HAW HAMBURG

BACHELOR THESIS
Valeria Silver

Ein NLP-basierter Ansatz zur Analyse des Informationsflusses in russischen Medien

FAKULTÄT TECHNIK UND INFORMATIK
Department Informatik

Faculty of Engineering and Computer Science
Department Computer Science

HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN HAMBURG
Hamburg University of Applied Sciences

Valeria Silver

Ein NLP-basierter Ansatz zur Analyse des Informationsflusses in russischen Medien

Bachelorarbeit eingereicht im Rahmen der Bachelorprüfung
im Studiengang *Bachelor of Science Angewandte Informatik*
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuender Prüfer: Prof. Dr. Marina Tropmann-Frick
Zweitgutachter: Prof. Dr. Stefan Sarstedt

Eingereicht am: 12. Oktober 2023

Valeria Silver

Thema der Arbeit

Ein NLP-basierter Ansatz zur Analyse des Informationsflusses in russischen Medien

Stichworte

Linguistische Datenverarbeitung (NLP), Eigennamenerkennung (NER), Kreml, Ukraine, russische Medien

Kurzzusammenfassung

Seit dem Kriegsbeginn in der Ukraine im Jahr 2022 wurden unabhängige Medien und Journalisten in Russland unter erhöhtem Druck gesetzt. Viele von denen wurden als Propaganda bezeichnet, sollten das Land verlassen oder ihre Tätigkeit komplett einstellen. Andererseits sollten pro-kremlische Medien sich an die neuen Gesetze anpassen. Das hat zu einer Reduzierung der oppositionellen Nachrichtenverlage und dadurch zu einer Verringerung der Meinungsvielfältigkeit geführt. In dieser Bachelorarbeit wird das Verhaltensmuster von pro-kremlischen und oppositionellen Medien im Jahr 2022 analysiert. Unter dem Einsatz von Techniken der natürlichen Sprachverarbeitung (NLP) wird eine umfassende Datenanalyse vorgenommen.

Valeria Silver

Title of Thesis

An NLP-based approach for analyzing information flow in Russian media

Keywords

Natural Language Processing (NLP), Named Entity Recognition (NER), Kremlin, Ukraine, russian media

Abstract

Since the outbreak of the war in Ukraine in 2022, independent media and journalists in Russia have been put under increased pressure. Many of them were labeled as propagandists, had to leave the country, or cease their activities entirely. In contrast, pro-Kremlin media adapted to new regulations, leading to a decline in the number of opposition news outlets and, consequently, a reduction in the diversity of opinions.This bachelor thesis analyzes the behavioral patterns of both pro-Kremlin and opposition media during 2022. Leveraging natural language processing (NLP) techniques, a comprehensive data analysis is conducted.

Inhaltsverzeichnis

Abbildungsverzeichnis **vi**

Tabellenverzeichnis **ix**

Abkürzungen **x**

1 Einleitung **1**
1.1 Motivation . . . 1
1.2 Aufgabenstellung . . . 1
1.3 Aufbau der Bachelorarbeit . . . 2

2 Theoretische Grundlagen **3**
2.1 Klassischer Ansatz der NLP . . . 3
2.1.1 Textvorverarbeitung . . . 3
2.1.2 Lexikasche Analyse . . . 5
2.1.3 Syntaktisches Parsing . . . 6
2.1.4 Semantische Analyse . . . 7
2.2 Empirische und statistische Ansätze . . . 8
2.2.1 Empirische Ansätze . . . 8
2.2.2 Statistische Ansätze . . . 9
2.2.3 Named Entity Recognition . . . 10
2.3 Sprachmodelle . . . 11
2.3.1 „Transformer“-Architektur . . . 11
2.3.2 BERT . . . 12
2.3.3 GPT . . . 13
2.4 Diskussion . . . 13

3 Daten **15**
3.1 Politischer Kontext . . . 15
3.1.1 Relevante Begriffe . . . 16
3.1.2 Neusprech . . . 19
3.2 Auswahl der Datenquellen . . . 20
3.3 Informationskanäle . . . 21
3.3.1 VKontakte . . . 22
3.3.2 Telegram . . . 23

3.4 Datensammlung und Datenverarbeitung 23
3.4.1 Datensammlung 23
3.4.2 Datenbereinigung 27
3.4.3 Übersetzung 28
3.5 Diskussion 29

4 Technischer Hintergrund **30**
4.1 Auswahl der Bibliotheken 30
4.2 Verwendung von Tools 32
4.3 Pipeline 33
4.4 Diskussion 37

5 Experimente **38**
5.1 Voranalyse der Eingangsdaten 38
5.1.1 Überblick über die Daten 38
5.1.2 Anzahl der Artikel pro Tag 40
5.1.3 Worthäufigkeit pro Medium 43
5.2 Analyse des Kontextes ausgewählter Wörter 45
5.2.1 Krieg 46
5.2.2 Spezialoperation 49
5.2.3 Diskussion 52
5.3 Untersuchung der Zusammenhänge zwischen Entitäten 53
5.3.1 Graphenvorbereitung 53
5.3.2 Graphenanalyse 56
5.4 Diskussion 61

6 Fazit **62**
6.1 Zusammenfassung 62
6.2 Ausblick 63

Literaturverzeichnis **65**

A Anhang **71**
A.1 Die Architektur des Transformer-Modells 71
A.2 Die Anzahl der Nachrichten pro Tag bei allen Medien im Jahr 2022 . . . 72
A.3 Worthäufigkeit der ausgewählter Wörter in allen Medien im Jahr 2022 . 78
A.4 Verwendungshäufigkeit des Wortes „Krieg“ in allen Medien im Jahr 2022 81
A.5 Verwendungshäufigkeit des Wortes „Spezialoperation“ in allen Medien im Jahr 2022 87
Selbstständigkeitserklärung 93

Abbildungsverzeichnis

2.1 Die Phasen der Analyse bei der Natural language processing (NLP) [29] 4
2.2 Syntaxbaum des Satzes „the old man a ship“ [29] 6

3.1 Führende soziale Medienplattformen in Russland im 3. Quartal 2022 nach monatlicher Durchdringungsrate [52] 22
3.2 Ablaufdiagramm der Datenextrahierung mithilfe des VK API 25
3.3 Ablaufdiagramm der Datenextrahierung mithilfe des Telegram API . . . 26

4.1 Prozessablauf der Daten vom Datensammlung bis zur Visualisierung . . 33
4.2 Ablaufdiagramm des Skripts zur Identifikation häufig wiederholter Wörter 34
4.3 Ablaufdiagramm des Skripts zur Analyse ausgewählter Wörter 35
4.4 Ablaufdiagramm des Skripts zur Analyse des Kontexts ausgewählter Wörter 36
4.5 Ablaufdiagramm des Skripts zur Untersuchung der Zusammenhänge zwischen Entitäten 36

5.1 Anzahl der Nachrichten pro Tag bei Meduza im September 2022 41
5.2 Anzahl der Nachrichten pro Tag bei Mediazona im September 2022 . . . 41
5.3 Anzahl der Nachrichten pro Tag bei Ria im September 2022 42
5.4 Anzahl der Nachrichten pro Tag beim Ersten Kanal im September 2022 42
5.5 Diagramm der Verwendungshäufigkeit von ausgewählten Wörtern im Februar und März 2022 45
5.6 Wortwolke des Wortes „Krieg“ bei Ria im Jahr 2022 47
5.7 Wortwolke des Wortes „Krieg“ beim Ersten Kanal im Jahr 2022 47
5.8 Wortwolke des Wortes „Krieg“ bei Meduza im Jahr 2022 48
5.9 Wortwolke des Wortes „Krieg“ bei Mediazona im Jahr 2022 48
5.10 Wortwolke des Wortes „Spezialoperation“ bei Ria im Jahr 2022 50
5.11 Wortwolke des Wortes „Spezialoperation“ beim Ersten Kanal im Jahr 2022 50
5.12 Wortwolke des Wortes „Spezialoperation“ bei Meduza im Jahr 2022 . . . 51
5.13 Wortwolke des Wortes „Spezialoperation“ bei Mediazona im Jahr 2022 . 51
5.14 Datengraph ohne Kanten, Darstellung in Gephi 54
5.15 Datengraph mit Kanten, Darstellung in Gephi 54
5.16 Datengraph nach der Anwendung von Force Atlas-2 Layout Algorithmus 54
5.17 Datengraph mit der farblichen Rangordnung 55
5.18 Datengraph mit der farblichen Größenranganordnung 55
5.19 Datengraph nach der Anwendung vom Grad-Filter 55

5.20 Datengraph auf Basis der Daten aller vier Medien 56
5.21 Datengraph auf Basis der Daten von Meduza 57
5.22 Datengraph auf Basis der Daten von Mediazona 58
5.23 Datengraph auf Basis der Daten von Ria 59
5.24 Datengraph auf Basis der Daten vom Ersten Kanal 60

A.1 Die Architektur des Transformer-Modells [60] 71
A.2 Die Anzahl der Nachrichten pro Tag bei allen Medien im Januar-Februar 2022 72
A.3 Die Anzahl der Nachrichten pro Tag bei allen Medien im März-April 2022 73
A.4 Die Anzahl der Nachrichten pro Tag bei allen Medien im Mai-Juni 2022 74
A.5 Die Anzahl der Nachrichten pro Tag bei allen Medien im Juli-August 2022 75
A.6 Die Anzahl der Nachrichten pro Tag bei allen Medien im September-Oktober 2022 76
A.7 Die Anzahl der Nachrichten pro Tag bei allen Medien im November-Dezember 2022 77
A.8 Worthäufigkeit der ausgewählter Wörter in allen Medien im Januar-April 2022 78
A.9 Worthäufigkeit der ausgewählter Wörter in allen Medien im Mai-August 2022 79
A.10 Worthäufigkeit der ausgewählter Wörter in allen Medien im September-Dezember 2022 80
A.11 Verwendungshäufigkeit des Wortes „Krieg“ in allen Medien im Januar-Februar 2022 81
A.12 Verwendungshäufigkeit des Wortes „Krieg“ in allen Medien im März-April 2022 82
A.13 Verwendungshäufigkeit des Wortes „Krieg“ in allen Medien im Mai-Juni 2022 83
A.14 Verwendungshäufigkeit des Wortes „Krieg“ in allen Medien im Juli-August 2022 84
A.15 Verwendungshäufigkeit des Wortes „Krieg“ in allen Medien im September-Oktober 2022 85
A.16 Verwendungshäufigkeit des Wortes „Krieg“ in allen Medien im November-Dezember 2022 86
A.17 Verwendungshäufigkeit des Wortes „Spezialoperation“ in allen Medien im Januar-März 2022 87
A.18 Verwendungshäufigkeit des Wortes „Spezialoperation“ in allen Medien im April-Mai 2022 88
A.19 Verwendungshäufigkeit des Wortes „Spezialoperation“ in allen Medien im Juni-Juli 2022 89
A.20 Verwendungshäufigkeit des Wortes „Spezialoperation“ in allen Medien im August-September 2022 90

