutschen Gesamtbevölkerung beträgt 0,5%. Bei den 489 Usern, die allein zur SPD posten, konnte eine explorative Durchsicht eines Samples von 50 Usern, bereits 6 (12%) User identifizieren die Parteimitglied der SPD sind. Bei einer Filterung der Tweets anhand der User-Beschreibung konnten bei den SPD-bezogenen Tweets 6,1% der Tweets diesen Usern zugeordnet werden. Die Werte für die anderen Parteien befinden sich auch im einstelligen Prozentbereich. Damit ist zwar davon auszugehen, dass die Parteiaccounts überrepräsentiert sind, aber bei den Tweetzahlen nicht dermaßen ins Gewicht fallen, dass bei einer Entfernung die Datenmenge zu stark reduziert wär.

Eine besondere Problematik in diesem Experiment ergibt sich durch die Notwendigkeit die Umfrageergebnisse für die 6 Parteien im Bundestag (SPD, CDU, Grüne, FDP, Linke, AfD) vorherzusagen, statt einem Wahlergebnis zwischen 2 Kandidaten. Studien zur Prognose von Umfrageergebnissen wurden häufig in Staaten mit Mehrheitswahlrecht, in denen Wahlentscheidungen zwischen 2 Personen bzw. Parteien stattfinden. O'Connor et al. [2010] führte dies am Beispiel der amerikanischen Präsidentenwahl 2008 zwischen Barack Obama und John McCain durch. Dabei zeigten sich bereits bei dem einfacheren Setting dieser Studie Probleme eine Korrelation zwischen Sentiment Scores und Wahlergebnisse zu finden. In Fernández et al. [2017] wurde eine Studie zur

Vorhersage der Wahlumfrage im Rahmen der französischen Präsidentschaftswahl 2017 durchgeführt. Bei dieser Wahl traten 11 Kandidaten an Le Monde [2017], aber in der Studie wurde eine Wahlumfrage zur Stichwahl zwischen Emmanuel Macron und Marine LePen erfolgreich approximiert. Ein ähnliches Setting findet sich in Deutschland allerdings nur für Kommunalwahlen. Für den Bundestag am stärksten relevant ist die Zweitstimme, mit der die Wähler ihre Parteipräferenz angeben. Dementsprechend müssen durch die Verfahren zur Bestimmung der Sentimentscores auf Twitter die Wahlanteile von aktuell 6 Parteien geschätzt werden, was komplizierter ist als die Wahlstimmenanteile für 2 Kandidaten zu schätzen.

Kapitel 7

Fazit

In dieser Arbeit wurde eine Methodik zur Approximation von Umfragedaten entwickelt. Dazu wurden 4 verschiedene Resamplingverfahren eingesetzt. Bei der Tweet-basierten Auswahl (Verfahren 1) wurde die Gesamtmenge geocodierter Tweets verwendet. Bei der Stratifiziert Tweet-basierten Auswahl (Verfahren 2) wurde diese Menge an Tweets stratifiziert gesampelt, um eine gleichmäßigere, räumliche Verteilung an Tweets zu erhalten. Bei der Populations-basierten Auswahl (Verfahren 3) wurden alle Tweets von Usern verwendet, deren Ort maschinell bestimmt wurde. Bei der Populations-geocodierten Auswahl (Verfahren 4) erfolgt die Auswahl der User nach dem gleichen Verfahren, allerdings wurden hierbei nur geocodierte Tweets der User verwendet.

In der vorliegenden Arbeit wurde in 2 Experimenten diese Methodik zum Resampling von Twitter-Daten getestet. Dazu wurden in Experiment 1 Sentimentzeitreihen von Twitter mit Umfragedaten zu psychischen Belastungen in Folge der Corona-Epidemie verglichen und in Experiment 2 Sentimentzeitreihen von Twitter zu politischen Parteien mit Daten einer Wahlumfrage verglichen.