A.21 Verwendungshäufigkeit des Wortes „Spezialoperation“ in allen Medien im Oktober-November 2022 91
A.22 Verwendungshäufigkeit des Wortes „Spezialoperation“ in allen Medien im Dezember 2022 92

Tabellenverzeichnis

3.1 Beispiel der Kennzeichnung des ausländischen Agenten [11] 17
3.2 Beispiele von Neusprech aus russischen Medien 20
3.3 Übersicht verschiedener Quellen mit ihrer jeweiligen politischen Ausrichtung und der Anzahl der gesammelten Datensätze 20
3.4 Beispiel zur Struktur der gesammelten Daten 24
3.5 Beispiele für automatische Übersetzung und die tatsächliche Bedeutung der Worte 29

4.1 Liste der Repositorien (Unterbibliotheken) des Natasha-Projekts [8] . . . 31
4.2 Auflistung der Bibliotheken zur Tokenisierung und deren Eigenschaften . 32

5.1 Aufzählung der im Jahr 2022 von jedem Medium verwendeten Wörter . 39
5.2 Anzahl der Verneinungen in den Nachrichtentexten 39
5.3 Anzahl der Publikationen vor und am Tag des Kriegsbeginns 40
5.4 Veränderung der durchschnittlichen Anzahl der veröffentlichten Publikationen vor und nach Kriegsbeginn 40
5.5 Zusammenfassung der zehn am häufigsten wiederholten Wörter in jedem Medium 43
5.6 Zusammenfassung der zehn am häufigsten wiederholten Wörter in jedem Medium (überarbeitet) 44

Abkürzungen

API Application Programming Interface

BERT Bidirectional Encoder Representations from Transformers

CFG Context-free grammar

CRF Conditional Random Field

GPT Generative Pre-trained Transformer

HMM Hidden Markov Model

MLM Masked Language Modeling

NER Named-entity recognition

NLP Natural language processing

NSP Next Sentence Prediction

1 Einleitung

1.1 Motivation

In den letzten Jahren hat sich die Politik zwischen Russland und der Ukraine zunehmend distanziert und verschärft. Der Beginn des Krieges in der Ukraine im Jahr 2022 verstärkt den Fokus auf die Russische Aktivitäten auf nationaler und internationaler Ebene. Die nationalen Medien haben bedeutenden Einfluss auf die gesamte Meinung der russischen Bevölkerung. Die wesentliche Bedeutung von Medien in Bezug auf aktuelle politische Änderung darf nicht unterschätzt werden.

Infolge des Kriegsbeginns wurden unabhängige Medien und Journalisten in Russland unter erhöhtem Druck gesetzt. Viele von denen wurden als Propaganda bezeichnet, sollten das Land verlassen oder ihre Tätigkeit komplett einstellen. Das hat zu einer Reduzierung der oppositionellen Nachrichtenverlage und dadurch zu einer Verringerung der Meinungsvielfältigkeit geführt. Angesichts dauerhafter repressiver Maßnahmen und der Einführung neuer Gesetzgebungen in Russland hat sich die Medienvielfältigkeit deutlich transformiert.

Die Motivation dieser Abschlussarbeit ergibt sich unter anderem aus zwei Hauptaspekten:

1. Ein tieferes Verständnis des Medieneinflusses in der Gesellschaft vermittelt die Erkenntnis der Mechanismen von der Meinungsbildung und der Informationsvermittlung.
2. Durch eine Analyse von Verhaltensänderungen der Medien während der politischen Krisenzeit können Erkenntnisse darüber vermittelt werden, wie Medieninstitutionen sich an neue globalen Änderungen anpassen und welche Strategien sie verfolgen, um ihre Berichterstattung fortzusetzen.

1.2 Aufgabenstellung

Das Hauptziel dieser Bachelorarbeit besteht darin, das Verhaltensmuster von pro-kremlischen und oppositionellen Medien im Jahr 2022 im Kontext des Kriegsausbruchs in der Ukraine zu analysieren. Zu diesem Zweck werden Daten von jeweils zwei pro-kremlischen

und zwei oppositionellen russischen Medienquellen gesammelt. Unter dem Einsatz von Techniken der natürlichen Sprachverarbeitung (NLP) wird eine umfassende Datenanalyse vorgenommen. Die Analyse erfolgt in drei Phasen, wie Voranalyse der Eingangsdaten, die Analyse des Kontextes ausgewählter Wörter und die Untersuchung der Zusammenhänge zwischen den Entitäten. Die resultierenden Erkenntnisse werden anschließend visualisiert, um eine fundierte Bewertung und Interpretation zu ermöglichen.

1.3 Aufbau der Bachelorarbeit

Die Bachelorarbeit beschreibt eine grundlegende Analyse der Nachrichtentexten aus den russischen Medien für das Jahr 2022.

Die „Einleitung“ bietet einen kurzen Überblick über die Motivation dieser Arbeit und stellt die Ziele dieser Studie vor.

Das Kapitel 2 „Theoretische Grundlagen“ setzt sich zum Ziel, die theoretischen Grundlagen der NLP zu erläutern. Dabei wird der klassische Ansatz sowie empirische und statistische Ansätze in der NLP erklärt. Am Ende dieses Kapitels wird ein kurzer Überblick über die wichtigsten Sprachmodelle und ihre Architekturen gegeben.

Das Kapitel 3 „Daten“ nimmt eine zentrale Stellung in dieser Arbeit ein. Es beschreibt den politischen Kontext von Russland, der für eine grundlegende Analyse unerlässlich ist. In diesem Teil wird zudem die Auswahl der Datenquellen und der Informationskanäle für die Untersuchung begründet. Weiterhin enthält das Kapitel ausführliche Erläuterungen zu den technischen Aspekten der Datensammlung aus den jeweiligen Informationskanälen und erläutert den Prozess der Datenbereinigung sowie deren Übersetzung auf Deutsch.

Das Kapitel 4 „Technischer Hintergrund“ vertieft die technischen Aspekte der durchgeführten Analyse. Hierbei werden die für die Implementierung ausgewählten Bibliotheken detailliert vorgestellt sowie die Entscheidungsgründe für spezifische Visualisierungstools erläutert. Zum Abschluss bietet das Kapitel eine umfassende Darstellung des Datenverarbeitungsprozesses. Dabei werden auch die eingesetzten Algorithmen zur Datenanalyse ausführlich besprochen.

Das abschließende Kapitel 5 „Experimente“ evaluiert die während des Analyseverlaufs erstellten Visualisierungen. Die gesamte Untersuchung gliedert sich in drei wesentliche Phasen: eine Voranalyse der Eingangsdaten, die Analyse des Kontextes ausgewählter Wörter und die Untersuchung der Zusammenhänge zwischen Entitäten. Innerhalb dieses Kapitels erfolgt eine tiefgehende Erläuterung zu jeder dieser Phasen.

Mit einer Erläuterung der Möglichkeiten für weitere Analysen schliesst das Kapitel 6 „Fazit“ die Bachelorarbeit ab.

2 Theoretische Grundlagen

Natural Language Processing (NLP), oder natürliche Sprachverarbeitung, bezieht sich auf den Einsatz von Algorithmen und Technologien, um menschliche Sprache zu verstehen, interpretieren und generieren, um dadurch maschinelles Lernen und künstliche Intelligenz in der Verarbeitung von Sprachdaten zu ermöglichen [29].

In diesem Kapitel wird ein Überblick über die zentralen Ansätze im Bereich NLP gegeben. Traditionelle regelbasierte Methoden dienen als Ausgangsbasis. Im weiteren Verlauf liegt der Fokus besonders auf den modernen Verfahren, wobei die Transformer-Architektur besondere Beachtung findet. Auf Basis dieser Architektur wurden Modelle wie Bidirectional Encoder Representations from Transformers (BERT) und Generative Pre-trained Transformer (GPT) entwickelt, die heute als sehr leistungsfähig gelten. Das allgemeine Verständnis dieser Modelle ist essentiell, um aktuelle Entwicklungen in der Sprachtechnologie nachzuvollziehen.

2.1 Klassischer Ansatz der NLP

Der klassische Ansatz in der NLP basiert auf festgelegte Regeln, die von Linguistikexperten entwickelt wurden. Der gesamte Prozess der klassischen NLP ist in in der Abbildung 2.1 dargestellt. Dieser fasst mehrere Phasen der Datenbearbeitung um, wie die Textvorverarbeitung, lexikalische Analyse, syntaktisches Parsing und semantische Analyse [29].

Jeder dieser Schritte stellt eigene Anforderungen und spezifische Herausforderungen dar, welche im nachfolgenden Abschnitt näher beleuchtet werden.

2.1.1 Textvorverarbeitung

In der Welt der NLP spielt die Textvorverarbeitung eine entscheidende Rolle. Sie bildet den ersten Schritt, bei dem Symbole, Worte und Sätze identifiziert werden. Diese erkannten Einheiten gelten als primäre Bausteine, die für folgenden Verarbeitungsschritte im NLP-Prozess vorgesehen sind.

Die Methodik der Textvorverarbeitung lässt sich grundsätzlich in zwei Hauptabschnitte unterteilen:

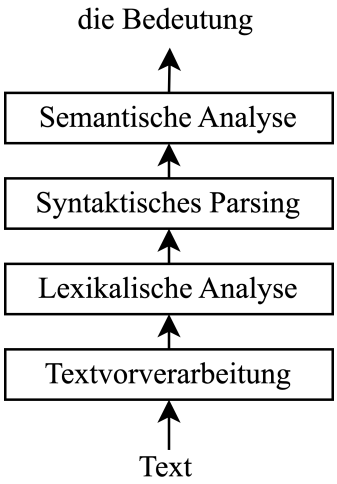


Abbildung 2.1: Die Phasen der Analyse bei der NLP [29]

1. Die Aufbereitung des Dokuments, worin rohe digitale Dateien in geordnete und wohldefinierte Textdokumente überführt werden.

2. Textsegmentierung (Satzsegmentierung und Tokenisierung) [29].

Innerhalb der Segmentierung gibt es zentrale Ansätze wie die „Satzsegmentierung“ und die „Tokenisierung“. Bei der Satzsegmentierung geht es primär darum, den Text in einzelne Sätze zu unterteilen. Hierbei wird es oft auf typische Indikatoren wie Satzzeichen zurückgegriffen [29]. Die Tokenisierung hingegen befasst sich mit der Unterteilung des Textes in individuelle Worte oder „Token“. Nach dieser Phase wird in der Regel eine Textnormalisierung durchgeführt. Dabei werden verschiedene Varianten eines Wortes, z.B. „Mr.“ im Vergleich zu „Mister“, zu einer standardisierten Form zusammengeführt [43].

Trotz dieser strukturierten Vorgehensweise ist die Segmentierung von natürlichen Sprachen mit erheblichen Herausforderungen verbunden. Diverse Sprachen weisen Mehrdeutigkeiten auf, insbesondere wenn es um die Trennung von Wörtern oder Sätzen geht [43]. Ein weiteres Beispiel sind Abkürzungen, die in vielen europäischen Sprachen sowohl als Wortgrenze als auch als Satzende interpretiert werden können, was den Segmentierungsprozess zusätzlich erschwert [29].

Hauptprobleme und Besonderheiten des Tokenisierungsprozesses können in der Verarbeitung von natürlicher Sprache wie folgt in mehrere Teile aufgespaltet werden [28]:

1. *Leerzeichen.* In Sprachen wie Chinesisch und Thailändisch werden Wörter ohne klare Grenzen geschrieben. Somit erfolgt die Tokenisierung anders, weil zusätzliche lexikalische und morphologische Informationen benötigt werden.

2. *Zeichensetzung.* Ein Tokenizer muss unterschiedliche Verwendungen von Satzzeichen, wie den Dezimalpunkt oder die Kennzeichnung von Abkürzungen, erkennen.

Einige Sprachen, wie die Thailändische, nutzen wenige oder keine Satzzeichen zur Satzgrenzen-Kennzeichnung.

3. *Mehrteilige Wörter.* Bindestriche in Sprachen, wie Französisch, schaffen grammatikalische Strukturen, wie „va-t-il“ oder „c'est-à-dire“. Tokenisierungsprobleme können auch durch Bindestriche am Zeilenende entstehen.

4. *Mehrwortausdrücke.* Ausdrücke, wie „in spite of“, können als ein Token betrachtet werden, genau wie „despite“. Der englische Ausdruck „no one“ kann je nach Kontext als ein oder zwei Tokens betrachtet werden [28].

In Situationen, in denen die korrekte Darstellung eines Tokens während der Tokenisierung unklar ist, sollten Entscheidungen hierüber den fortgeschrittenen Phasen der natürlichen Sprachverarbeitung überlassen werden [29].

2.1.2 Lexikasche Analyse

Im Rahmen der Textverarbeitung liegt der Hauptfokus auf der Satztrennung und der Bestimmung von Wortgrenzen, die die Tokenisierung umfasst. Der vorliegende Abschnitt erläutert die Analyse von Tokens [29].

Die morphologische Komplexität vieler Wörter macht die lexikalische Analyse zu einem wichtigen Schritt. Wörter können als individuelle Textelemente oder als Einheiten, die verwandte Formgruppen repräsentieren, betrachtet werden [16]. Die Umwandlung von Wortformen in ihre Basisform, auch Lemma genannt, wird als Lemmatisierung bezeichnet. Als Beispiel dient das Verb „deliver“ mit den Wortformen {delivers, deliver, delivering, delivered}. Hier ist „deliver“ das Lemma. Dieser Prozess verknüpft Wortformen mit relevanten semantischen und syntaktischen Informationen, die in bestimmten Anwendungen, wie in der maschinellen Übersetzung, zur Erhöhung der Präzision beitragen können [29].

Um morphologische Wortvarianten zu steuern, kommen endliche Automaten zum Einsatz [16]. Sie ermöglichen die Abbildung von Grundformen zu ihren korrespondierenden Oberflächenformen und umgekehrt. Dabei berücksichtigen sie die morphologische Komplexität des Wortes. Verschiedene Lemmaklassen und zugehörige Affixe spielen hierbei eine Rolle. Zum Beispiel, wandelt der Suffix „-lich“ oftmals ein Substantiv oder ein Adjektiv in ein Adverb oder ein neues Adjektiv um, wodurch sich die grammatische und oft auch die semantische Rolle des Wortes im Satz verändert. Endliche Automaten vereinfachen die Morphologieanalyse, was insbesondere für Sprachen mit einer komplexen Morphologie von Vorteil ist. Trotz einiger herausfordernder Ausnahmen in Sprachen erhöhen endliche Automaten die Genauigkeit der Analyse [18].

2.1.3 Syntaktisches Parsing

Das syntaktische Parsen stellt einen Prozessschritt in der Verarbeitung natürlicher Sprache dar, wodurch die Struktur von Sätzen analysiert und die Beziehungen zwischen ihren Komponenten erkannt werden können. Aufgrund der Mehrdeutigkeiten und langen Abhängigkeiten in natürlichen Sprachen stellt sich diese Aufgabe als besonders herausfordernd dar. Durch das Parsen wird es möglich, natürliche Sprache anhand von grammatikalischen Strukturen zu analysieren [29].

Context-free grammar (CFG), die im Jahr 1956 von Chomsky eingeführt wurden, bilden das Fundament in diesem Forschungsbereich. Sie setzen sich aus diversen Symbolen und Regeln zusammen und sind essenziell für zahlreiche Methoden des syntaktischen Parsings. Je nach spezifischem Algorithmus können Varianten der CFG angewendet werden. Dabei können Veränderungen innerhalb der Grammatik direkte Auswirkungen auf deren Struktur haben [59].

Dennoch ist die Leistung der CFG in der Analyse von natürlichen Sprachen begrenzt. Diese Tatsache führte zur Entwicklung fortschrittlicheren Grammatiken, wie Regulus, Generalized phrase structure grammar (GPSG), Combinatory categorial grammar (CCG) und Anderen. Die erweiterten Grammatiken ermöglichen eine größere Aussagekraft und Effizienz in der Analyse [58].

Ein gängiges Mittel zur Darstellung der syntaktischen Struktur eines Satzes ist der Syntaxbaum. Dieser veranschaulicht alle Ableitungsschritte des Satzes, beginnend mit Wurzelknoten. Als Beispiel lässt sich ein Baum für den Satz „the old man a ship“ in der Abbildung 2.2 betrachten [29]. In dieser Darstellung kann das Wort „man“ als Nomen (the old man) und als Verb (man a ship) interpretiert werden. Dies veranschaulicht, wie dieselben Wörter in unterschiedlichen syntaktischen Rollen fungieren können.

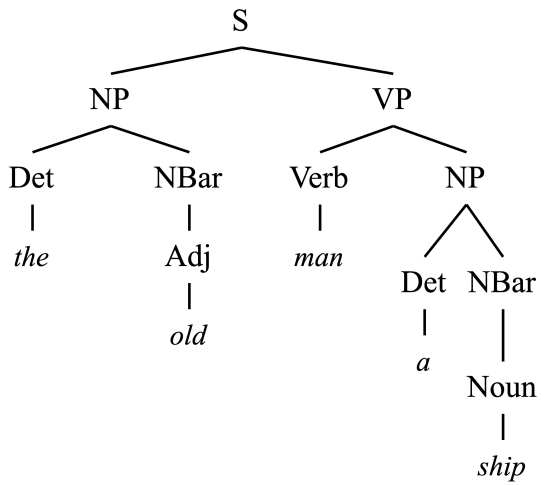


Abbildung 2.2: Syntaxbaum des Satzes „the old man a ship“ [29]

Im Bereich des Parsings wird grundsätzlich zwischen einem Erkenner und einem Parser unterschieden. Während der Erkenner die Grammatik eines Satzes feststellt, erzeugt der Parser strukturelle Darstellungen wie Parse-Bäume, die eine Interpretation des Textes ermöglichen. Die zugrundeliegende Grammatik definiert den Rahmen der möglichen Zustände. Wenn ihre Regeln erfolgreich angewendet werden können, gilt ein Satz als grammatikalisch. Dabei verfolgt jeder Parser einen eigenen Ansatz hinsichtlich der Regelanwendung, dem Umgang mit Mehrdeutigkeit und der Ausrichtung der Analyse [58].

Innerhalb der syntaktischen Analyse lassen sich drei Hauptprobleme hervorheben [59]:

1. *Stabilität*: Das System soll in der Lage sein, mangelhafte Daten zu verarbeiten, ohne nur Fehler zu melden. Es soll versuchen, aus nicht-regelkonformen Texten sinnvolle Informationen zu ziehen. Dabei können einige Methoden wie partielles Parsen oder die Verarbeitung von „lärmenden“ Daten helfen.

2. *Mehrdeutigkeitsauflösung*: Bei der Satzanalyse können verschiedene Interpretationen auftreten. Um diese Mehrdeutigkeit zu bewältigen, sind spezialisierte Grammatiken oder der Einsatz statistischer Ranking-Verfahren hilfreich.

3. *Effizienz*: Das Parsen mittels einer CFG kann sich als komplex und zeitaufwändig erweisen. Auch wenn es theoretische Einschätzungen zur Komplexität gibt, ist es essenziell, in der Praxis mit realen Daten zu arbeiten und dabei die Einzelheiten der jeweiligen Implementierung zu berücksichtigen [59].

2.1.4 Semantische Analyse

Das Verständnis sprachlicher Äußerungen ist in der Linguistik eine Herausforderung. Es hängt nicht nur von der lexikalischen Bedeutung der Wörter ab, sondern auch von den Faktoren wie Syntax und Kontext. Ein zentrales Ziel der semantischen Analyse, die sich mit der Bedeutung von Wörtern und Sätzen beschäftigt, liegt darin, die kognitiven Prozesse der Sprachnutzer abzubilden [19]. Oftmals erfordert die Analyse eine Übersetzung in eine Meta-Sprache. Die Herangehensweisen an diese Meta-Sprache können jedoch variieren. Viele Ansätze fokussieren sich auf die Festlegung von Wahrheitsbedingungen. Dennoch können sie nicht stets alle Nuancen der Sprache berücksichtigen, insbesondere in der Interaktion zwischen Mensch und Maschine [21].

Traditionell wird zwischen lexikalischer und superlexikalischer Semantik unterschieden. Während sich die lexikalische Semantik auf die Bedeutung einzelner Wörter konzentriert, betrachtet die superlexikalische Semantik größere Texteinheiten und deren übergeordnete Bedeutungsstrukturen. Im Laufe der semantischen Analyse zeigt sich jedoch, dass diese Grenzen zunehmend uneindeutig werden [19]. Der Grund dafür ist die enge Verbindung zwischen der Bedeutung eines Wortes und seiner grammatischen Struktur. Das hat zur Entstehung des Begriffs „Lexikogrammatik“ geführt. In solchen

Analysen stehen nicht nur die Bedeutungen der Wörter im Fokus. Die grammatischen Konstruktionen spielen dabei auch eine entscheidende Rolle [29].

Eine der zentralen Herausforderungen und Aufgaben in der semantischen Analyse ist die Vermeidung von zirkulären Definitionen, bei denen ein Konzept durch ein anderes und umgekehrt erklärt wird. Ebenso entscheidend ist die Auswahl zugrundeliegender, unabhängiger Elemente für eine detaillierte Untersuchung. Diese Themen werden in der Wissenschaft intensiv diskutiert, insbesondere in Bezug auf die Tiefe und Genauigkeit der Analyse [21].