Die Ergebnisse aus dem Experiment zu Corona sind selber gemischt. Es ließen sich in geringem Maß Verbesserungen im Vergleich mit dem Baseline-Verfahren (Tweet-basiert) in der Übereinstimmung zwischen den Zeitreihen feststellen. Die Ergebnisse variieren hierbei zwischen den Verfahren. Die Verfahren mit der Bestimmung der User-Orte lieferten geteilte Resultate. Teilweise ließen sich höhere Werte für die Korrelationen finden, aber auch negative Werte, die auf einen gegenläufigen Zusammenhang hindeuten. Zudem zeigte sich, dass die Resamplingverfahren mit Keyword-gefilterten Tweets nicht zwingend besser abschnitten als die gleichen Resamplingverfahren ohne Keyword-gefilterte Tweetauswahl. Dementsprechend lässt sich nicht allgemein sagen, dass die Keyword-basierten Verfahren die Umfrage besser approximierten.

Anschließend wurde die Methodik in einem zweiten Experiment auf Wahl-

umfragen angewandt. Dabei wurden wiederum Sentimentzeitreihen von Twitter mit Daten aus einer Wahlumfrage verglichen. Bei den Ergebnissen zeigte sich, unter Verwendung aller Metriken, nur eine sehr geringe Übereinstimmung zwischen den beiden Zeitreihen. Dabei war die Streuung der Ergebnisse für die verschiedenen Resampling-Verfahren sehr gering. Dies deutet daraufhin, dass für die gegebenen Wahldaten mit den Resamplingverfahren keine Annäherung der Ergebnisse an die Umfragedaten erreicht werden kann.

Bei einem Vergleich der Ergebnisse aus den beiden Experimenten lässt sich feststellen, dass für die Analyse der Corona-bezogenen Daten teilweise bessere Annäherungen an die Umfragedaten erreicht wurden als bei dem Experiment zu Wahldaten. Besonders bei den Korrelationen ergaben sich höhere Werte im ersten Experiment. Dies betrifft die dem Verfahren Populations-geocodiert (mit allen Tweets), sowie Populations-basiert(mit Corona-bezogenen Tweets). Gemeinsam ist beiden Experimenten, dass die Bewertung von Tweets durch die Sentimentlabelung erfolgte. Eine Sichtung einer (kleineren) Datenmenge zeigte, dass die Genauigkeit der Klassifikation sich im oben mittleren Bereich bewegte, aber es hier noch Verbesserungsmöglichkeiten gibt. Das Problem dabei ist, dass Ungenauigkeiten bei der Sentimentklassifikation Auswirkungen auf die nachfolgenden Verarbeitungsschritte, wie die Aggregierung der Sentimentscores zu Zeitreihen, haben. Damit werden bestehende Fehler in die anderen Schritte der Vorprozessierung übertragen.

Als Schlussfolgerung lässt sich ziehen, dass Resampling insgesamt nur bedingt bei der Approximation der Umfragedaten hilfreich war. Für die Approximierung von Wahlumfragen ließen sich bei den gegebenen Daten keine positiven Resultate finden. Jedoch scheint die Resamplingmethodik für Zeitreihen zu psychischer Gesundheit und ähnlichen Themen möglicherweise hilfreicher, da für die Corona-bezogenen Belastungen bessere Resultate erzielt wurden.

Eine Annahme dieser Arbeit ist es, dass die zu untersuchenden Trends so weit verbreitet sind, dass sich deren Einfluss sowohl auf Personen im Twitter-Sample, als auch im Umfragen-Sample erstreckt. Mit den verfügbaren Daten ist kein Matching von Personendatensätzen möglich (vgl. Beuthner et al. [2021]). Dabei ist unklar, ob es eine Schnittmenge an Personen gibt, die in beiden Stichproben (Twittersample und Umfragensample) enthalten ist. Die Möglichkeit ist gegeben, dass hierbei 2 verschiedene Stichproben von Personen verglichen werden, die sich sehr stark voneinander unterscheiden.

Der Fokus dieser Arbeit war das Resampling bezogen auf die geografische Verteilung der User. Dazu wurden Stratifikation und ein Verfahren zur maschinellen Ortsbestimmung angewandt. Allerdings lässt sich Resampling prinzipiell auch anhand soziodemografischer Variablen, wie Alter oder Geschlecht, durchführen. Dazu werden allerdings User-Daten mit Labels zu diesen Variablen benötigt. Jedoch stellt sich bei Twitter die Problematik, dass es hierzu

keine Angaben gibt, anders als bei Standorten von denen Tweets abgesetzt werden. Somit muss eine maschinelle Bestimmung dieser User-Eigenschaften vorgenommen werden. In zukünftigen Arbeiten könnte an diesem Punkt angesetzt werden, um zu untersuchen, ob dabei informativere Ergebnisse erzielt werden können.