Neben den rein linguistischen Aspekten ergibt sich auch die Frage nach dem Verhältnis zwischen dem sprachlichem Inhalt und dem enzyklopädischen Wissen. Im täglichen Leben ist die Sprachverwendung eng mit kulturellen Normen und dem Allgemeinwissen verknüpft. All dies erschwert zusätzlich die Analyse. Forschungen in diesem Bereich ziehen oft Interpretationen tatsächlicher Sprachnutzer heran [21].

Abschließend sollte im Kontext von NLP die Herausforderung der Mehrdeutigkeit von Wörtern erwähnt werden. Solche Phänomene wie Homonymie, wo dieselbe Wortform verschiedene Bedeutungen aufweist, und Polysemie, bei der ein Wort mehrere Bedeutungen hat, sind von besonderem Interesse. Die Polsemie, mit ihren vielen Bedeutungsfacetten, kann erhebliche Schwierigkeiten bereiten. Zudem erhöht die figurative Sprache die Komplexität bei der Interpretation [19].

2.2 Empirische und statistische Ansätze

Der im vorherigen Abschnitt besprochenen klassischen Ansatz hat seine Grenzen, insbesondere in Bezug auf Skalierung und Anpassung. Das Erstellen und Aktualisieren von Regeln erfordert bei umfangreichen Datenmengen einen hohen Arbeitsaufwand. Zudem können klassische Methoden die Mehrdeutigkeit natürlicher Sprache oft nicht effektiv verarbeiten und sich nur schwer an neue Domänen oder Textstile anpassen [29].

Im Gegensatz zu vordefinierten Regeln verwenden empirische und statistische Ansätze in NLP das maschinelle Lernen und statistische Analysen, um Muster in großen Datenmengen zu erkennen. Sie stützen sich auf die Untersuchung umfangreicher Textdaten und entwickeln Modelle basierend auf diesen Erkenntnissen [49].

2.2.1 Empirische Ansätze

Der empirische Ansatz in NLP fußt auf den Einsatz realer Sprachdaten. Im Gegensatz zu Ansätzen, die sich lediglich auf theoretische Modelle oder intuitive Annahmen stützen, legt die empirische Methode besonderen Wert auf authentische Kommunikationsbeispiele. Sie fokussiert sich auf die Analyse der Sprache durch Texte, Dialoge und Gesprächsaufzeichnungen [49].

Das Ziel dieser Ansätze ist es, Modelle zu entwickeln, die echte sprachliche Strukturen und Muster präzise abbilden. Dabei basieren sie sich auf die empirische Daten. Diese Modelle können potenziell für eine Vielzahl von Aufgaben, wie dir Maschinenübersetzung, die automatische Textannotation und die Textzusammenfassung, verwendet werden [43].

Ein zentraler Schritt in empirischer Modellierung der Sprache besteht aus der Sammlung und der Annotation großer Sprachkorpora. Diese Korpora dienen zusätzlich als Trainings- und Bewertungsdaten für Modelle. Die auf dem empirischen Ansatz basierenden Modelle, weisen oft eine höhere Robustheit und Generalisierungsfähigkeit auf, weil sie an Daten trainiert werden, die eine breite Palette sprachlicher Phänomene abdecken [43].

Bei der Anwendung des empirischen Ansatzes ist es notwendig, unterschiedliche Methoden und Werkzeuge zur Verarbeitung großer Datenmengen zu nutzen, wobei manche Techniken aus der klassischen NLP zum Einsatz kommen. Der empirische Ansatz zeigt seine Stärken besonders in den Bereichen, in denen traditionelle, theoretische Modelle Schwierigkeiten haben, die komplexen Facetten der realen Sprache vollständig zu erfassen. Dies trifft beispielsweise auf Forschungsfelder wie die Soziolinguistik, Dialektologie oder die Untersuchung neuer sprachlicher Phänomene wie dem Internet-Slang zu [49].

Es ist hervorzuheben, dass empirische Methoden häufig in Verbindung mit anderen Techniken eingesetzt werden. Als Beispiele lassen sich die statistische, symbolische oder auf maschinellem Lernen basierende Ansätze zeigen [29].

2.2.2 Statistische Ansätze

Empirische und statistische Ansätze in der NLP stehen oft in enger Verbindung. Sie basieren jedoch auf unterschiedlichen grundlegenden Prinzipien [49]. Während der empirische Ansatz sich von strengen, aus linguistischen Theorien abgeleiteten Regeln befreit, nutzt der statistische Ansatz mathematische und statistische Modelle, um Sprachdaten zu interpretieren. Der statistische Ansatz versucht, durch die Analyse großer Datensätze, Muster und Trends in der Sprache zu identifizieren [34].

In der NLP stellen das Hidden Markov Model (HMM) einen signifikanten statistischen Ansatz dar. Dies wird insbesondere für Anwendungen in der morphologischen Analyse und der Spracherkennung herangezogen. Parallel dazu haben sich die Conditional Random Field (CRF) als effektive Instrumente für die Informationsgewinnung herauskristallisiert, vor allem in Bezug auf die Named-entity recognition (NER)-Aufgabe [29].

Im Rahmen des statistischen Ansatzes sind n-Gramm-Modelle besonders dafür bekannt, Wortsequenzfrequenzen in umfangreichen Textkorpora zu analysieren. Die aus solchen Analysen gewonnenen Daten werden insbesondere in Anwendungsbereichen wie Spracherkennung und Maschinenübersetzung verwendet [29]. Ein signifikantes Beispiel für

den Einsatz des statistischen Ansatzes in der NLP stellt die statistische maschinelle Übersetzung dar. Diese basiert auf der Auswertung von parallelen Textkorpora [34].

Der statistische Ansatz in der NLP zeigt sich durch seine Adaptivität gegenüber echten Sprachdaten. Das macht den Ansatz besonders wertvoll, weil traditionelle Sprachregeln schwer zu definieren sind oder eine umfangreiche Datenbasis vorhanden sein soll [34]. Doch wie alle Methoden hat auch dieser Ansatz seine Beschränkungen. Solche Modelle erfordern oft umfangreiche Datenmengen für das Training und können bei seltenen oder neuen Sprachphänomenen an ihre Grenzen stoßen. Außerdem können ihre Ergebnisse manchmal aus linguistischer Sicht als ungenau oder sogar untypisch erscheinen [43].

In moderner NLP ergänzen sich trotz ihrer unterschiedlichen Schwerpunkte empirische und statistische Ansätze gegenseitig. Die empirische Analyse kann als Basis für die Sammlung und Interpretierung realer Daten fungieren, während statistische Verfahren diese Daten zur Modellierung und Vorhersage nutzen können [43]. Diese Zusammenarbeit führt oft zu präziseren und verlässlicheren Ergebnissen in diversen NLP-Anwendungen.

2.2.3 Named Entity Recognition

Nach der Betrachtung der Unterschiede und Besonderheiten des empirischen und statistischen Ansatzes in NLP wendet sich die Aufmerksamkeit dem spezifischen Bereich des Named Entity Recognition (NER) zu. In dieser Bachelorarbeit wird der Fokus auf NER-Systemen gelegt. Dabei geht es um die Identifikation und Klassifikation von Entitäten im Text wie beispielsweise Namen von Personen, Organisationen und Orten [32]. NER-Systeme sind ein wesentlicher Bestandteil vieler NLP-Ansätze und werden häufig unter Verwendung von Methoden des maschinellen Lernens und des Deep Learning entwickelt [29].

In diesem Kontext basiert der empirische Ansatz auf einer initialen Analyse realer Texte, um die Darstellung benannter Entitäten in der Sprache zu erfassen [49]. Zum Training und Testen von NER-Modellen dienen Textkorpora, welche Markierungen für benannte Entitäten enthalten [32].

In der Vergangenheit wurden statistische Methoden, wie HMM und CRF intensiv für NER-Aufgaben eingesetzt. Nachdem sie an annotierten Korpora trainiert wurden, erfolgte die Bewertung ihrer Effizienz anhand von Testdaten [43].

Obwohl klassische NER-Methoden ihre Berechtigung hatten, verzeichnen moderne Deep Learning-Ansätze besondere Erfolge in diesem Bereich. Hierzu zählen Modelle, die auf der Transformer-Architektur aufbauen, wie beispielsweise BERT. Diese fortgeschritten Modelle werden mithilfe empirischer Daten trainiert. Wie es sich anhand der Ergebnisse ersichtlich macht, erreichen sie häufig bessere Genauigkeiten und sind zuverlässiger als traditionelle statistische Ansätze [46].

Zusammenfassend lässt sich sagen, dass die NER-Aufgabe im Bereich des NLP mithilfe verschiedener methodologischer Ansätze adressiert werden kann. Welcher Ansatz und welches Modell gewählt wird, hängt von den spezifischen Anforderungen der Studie, der vorhandenen Datenmenge und den angestrebten Zielen ab [29].

2.3 Sprachmodelle

Auf Grundlage von klassischen, empirischen und statistischen Ansätzen in NLP wurden in den letzten Jahren bedeutende Fortschritte im Bereich Deep Learning und neuronaler Netzwerke erreicht. Diese technologischen Entwicklungen führten zur Entstehung von Sprachmodellen, die in der Lage sind, Texte mit bisher unerreichter Präzision und Flexibilität zu verarbeiten. Modelle wie BERT und GPT stellen in diesem Zusammenhang den Beginn einer neuen Ära der NLP dar. Sie verwenden innovative Techniken zur Verarbeitung von Kontexten und umfangreichen Datenmengen [46]. Im nachfolgenden Abschnitt wird es detailliert auf die Architektur der Transformer eingegangen und die Funktionsweisen von BERT und GPT im Vergleich zu traditionellen Ansätzen diskutiert.

2.3.1 „Transformer“-Architektur

Das Transformer-Modell wurde im Jahr 2017 im Artikel „Attention is All You Need“ vorgestellt [60]. Es wurde dabei eine Erhöhung der Leistung in Aufgaben der natürlichen Sprachverarbeitung erzielt. Als Beispiel dazu lässt sich die Sprachübersetzung und Textzusammenfassung nennen. Die Kernidee des Transformers ist der sogenannte Aufmerksamkeitsmechanismus. Dieser erlaubt dem Modell, den Fokus während der Informationsverarbeitung gezielt auf relevante Bereiche der Eingabedaten zu legen.

In der Abbildung A.1 im Anhang ist die Funktionsweise der Transformer-Architektur grafisch dargestellt. Die grundlegende Transformer-Struktur besteht aus zwei Hauptkomponenten: einem Encoder und einem Decoder. Der Encoder bearbeitet die Eingabesequenz und repräsentiert diese als Vektor, während der Decoder zur Generierung der Ausgabe dient. Ein charakteristisches Merkmal des Transformers ist, dass er keine eigene Sequenzauffassung besitzt. Daher wird den Wortvektoren eine Positionskodierung hinzugefügt, um sicherzustellen, dass das Modell die Stellung jedes Wortes innerhalb der Sequenz berücksichtigt [60].

Der nächste Schritt umfasst die Multi-Head Attention. In diesem Verfahren berücksichtigt jedes Wort alle anderen Wörter im Satz, um deren relative Bedeutung zu bestimmen. Dabei kommen drei verschiedene Vektoren zum Einsatz: Anfragen (queries), Schlüssel (keys) und Werte (values). Um diese Vektoren zu erzeugen, werden die ursprünglichen Wortvektoren mit spezifischen Gewichtsmatrizen multipliziert. Der eigentlichen Aufmerksamkeitsmechanismus basiert auf dem Skalarprodukt von Anfrage- und Schlüssel-

vektoren. Dieses Produkt wird skaliert und anschließend einer Softmax-Funktion zugeführt, wodurch die Aufmerksamkeitsgewichte entstehen. Durch Multiplikation dieser Gewichte mit den Wertvektoren ergibt sich eine gewichtete Darstellung, die den Kontext jedes Wortes abbildet. Da dieser Vorgang parallel in verschiedenen „Köpfen“ erfolgt (daher „Multi-Head“), können unterschiedliche Beziehungsaspekte zwischen den Wörtern erkannt werden [60].

Nachdem die gewichteten Darstellungen aus allen Attention-Köpfen kombiniert wurden, werden sie durch ein Feed-Forward-Netzwerk geleitet. Nach jedem Hauptblock (Aufmerksamkeit und Feed-Forward-Netzwerk) werden Restverbindungen (residual connections) und Schichtnormalisierung (layer normalization) angewendet [46].

Die oben genannten Schritte werden mehrmals wiederholt. Die Anzahl der Wiederholungen hängt von der spezifischen Konfiguration des Transformers ab. Am Ausgang der letzten Transformer-Schicht stehen die Daten für die weitere Verarbeitung oder Klassifikation bereit [46].

2.3.2 BERT

Der Bidirectional Encoder Representations from Transformers (BERT), ein von Google 2018 eingeführtes Modell, basiert auf die Transformer-Architektur. Sein bidirektionaler Aufmerksamkeitsmechanismus ermöglicht eine effektive Verarbeitung des Kontexts im Text [22]. Indem BERT Daten aus beiden Richtungen gleichzeitig verarbeitet, kann es den Kontext jedes Wortes genauer erfassen.

Da das Hauptziel von BERT besteht darin, ein Sprachmodell zu entwickeln, das lediglich den Encoder der Transformer-Architektur verwendet. Beim Trainieren von Sprachmodellen besteht häufig die Herausforderung, das richtige Vorhersageziel festzulegen. Viele Modelle versuchen das nächste Wort in einer Sequenz vorherzusagen. Jedoch wird dadurch das kontextbezogene Lernen eingeschränkt. Um dieses Problem zu lösen, verwendet BERT zwei Trainingstrategien: Masked Language Modeling (MLM) und Next Sentence Prediction (NSP) [22].

Beim MLM werden vor der Verarbeitung durch BERT 15% der Wörter in einer Sequenz durch Token [MASK] ersetzt. Das Ziel des Modells ist es dann, den ursprünglichen Wert dieser maskierten Wörter basierend auf dem Kontext der anderen, nicht-maskierten Wörter, korrekt vorherzusagen [46].

Beim NSP werden BERT zwei Sätze präsentiert. Das Modell muss dann bestimmen, ob der zweite Satz logisch auf den ersten folgt. Während des Trainings besteht eine Hälfte der Beispiele aus konsekutiven Sätzen des Originaltexts, also positiven Beispielen. Die andere Hälfte setzt sich hingegen aus zufällig gewählten, nicht zusammenhängenden Sätzen zusammen, den negativen Beispielen [46].

BERT hat die NLP-Welt durch seine Fähigkeit revolutioniert, den Kontext eines Wortes aus beiden Richtungen zu erfassen und somit ein tiefgreifendes Verständnis für die Semantik zu entwickeln. Das Modell lässt sich vielseitig an verschiedene NLP-Aufgaben anpassen, von Textklassifikation bis NER, und übertrifft dabei oft die Effizienz klassischer Methoden. Da BERT bereits auf umfangreichen Datenmengen vortrainiert wurde, entfällt der Bedarf an manueller Annotation und Feature-Erstellung. Das macht es zu einer flexiblen Lösung für unterschiedliche NLP-Aufgaben [46].

2.3.3 GPT

Das andere wichtige Sprachmodell, der Generative Pre-trained Transformer (GPT) wurde im Jahr 2018 von OpenAI vorgestellt. Ihm liegt ebenfalls die Transformer-Architektur zugrunde. Im Vergleich zu BERT setzt GPT jedoch einen unidirektionalen Aufmerksamkeitsmechanismus ein, um einen Text zu generieren [48].

Da GPT für die Textgenerierung konzipiert wurde, benötigt er nur den Decoder aus der Transformer-Architektur. Im Unterschied zu BERT-Modell, das fehlende Wörter vorhersagt oder den Kontext zwischen zwei Sätzen bestimmt, setzt GPT auf eine Sprachmodellierungsaufgabe. Das Hauptziel dabei ist es, basierend auf dem gegebenen Textkontext das darauffolgende Wort vorherzusagen. Dabei analysiert GPT den Text sequenziell von links nach rechts [46].

Während des Trainingsprozesses verarbeitet GPT eine gegebene Wortsequenz und prognostiziert das darauf folgende Wort unter Berücksichtigung des Kontextes der bisherigen Wortfolge. Durch diese Eigenschaft ist das Modell besondes für Anwendungen in der Textgenerierung geeignet, da es kontinuierlich Informationen aus dem gesamten vorhergehenden Text nutzt, um kohärente und kontextbezogene Antworten zu generieren [46].

GPT stellt durch seinen Ansatz zur unidirektionalen Modellierung sowie durch das Training an umfangreichen, nicht annotierten Datenmengen einen signifikanten Fortschritt in der Textgenerierung und in weiteren NLP-Aufgaben dar. Dieses Modell findet Anwendung in einer Vielzahl von Bereichen, von Übersetzungen und Fragenbeantwortungssystemen bis hin zur Erzeugung von Originalinhalten. Es zeichnet sich durch eine bemerkenswerte Performance im Vergleich zu herkömmlichen Methoden aus. Ähnlich wie BERT verringert GPT den Bedarf an manueller Datenannotation und stellt somit ein effizientes Werkzeug für die Verarbeitung natürlicher Sprache dar [46].

2.4 Diskussion

Am Ende dieses Kapitels ist es hervorzuheben, dass die natürliche Sprachverarbeitung mehrere Entwicklungsphasen durchlaufen hat. Das Spektrum der Ansätze variiert von

klassischen bis hin zu statistischen und empirischen Methoden. Diese Entwicklungen legten den Grundstein für den Aufbau komplexer Sprachmodelle wie BERT und GPT, die sich auf die Transformer-Architektur stützen. Obwohl jeder Ansatz seine spezifischen Vorteile und Grenzen besitzt, zeichnen sich moderne Modelle durch ihre bemerkenswerte Anpassungsfähigkeit und Effizienz in der Lösung von NLP-Aufgaben aus. Insbesondere hat die Transformer-Architektur das Verständnis und die Nutzung von kontextuellen Informationen im Text revolutioniert, was zu wesentlichen Fortschritten in diversen Anwendungen beigetragen hat.

Dabei ist die Unterscheidung zwischen den verschiedenen Ansätzen nicht immer eindeutig. Häufig werden zur Erreichung des Ziels Methoden aus verschiedenen Paradigmen kombiniert. Der gegenwertige Stand der NLP spiegelt somit eine fortwährende Entwicklung von Techniken wider, wobei jeder Ansatz zum Wachstum dieses spannenden und bedeutsamen Feldes beiträgt.

In der vorliegenden Bachelorarbeit wird kein eigenes Sprachmodell entwickelt. Stattdessen liegt der Schwerpunkt auf der Untersuchung klassischer und statistischer Methoden zur Analyse des Informationsflusses in den russischen Medien. Die weitere Vertiefung in der Aufgabestellung wird in den folgenden Kapiteln erläutert.

3 Daten

Nach der theoretischen Erläuterung im vorherigen Kapitel wird der Fokus auf die praktische Seite der Studie gelegt, insbesondere auf die Bedeutung der Daten. In dieser Arbeit spielen die Daten nicht nur die Rolle eines Werkzeugs, sondern nehmen eine zentrale Stelle im gesamten Forschungsprozess ein. Ein umfassender Ansatz zur Datensammlung erfordert ein tiefes Verständnis für die Motivation bei der Auswahl bestimmter Informationsquellen, die Analyse des politischen Bildes von Russland während des Untersuchungszeitraums, das Verständnis der Besonderheiten der Interaktion mit der Application Programming Interface (API) von dominanten sozialen Plattformen und, ebenso wichtig, die kompetente Anwendung von Methoden zur Datenbereinigung. Es sollte auch die Bedeutung der effektiven Nutzung von NER für eine tiefe Textanalyse hervorgehoben werden. In diesem Kapitel wird eine detaillierte Darstellung aller erwähnten Schritte und Besonderheiten der durchgeführten Studie erläutert.

3.1 Politischer Kontext

Eine umfassensde Analyse der politischen Situation von Russland und der in letzter Zeit verabschiedeten repressiven Gesetzgebung ist sowohl für dieses Kapitel als auch für die gesamte Studie von Bedeutung. In diesem Abschnitt wird auf Gesetze eingegangen, die die Mediendarstellung im Jahr 2022 maßgeblich geprägt haben. Dies wird insbesondere dazu beitragen, die Unterschiede in den Ergebnissen der Experimente, welche im Kapitel 5 diskutiert werden, zwischen oppositionellen und pro-kremlischen Medien besser zu verstehen.

In autoritären Staaten stehen Massenmedien meist unter strenger staatlicher Aufsicht. Es wird von ihnen erwartet, dass sie der Linie des Regimes konsequent folgen [47]. Russland stellt hier keine Ausnahme dar. Angesichts der sich verschärften politischen Lage und der zunehmenden Zensurbestrebungen differenzieren sich die Medien in zwei Gruppen: einige unterstützen das Regime und erhalten staatliche Förderung, während andere Kritik am gegenwärtigen Regime üben [25].

Im Jahr 2022 standen oppositionelle Medien in Russland vor großen Herausforderungen durch staatliche Repressionen. Um ihre Arbeit fortsetzen zu können, mussten sie sich nicht nur an die neue zensurbehaftete Gesetzgebung anpassen, sondern auch ihrer journalistischen Integrität treu bleiben. Im Gegenzug waren Medien, die Putin unterstützen,

ebenso strikten Zensurrichtlinien unterworfen. Deswegen wird eine Beschreibung eines Ereignisses oft in völlig unterschiedlichen Kontext dargestellt.