Anhang A

Keywordlisten

A.1 Experiment 1

Aufgrund der hohen Anzahl der Filterterme wurden diese nicht in den Anlagen angegeben, sondern befinden sich im Git-Repository dieser Arbeit. Link: https://git.webis.de/code-teaching/theses/thesis-halbauer/-/blob/main/data/ressources/corona_

A.2 Experiment 2

Verwendete Keywords/Hashtags zur Filterung von Tweets zu Parteien:

„#spd", „@spdde", „spd", „#scholz", „@OlafScholz", „scholz", „#cdu", „#union", „cdu", „union", „#laschet", „@ArminLaschet", „laschet", „#gruene",„#diegruenen", „gruene", „diegruenen", „#baerbock", „@ABaerbock",„baerbock", „#fdp", „fdp", „#lindner", „lindner", „@c_lindner", „#linke", „#dielinke", „dielinke", „linke", „#wissler", „#bartsch", „wissler", „bartsch", „@dietmarbartsch", „@janine_wissler", „#alternativefuerdeutschland", „afd", „alternativefuerdeutschland", „#weidel", „#chrupalla", „weidel", „chrupalla", „@alice_weidel", „@tino_chrupalla", #afd",

Literaturverzeichnis

Brooke Auxier and Monica Anderson. Social media use in 2021. Technical report, Pew Research Center, April 2021.

Tarek Al Baghal, Alexander Wenz, Luke Sloan, and Curtis Jessop. Linking twitter and survey data: asymmetry in quantity and its impact. *EPJ Data Science*, 10(1), June 2021. doi: 10.1140/epjds/s13688-021-00286-7. URL `https://doi.org/10.1140/epjds/s13688-021-00286-7`.

Natalie Beisch and Carmen Schäfer. Ergebnisse der ard/zdf-onlinestudie 2020. internetnutzung mit großer dynamik:medien, kommunikation, social media. *Media Perspektiven*, September 2020.

Cornelia Betsch, Lothar Wieler, Michael Bosnjak, Michael Ramharter, Volker Stollorz, Saad Omer, Lars Korn, Philipp Sprengholz, Lisa Felgendreff, Sarah Eitze, and Philipp Schmid. Germany covid-19 snapshot monitoring (cosmo germany): Monitoring knowledge, risk perceptions, preventive behaviours, and public trust in the current coronavirus outbreak in germany. *PsychArchives*, Mar. 2020. doi: 10.23668/psycharchives.2776. URL `https://psycharchives.org/en/item/e5acdc65-77e9-4fd4-9cd2-bf6aa2dd5eba`.

Christopher Beuthner, Johannes Breuer, and Stefan Jünger. Data linking - linking survey data with geospatial, social media, and sensor data. *GESIS Survey Guidelines*, 2021. doi: 10.15465/GESIS-SG_EN_039. URL `http://shorturl.at/ETU69`.

Yu Cui and Quing He. Inferring twitters' socio-demographics to correct sampling bias of social media data for augmenting travel behavior analysis. *Journal of Big Data Analytics in Transportation*, 3(2):159–174, March 2021. doi: 10.1007/s42421-021-00037-0. URL `https://doi.org/10.1007/s42421-021-00037-0`.

Andreas Diekmann. *Empirische Sozialforschung: Grundlagen, Methoden, Anwendungen*. Rororo Rowohlts Enzyklopädie. Rowohlt Taschenbuch Verlag, Reinbek bei Hamburg, 2018.

Javi Fernández, Fernando Llopis, Yoan Gutiérrez, Patricio Martínez-Barco, and Álvaro Díez. Opinion mining in social networks versus electoral polls. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 231–237, Varna, Bulgaria, September 2017. INCOMA Ltd. doi: 10.26615/978-954-452-049-6_032. URL `https://doi.org/10.26615/978-954-452-049-6_032`.