Es ist auch wichtig auf die Gefahren für Journalisten einzugehen, insbesondere für diejenigen, die über die Ukraine berichten. Im Oktober 2022 zeigten zwei Journalistinnen, in München und in Tiflis (Georgien) Anzeichen einer Vergiftung auf. Es wird angenommen, dass die russischen Geheimdienste darin beteiligt waren. Diese und andere Fälle werden von internationalen Forschungsgruppen wie „The Insider“ und „Bellingcat“ untersucht [66]. Dies zeigt, dass russische Journalisten, selbst wenn sie im Ausland tätig sind, erheblichen Risiken ausgesetzt sind.

3.1.1 Relevante Begriffe

Angesichts der gegenwärtig schwierigen politischen Situation in Russland sind bestimmte Begriffe und Konzepte von besonderer Relevanz, da sie die Medienarbeit beeinflussen. Zu solchen gehört das Konzept des „ausländischen Agenten“. Dieses hat in der russischen Politik während der letzten Jahre zunehmend an Bedeutung gewonnen.

Ein Medium kann den Status eines ausländischen Agenten zugewiesen bekommen, wenn es „gedruckte, audio-, audiovisuelle und andere Nachrichten und Materialien verbreitet, die für eine unbegrenzte Anzahl von Personen vorgesehen sind“ [2]. Hierzu zählen insbesondere Veröffentlichungen im Internet, wie sie in Online-Medien oder sozialen Netzwerken zu finden sind. Hierbei ist nicht nur die ursprüngliche Veröffentlichung relevant, sondern auch das Teilen und Weiterleiten solcher Inhalte. Eine weitere Voraussetzung für diese Einstufung ist der Erhalt der Finanzierung aus dem Ausland. Bemerkenswert dabei ist, dass der genaue Zweck dieser Finanzierung nicht definiert wird. Sogar eine freiwillige Spende von einem Freund oder Familienmitglied könnte hierunter fallen. Die Höhe des Betrags ist dabei unerheblich [35].

Auch Einzelpersonen können in Russland den Status eines ausländischen Agenten erhalten. Voraussetzung dafür sind politische Aktivitäten oder eine gezielte Informationsbeschaffung, die potenziell gegen die Sicherheitsinteressen von Russland wirken könnte. Gleich wie bei Medien entscheidet auch bei Einzelpersonen der Empfang finanzieller Unterstützung aus dem Ausland über die Einstufung [35].

Mit der Einstufung als ausländischer Agent sind in Russland verschiedene Pflichten und Beschränkungen verbunden. Solche Medien, die diesen Status erhalten, sind dazu verpflichtet, innerhalb eines Monats nach ihrer Einstufung eine juristische Einheit zu gründen. Zudem müssen sie alle drei Monate einen Bericht über ihre Aktivitäten und Finanzen einreichen und sich jährlich einer Pflichtprüfung unterziehen. Eine zusätzliche Verpflichtung besteht darin, jeglichen veröffentlichten Inhalt mit einem Hinweis auf ihren Status zu versehen. Das betrifft alle Artikeln und Nachrichen sowohl in der eigenen Website, als auch in sozialen Netzwerken. Ein Beispiel für die Kennzeichnung ist in der Tabelle 3.1 zu finden. Diese Kennzeichnung muss in einer Schriftgröße verfasst sein,

die doppelt so groß ist wie der Haupttext und muss jede Veröffentlichung begleiten, unabhängig von ihrer Länge. Bei Nichtbeachtung dieser Vorschriften drohen Geldbußen, bei wiederholten Verstößen kann eine Strafverfolgung eingeleitet werden.

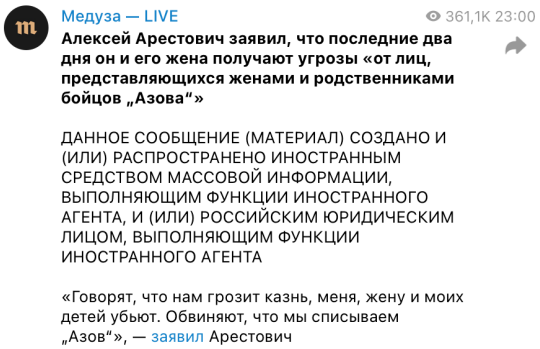
Original der Nachricht mit Kennzeichnung	Übersetzung
	Meduza - LIVE Alexej Arestowitsch erklärte, dass er und seine Frau in den letzten beiden Tagen Drohungen „von Personen, die sich als Frauen und Verwandte der Kämpfer von 'Azov'“ erhielten. DIESE NACHRICHT (MATERIAL) WURDE ERSTELLT UND (ODER) VERBREITET VON EINEM AUSLÄNDISCHEN MASS-MEDIENUNTERNEHMEN, DAS DIE FUNKTIONEN EINES AUSLÄNDISCHEN AGENTEN AUSÜBT, UND (ODER) EINEM RUSSISCHEN RECHTLICHEN SUBJEKT, DAS DIE FUNKTIONEN EINES AUSLÄNDISCHEN AGENTEN AUSÜBT. „Sie sagen, dass uns die Hinrichtung droht, meine Frau und meine Kinder werden getötet. Sie beschuldigen uns, 'Azov' auszulöschen“, sagte Arestowitsch.

Tabelle 3.1: Beispiel der Kennzeichnung des ausländischen Agenten [11]

Der Status „ausländischer Agent“ beeinflusst maßgeblich die Arbeitsweise von Medien in Russland. Werbetreibende tendieren dazu, Beziehungen zu diesen Medien zu meiden, da eine Zusammenarbeit zu einem ähnlichen Status führen kann. Ein Großteil der unabhängigen russischen Medien finanziert sich über Werbung und Spenden. Das Fehlen von Werbetreibenden aufgrund dieses Labels kann daher ihre Finanzierung erheblich beeinträchtigen [51]. Während öffentliche Unterstützung oder Erwähnung eines ausländischen Agenten dazu führen kann, dass jemand ebenfalls diesen Status erhält, sind direkte Spenden an solche Medien rechtlich unbedenklich.

Die zwei anderen für diese Studie relevanten Begriffe sind der Status von „unerwünschten“ und „extremistischen“ Organisationen. Eine „unerwünschte Organisation“ wird als solche betrachtet, wenn ihre Aktivitäten die verfassungsmäßige Ordnung des russischen Regimes bedrohen. Der Status kann willkürlich zugewiesen werden. Einmal als unerwünscht eingestuft, dürfen diese Organisationen in Russland nicht mehr aktiv sein. Jegliche Form

der Zusammenarbeit birgt das Risiko von Strafen, die von Geldbußen bis hin zu Freiheitsstrafen reichen können [31].

Das Weiterverbreiten oder Verlinken von Inhalten eines als unerwünscht eingestuften Mediums kann weitreichende rechtliche Konsequenzen haben. Selbst ältere Erwähnungen, etwa vor einem Jahr, können als Anlass für rechtliche Schritte dienen. Handlungen, die als Unterstützung solcher Medien interpretiert werden könnten – wie das Abonnieren von Newslettern, das Kommentieren von Artikeln oder der Kauf von Merchandise – birgen Risiken [31]. Einige der als unerwünscht betrachteten Organisationen in Russland sind zum Beispiel der Deutsch-Russischer Austausch e.V., The Insider und der Ukrainian World Congress [6].

Der Status einer „extremistischen Organisation“ stellt in Russland eine noch schwerwiegendere Einstufung dar. Jegliche Zusammenarbeit mit solchen Organisationen ist verboten, und ihre Auflösung erfolgt automatisch durch gerichtliche Anordnung. Die Konsequenzen für Unterstützer oder Mitwirkende sind gravierend. Es reicht von Wahlverboten bis hin zu strengen Strafen für das Teilen von Materialien oder das Zeigen von Symbolen [31]. In Russland werden verschiedene Gruppen und Unternehmen als extremistische Organisationen eingestuft. Dazu gehört die Terrorgruppe Islamischer Staat. Seit dem 21. März 2022 zählt auch das Unternehmen Meta dazu, zu dem Facebook, Instagram, WhatsApp und Oculus gehören [7].

Der Artikel über Fälschungen und Diskreditierung, der kurz nach dem Kriegsanfang am 4. März 2022 als das sogenannte „Fake News-Gesetz“ eingeführt wurde, hat weitreichende Auswirkungen auf die Medienlandschaft in Russland. Dieses Gesetz ahndet das Verbreiten von „bewusst falschen Informationen über die Aktivitäten der russischen Streitkräfte“ sowie das „Diskreditieren des Einsatzes russischer Truppen“. Bei Zuwiderhandlungen drohen Geldstrafen oder Freiheitsstrafen von bis zu 15 Jahren [3].

Aufgrund des Gesetzes zogen viele russische Medienunternehmen ihre Berichterstattung über den Ukraine-Krieg zurück. Einige stellten ihre Arbeit ganz ein, andere entfernten relevante Artikel und vermieden künftige Berichte über militärische Aktionen. Diese Entscheidungen wurden oft mit der Befürchtung begründet, dass Journalisten strafrechtlich verfolgt werden könnten. Tatsächlich wurden mehr als 10 oppositionelle Medien in Russland wegen ihrer Kriegsberichterstattung blockiert [39].

Gemäß diesem Gesetz können selbst geringfügige Abweichungen von der offiziellen russischen Darstellung des Ukraine-Kriegs als Fälschungen angesehen werden. Das betrifft Kritik an russischen Militäraktionen, Social-Media-Posts über den Krieg oder das Teilen ausländischer Artikel. Bis Mitte April 2022 führte dies zu mehr als 20 Strafverfahren [41].

Als Diskreditierung werden „öffentliche Handlungen“ angesehen, die darauf abzielen, den Einsatz der russischen Armee herabzuwürdigen. Dies erstreckt sich auch auf Handlungen, die die Tätigkeiten russischer staatlicher Einrichtungen im Ausland in Frage

stellen. Ein Plakat mit dem Text „Nein zum Krieg“ kann in Russland als Diskreditierung interpretiert werden [41].

Die Risiken, denen sich Journalisten und Medien in Russland aussetzen, wenn sie gegen das derzeitige Regime beziehen, sind enorm. Sie arbeiten unter sehr eingeschränkten Bedingungen. Aufgrund der neuen und oft vage formulierten Gesetze haben sich viele Medien und Journalisten dazu entschieden, Russland zu verlassen. Die vagen Gesetzestexte lassen einen breiten Interpretationsspielraum zu und können je nach Wunsch der Behörden ausgelegt werden [41]. Solch eine gesetzliche Unbestimmtheit erzeugt ein Klima des ständigen Drucks auf die Medien. Jede Handlung kann, je nach den aktuellen Interessen der Regierung, mit einem Gesetz verknüpft und geahndet werden. In diesem Kontext ist es von Bedeutung, die Praxis des „Neusprechs“ in den russischen Medien zu erwähnen. Darauf wird im Folgenden näher eingegangen.

3.1.2 Neusprech

Neben einer verstärkten Finanzierung von regierungsnahen Medien und der Unterdrückung unabhängiger Medien hat die Einführung von Gesetzen über Fake News, ausländische Agenten und Extremismus wesentlich dazu beigetragen, neue Denkweisen und Werte in Russland zu etablieren. Ein weiteres wichtiges Mittel in diesem Zusammenhang ist der weitverbreitete Einsatz von Neusprech [40].

Der Begriff „Neusprech“ stammt aus George Orwells dystopischem Roman „1984“. In diesem Werk bezeichnet er eine Sprachform, die so gestaltet ist, dass nur parteikonforme Gedanken geäußert werden können [45]. Seit dem Jahr 2020 nutzt die russische Opposition diesen Begriff, um die Sprachweise der staatlichen Medien in Russland zu beschreiben. Neusprech dient dazu, die Realität zu verzerren und mithilfe einer gezielten Wortwahl die Tragweite von Ereignissen zu verschleiern.

Ein entscheidender Moment für den Einsatz von Neusprech war die Coronavirus-Pandemie. Während der Pandemie, insbesondere während des Lockdowns, mussten die Behörden der Bevölkerung kritische Informationen liefern. Als die Kritik an Wladimir Putin im Zuge des Lockdowns stieg, verschwand das Wort „Lockdown“ aus der Berichterstattung. „Todesfälle“ wurden zu „letalen Ausgängen“, und die „Quarantäne“ wurde in „Selbstisolation“ umbenannt [40]. Solche Begriffe werden von der Bevölkerung als weniger alarmierend empfunden [27].

Mit Beginn des russischen Angriffs auf die Ukraine hat der Einsatz des Neusprechs in den staatlichen Medien zugenommen. Ein „Zusammenstoß“ wurde zur „Annäherung“ und eine „Besetzung“ zur „Befreiung“. Weitere Beispiele für Neusprech sind in der Tabelle 3.2 aufgelistet.

Tatsächliche Bedeutung	Neusprech
Explosion	Klatschen
Der Fall der Wirtschaft	Negatives Wachstum der Wirtschaft
Flugzeugabsturz	Harte Landung
Tod im Krieg	Unbekannte Abwesenheit in der Militäreinheit

Tabelle 3.2: Beispiele von Neusprech aus russischen Medien

Ein prägnantes Beispiel für Neusprech ist der Begriff „Spezialoperation“ als Ersatz für das Wort „Krieg“. Statt des Wortes „Krieg“ nutzen die Medien den Ausdruck „Spezialoperation“ und seine Varianten, etwa „spezielle Kriegsoperation“, „Spezialmilitäroperation“ oder die Kurzform „SWO“. Seit Beginn des Krieges in der Ukraine wurde der Ausdruck „Russland hat einen Krieg begonnen“ tabuisiert und fällt unter das zuvor erläuterte „Fake News-Gesetz“.

Die allgegenwärtige Neusprech-Praxis in den Medien beschränkt sich nicht nur auf den simplen Austausch für die Regierung unpassender Begriffe. Es wird systematisch versucht, alternative Denkmuster auszulöschen. Neusprech verursacht bestimmte menschlichen Emotionen und arbeitet mit Dualismen wie „Freund vs. Feind“. In solchen verfälschten Texten ist es eine Herausforderung, die wahre Information zu entdecken. Leser müssen nicht nur die Fähigkeit, sondern auch den Willen haben, die Information richtig zu interpretieren. Das Fake News-Gesetz scheint solche Interpretationen sogar rechtlich zu vertreten [40] [27].

3.2 Auswahl der Datenquellen

Im aktuellen politischen Umfeld erscheint es als eine besonders wichtige Aufgabe, die Unterschiede zwischen den Medien der russischen Regime und den Medien mit den Kritischen Aussagen und Meinungen zu untersuchen. Im Rahmen dieser Studie wurden Artikel von vier verschiedenen Medien analysiert: zwei davon unterstützen die russische Politik und die restlichen zwei stellen sich dagegen. Eine kurze Beschreibung dieser Quellen ist in der Tabelle 3.3 zusammengefasst.

Quellenname	Politische Ausrichtung	Anzahl der Datensätze
Die Nachrichten des Ersten Kanals	Pro-kremlisch	8486
Ria News	Pro-kremlisch	48601
Meduza	Oppositionell	24578
Mediazona	Oppositionell	3076

Tabelle 3.3: Übersicht verschiedener Quellen mit ihrer jeweiligen politischen Ausrichtung und der Anzahl der gesammelten Datensätze

Aus dem russischen Medienumfeld wurden die folgenden Quellen ausgewählt: „Nachrichten des Ersten Kanals“ (im Folgenden, der Erste Kanal), „Ria News“ (im Folgenden, Ria), „Meduza“ und „Mediazona“. Die ersten beiden Medien tendieren zur Kreml-Fraktion, während die letzten beiden als oppositionell gelten.

Der Erste Kanal steht unter staatlicher Kontrolle und präsentiert sich als der führende Sender des Landes [30]. Die Nachrichtenagentur Ria ist Teil des Medienkonzerns „Russia Today“, welcher staatlich finanziert wird [23]. Seit dem 25. Februar 2023 steht „Russia Today“ auf der Sanktionsliste sämtlicher EU-Länder [24].

Meduza ist ein unabhängiges Nachrichtenportal in russischer und englischer Sprache, gegründet von russischen Journalisten, mit Sitz in Riga, Lettland. Mediazona ist ein unabhängiges russisches Medienunternehmen, das den Schwerpunkt auf Opposition gegen Wladimir Putin legt und von zwei Mitgliedern der Protestgruppe Pussy Riot ins Leben gerufen wurde.

Sowohl Meduza als auch Mediazona werden im Kontext der russischen Politik als oppositionelle Medien betrachtet. Im Jahr 2022 erlebten sie Repressionen nach dem russischen Angriff in der Ukraine. Im März desselben Jahres wurden beide Medien wegen ihrer Berichterstattung über den Konflikt in Russland blockiert [38]. Weiterhin wurden beide Medien im Jahr 2021 als ausländische Agenten eingestuft [36][17]. Im Januar 2023 erhielt Meduza in Russland den Status einer unerwünschten Organisation.

Die Entscheidung, welche Medien in dieser Abschlussarbeit analysiert werden, bestand aus mehreren Kriterien. Ursprünglich wurde der unabhängige russische Fernsehkanal „RainTV“ als Informationsquelle gewählt. Dieser Kanal war in den vergangenen Jahren vermehrt staatlichen Repressionen ausgesetzt. Diese führten dazu, dass RainTV seine Tätigkeit vom 4. März 2022 bis zum 17. Juli 2022 zwecks Neustrukturierung pausieren musste [33]. Ein solch ausgedehnter Zeitraum ohne verfügbare Daten würde das Ergebnis dieser Analyse erheblich beeinträchtigen. Daher fiel die Wahl schließlich auf Meduza, die im gesamten Jahr 2022 durchgehend aktiv blieben.

3.3 Informationskanäle

Im Rahmen dieser Abschlussarbeit bezieht sich die Datenanalyse auf Medien, die auf unterschiedlichen Plattformen wie Webseiten, VKontakte, Telegram und Youtube veröffentlicht sind. Aufgrund ihrer Dominanz in Bezug auf die Nutzerzahlen in Russland wurden VKontakte und Telegram als Hauptkanäle für diese Untersuchung ausgewählt. Beide Plattformen stammen vom russischen Unternehmer Pawel Durow.

3.3.1 VKontakte

Das soziale Netzwerk VKontakte, oft VK abgekürzt, ist in Russland vergleichbar mit Facebook. Trotz der Verfügbarkeit in über 90 Sprachen dominiert die Nutzung durch russischsprachige Benutzer. Bis zum dritten Quartal 2022 erreichte VK eine Durchdringungsrate von über 75% in Russland (siehe Abbildung 3.1) [52]. Die Durchdringungsrate beschreibt das Verhältnis zwischen der Zahl der Nutzer einer sozialen Netzwerkplattform und der Gesamtzahl der Personen in der Zielgruppe. Nach der Blockierung von Facebook und Instagram in Russland stieg die Nutzerzahl von VKontakte bis November 2022 sogar auf 85% [53]. Dieser Anstieg ist auf die Einstufung des Meta-Unternehmens als extremistisch zurückzuführen [4][1], welches am 11.10.2022 auf die Liste von Organisationen mit Verbindungen zu Terrorismus und Extremismus gesetzt wurde [44].

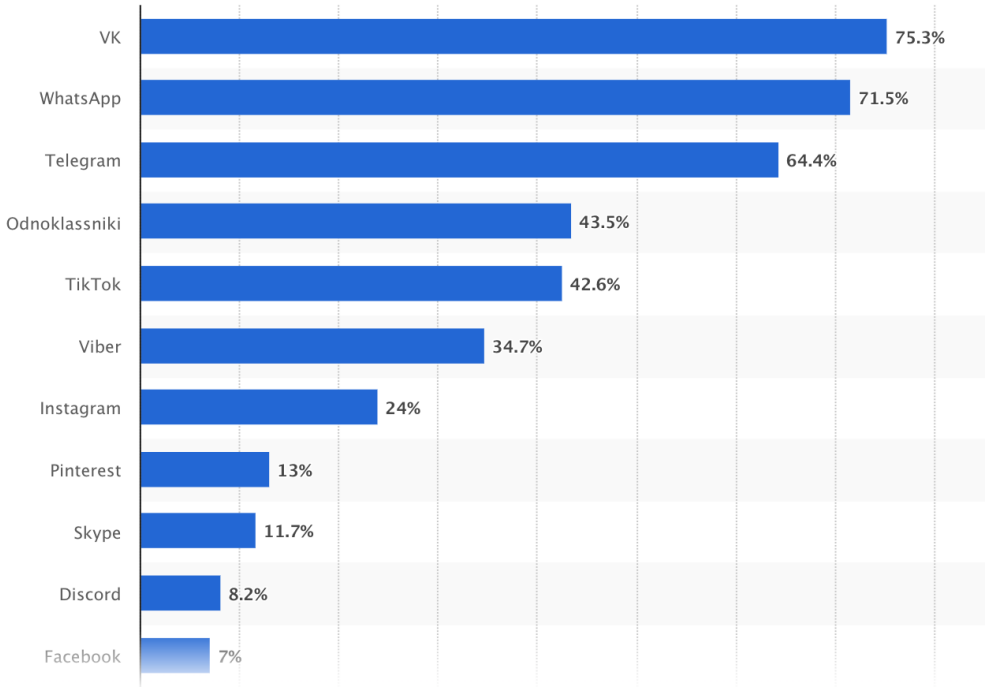


Abbildung 3.1: Führende soziale Medienplattformen in Russland im 3. Quartal 2022 nach monatlicher Durchdringungsrate [52]

Pawel Durow, der Gründer von VKontakte, sah sich im Jahr 2014 infolge der Annexion der Krim durch Russland gezwungen, seine Unternehmensanteile abzugeben und das Land zu verlassen. Durows Entscheidung stand in Zusammenhang mit seiner Weigerung, mit russischen Behörden zu kooperieren und beispielsweise die Seite des Oppositionsführers Alexej Nawalny zu schließen oder Daten von „Euromaidan“-Aktivisten (Protest im Zentrum von Kiew im Jahr 2013) herauszugeben [26].