GESIS. GLES-Design. `https://gles.eu/gles/das-gles-design/#waehlerinnen`, 2022a. Letzter Zugriff: 2022-07-20.

GESIS. GLES Rolling Cross-Section, 2022b. Letzter Zugriff: 2022-07-21.

GESIS. Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUS 2021. GESIS, Köln. ZA5280 Datenfile Version 1.0.0, https://doi.org/10.4232/1.13954, 2022c.

Oliver Guhr, Anne-Kathrin Schumann, Frank Bahrmann, and Hans Joachim Böhme. Training a broad-coverage german sentiment classification model for dialog systems. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1620–1625, Marseille, France, May 2020. European Language Resources Association. URL `https://www.aclweb.org/anthology/2020.lrec-1.202`.

Yuli Patrick Hsieh and Joe Murphy. *Total Twitter Error*, chapter 2, pages 23–46. John Wiley and Sons, Ltd, 2017. ISBN 9781119041702. doi: https://doi.org/10.1002/9781119041702.ch2. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119041702.ch2`.

Anna M. Kruspe, Matthias Häberle, and Iona Kuhn und Xiao Xiang Zhu. Cross-language sentiment analysis of european twitter messages duringthe COVID-19 pandemic. *CoRR*, abs/2008.12172, 2020. URL `https://arxiv.org/abs/2008.12172`.

Le Monde. Qui sont les candidats pour la présidentielle 2017 ? `http://www.shorturl.at/brtV2`, 2017. Accessed: 2022-08-11.

Leibniz-Institut für deutsche Sprache. Neuer Wortschatz rund um die Coronapandemie. `https://www.owid.de/docs/neo/listen/corona.jsp`, 2022. Letzter Zugriff: 2022-05-15.

Lewis Mitchell, Morgan R. Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M. Danforth. The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PLOS ONE*, 8(5):1–15, 05 2013. doi: 10.1371/journal.pone.0064417. URL `https://doi.org/10.1371/journal.pone.0064417`.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Fourth international AAAI conference on weblogs and social media*, 2010.

Ishaani Priyadarshini, Pinaki Mohanty und Raghvendra Kumar, Rohit Sharma, Vikram Puri, and Pradeep Kumar Singh. A study on the sentiments and psychology of twitter users during covid-19 lockdown period. *Multimedia Tools and Applications*, June 2021. doi: 10.1007/s11042-021-11004-w. URL `https://doi.org/10.1007/s11042-021-11004-w`.

Khairiyah Mohamed Ridhwan and Carol Anne Hargreaves. Leveraging twitter data to understand public sentiment for the covid-19 outbreak in singapore. *International Journal of Information Management Data Insights*, 1(2): 100021, 2021. ISSN 2667-0968. doi: https://doi.org/10.1016/j.jjimei.2021.100021. URL `https://www.sciencedirect.com/science/article/pii/S2667096821000148`.

Margarita Rodríguez-Ibáñez, Francisco-Javier Gimeno-Blanes, Pedro Manuel Cuenca-Jiménez, Sergio Muñoz-Romero, Cristina Soguero, and José Luis Rojo-Álvarez. On the statistical and temporal dynamics of sentiment analysis. *IEEE Access*, 8:87994–88013, 2020. doi: 10.1109/ACCESS.2020.2987207.

Rainer Schnell, Paul B. Hill, and Elke Esser. *Methoden der empirischen Sozialforschung*. Oldenbourg, München, 10., überarb. aufl. edition, 2013. ISBN 9783486728996. URL `http://d-nb.info/1043389342/04`.

Siegfried Schumann. *Repräsentative Umfrage*. Oldenbourg Verlag, 2012.

Marijn Wang, Shihan und Schraagen, Erik Tjong Kim Sang, and Mehdi Dastani. Dutch general public reaction on governmental covid-19 measures and announcements in twitter data, 2020. URL `https://arxiv.org/abs/2006.07283`.

Stefan Wojcik and Adam Hughes und Shawnee Cohn. Sizing up twitter users. Technical report, Pew Research Center, April 2019.

Feng Xu, Kuai und Wang, Xiaohua Jia, and Haiyan Wang. The impact of sampling on big data analysis of social media: A case study on flu and ebola. In *2015 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6, 2015. doi: 10.1109/GLOCOM.2015.7416974.