Am Ende 2021 wurde VKontakte einem anderen Besitzer übergeben, der dem engsten Kreis von Präsident Putin angehört. Das hatte eine stärkere Kreml-Einflussnahme auf das Netzwerk zur Folge [50]. Von 2015 bis 2019 wurden aufgrund von Inhalten auf

VKontakte 500 Strafverfahren eingeleitet, wobei einige Verurteilungen Haftstrafen von bis zu fünf Jahren nach sich zogen [62].

Für diese Untersuchung wurden hauptsächlich Materialien des Ersten Kanals von VKontakte verwendet. Die VK-Seite des Ersten Kanals hat mehr als eine Million Abonnenten (Stand 22.09.2023) [13]. Es ist anzumerken, dass Ria informativere Beiträge auf ihrer Telegram-Seite publiziert, während Meduza im März 2022 auf VKontakte blockiert wurde [37]. Mediazona publiziert ebenfalls VKontakte-Nachrichten, beschränkt sich dort jedoch auf minimale Informationsvermittlung.

3.3.2 Telegram

Diese Studie legt einen besonderen Fokus auf die Telegram-Plattform. Obwohl Telegram ursprünglich als ein Messenger-Dienst ähnlich wie WhatsApp konzipiert war, hat es sich im Laufe der Zeit zu einem sozialen Netzwerk entwickelt, das Gruppen, Kanäle sowie Übertragungen und Konferenzen bietet [20].

In Bezug auf die Nutzerzahlen steht Telegram nach WhatsApp an zweiter Stelle. Im dritten Quartal 2022 betrug die Durchdringungsrate von Telegram in Russland über 64% [52]. Zum Vergleich: WhatsApp lag bei 71,5% (siehe Abbildung 3.1). Dennoch hat Telegram in Nutzervolumen im Jahr 2022 erstmals WhatsApp überholt und erreichte Anfang 2023 einen Anteil von 60-80% [61].

Mit Beginn des russischen Angriffs auf der Ukraine ist Telegram rasch zu einer wichtigen Informationsquelle über den Kriegsverlauf geworden. Während regierungsnahe Medien diese Plattform für Desinformation und Propaganda nutzten, suchten viele Russen dort nach unabhängigen Quellen. Besonders wurde es relevant, wenn unabhängige Medien im Land eingeschränkt wurden. Laut Time-Magazin stellte Telegram seit dem 24. Februar einen Anstieg russischer Abonnenten um 48% fest [57].

Aus diesem Grund wurden die offiziellen Medienkanäle auf Telegram als Hauptquelle für diese Analyse herangezogen, insbesondere die Kanäle von Ria [12], Meduza [11] und Mediazona [10]. Zum Stichtag 22. September 2023 hatten diese Kanäle Abonnentenzahlen von 2,9 Millionen (Ria), 1,2 Millionen (Meduza) bzw. 188.000 (Mediazona) identifiziert.

3.4 Datensammlung und Datenverarbeitung

3.4.1 Datensammlung

Mit der Auswahl der Medien und deren Vertretungskanäle besteht der nächste Schritt darin, wie die Daten effizient gesammelt werden können und welche von Relevanz sind.

Datenstruktur

Bei der Analyse der gesammelten Daten zeigte sich ihre Komplexität, da sie neben Texten auch Medien wie Fotos und Videos enthalten. Zur Vereinfachung der Analyse wurde eine klare Datenstruktur festgelegt, welche wie folgt aussieht: Medien-ID, Nachrichten-ID, Veröffentlichungsdatum und Nachrichtentext. Die entsprechenden Datentypen hierzu sind: Integer, Integer, Date und String, wie in der Tabelle 3.4 dargestellt.

Die Unterscheidung zwischen den verschiedenen Medien ist für die Analyse von zentraler Bedeutung. Die jeweiligen Medien und ihre einzelnen Artikel müssen klar identifiziert werden können. Das Veröffentlichungsdatum der Nachrichten ist ebenfalls von Bedeutung, denn in Abhängigkeit von aktuellen Ereignissen können sich sowohl die Anzahl der Artikel als auch das lexikalische Spektrum der Texte ändern.

Owner_id: Integer	Message_id: Integer	Date: Date	Text: String
-49388814	901750	2022-12-31	31 декабря 2022 г. в 18 ч. 30 мин. на Ростовской АЭС произошло возгорание блочного трансформатора энергоблока №3, находящегося в плановом ремонте... *(Am 31. Dezember 2022 um 18:30 Uhr kam es im Kernkraftwerk Rostow zu einem Brand des Blocktransformators von Block 3, der sich in der geplanten Reparatur befindet.)*

Tabelle 3.4: Beispiel zur Struktur der gesammelten Daten

Mit der festgelegten Datenstruktur kann die technische Extraktion der Daten aus den Medienkanälen erfolgen.

Extrahierung der Daten mittels VK API

Für die Datensammlung der Nachrichten des Ersten Kanals auf VKontakte wurde das VK API herangezogen. Das VK API stellt eine standardisierte Schnittstelle dar, die es erlaubt, mittels HTTP-Anfragen auf Daten von Gruppen der analysierenden Plattform zuzugreifen [65].

Zur Gewährleistung von Datenschutzrichtlinien und gezieltem Datenzugriff benötigt das VK API einen spezifischen Zugangsschlüssel, den sogenannten access_token. Dieser Schlüssel kann durch ein Benutzerkonto bei VKontakte generiert und abgerufen werden [63].

Um eine effiziente und systematische Datensammlung zu gewährleisten, wurde ein Client in der Programmiersprache Python entwickelt. Dieser Client ist so konzipiert, dass er gezielte HTTP-Anfragen an den VK-Server sendet, wobei er auf die Methode „wall.get“ zurückgreift. Diese Methode erlaubt das Abrufen von Beiträgen einer Community-Seite [64]. Die Parameter, die bei diesen Anfragen übermittelt werden, beinhalten unter anderem den Zugangsschlüssel (access_token), die ID der betreffenden Community (owner_id), die gewünschte Beitragsanzahl (count) sowie einen Offset (offset). Der Prozess ist in der Abbildung 3.2 dargestellt.

Die vom Server bereitgestellten Daten sind im JSON-Format strukturiert. Sie umfassen mehrere Elemente, und unter Anderem den Datentyp (z.B. Beitrag oder Story), die ID der Dateneinheit, die Anzahl der Kommentare und den eigentlichen Nachrichtentext [65].

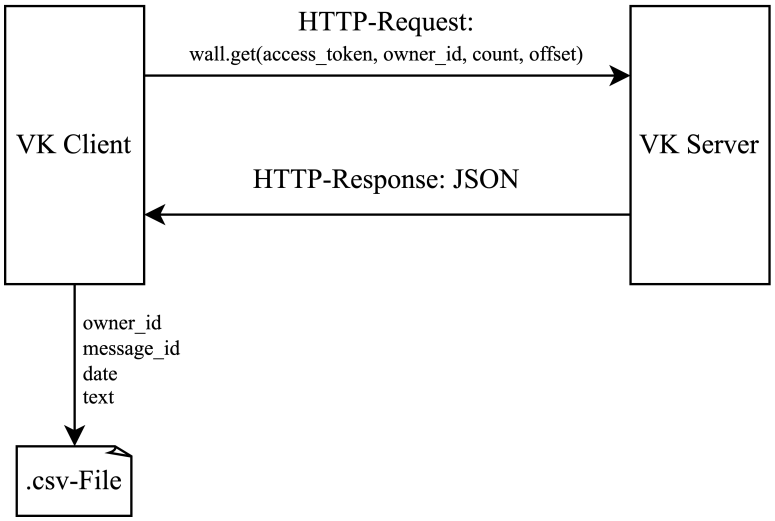


Abbildung 3.2: Ablaufdiagramm der Datenextrahierung mithilfe des VK API

Um die Daten zu analysieren, wurde die Pandas-Bibliothek von Python verwendet. Um das zu ermöglichen, wurde das Dateiformat „.csv“ gewählt. Alle relevanten Daten aus der JSON-Datei wurden dementsprechend in eine .csv-Datei transformiert.

Die im Rahmen dieses Projekts erfassten Daten beinhalten die Community-ID (owner_id) zur Quellenidentifikation, die eindeutige ID des Beitrags (id), das Veröffentlichungsdatum (date) sowie den eigentlichen Nachrichtentext (text). Hierbei ist es hervorzuheben, dass das in der Serverantwort gelieferte Datum als Timestamp vorliegt. Vor der Integration in die .csv-Datei wird dieser in ein allgemeines Datumsformat konvertiert.

Die Nutzung des VK API bringt mit sich spezifische Herausforderungen. Eine wesentliche Einschränkung ist die fehlende Möglichkeit, direkt Datumsparameter festzulegen. Datenabfragen erfolgen sequenziell, beginnend bei einen neusten Einträgen. Ein Abruf kann lediglich 100 Datensätze umfassen [65]. Der Offset-Parameter muss daher verwendet werden, um die Datenabfrage kontinuierlich fortzusetzen. Die Festlegung des ge-

nauen Offset-Werts, um den Abruf am 31.12.2022 zu beginnen, erfordert eine iterative Vorgehensweise durch manuelle Erprobung.

Da ein klar definierter Endzeitpunkt für den Datenabruf fehlte, war es notwendig, in jeder Iteration 1000 Datensätze abzurufen und anschließend das Datum zu überprüfen. Der Offset wurde dann angepasst und die Abfrage fortgesetzt, bis sämtliche Daten für das Jahr 2022 erfasst waren.

Extrahierung der Daten mittels Telegram API

Beim Einsatz des VK API konnten Daten lediglich vom Ersten Kanal gesammelt werden. Um Informationen aus weiteren Quellen über Telegram zu erhalten, kam das Telegram API zum Einsatz. Dabei handelt es sich um ein Hilfsmittel, das die Erstellung individueller Telegram-Clients ermöglicht. Es nutzt das MTProto-Protokoll von Telegram für den Austausch zwischen Client und Server, wobei jede Anfrage an die API ein spezifisches Format und bestimmte Parameter aufweist [55]. Zur Vereinfachung dieser Interaktion wurde die Telethon-Bibliothek eingesetzt. Sie bietet eine High-Level-Schnittstelle und verdeckt viele Details der MTProto-Implementierung [56].

Für die Autorisierung mittels Telegram API sind die Parameter api_id und api_hash erforderlich. Diese können direkt aus den Einstellungen des jeweiligen Telegram-Kontos bezogen werden [54].

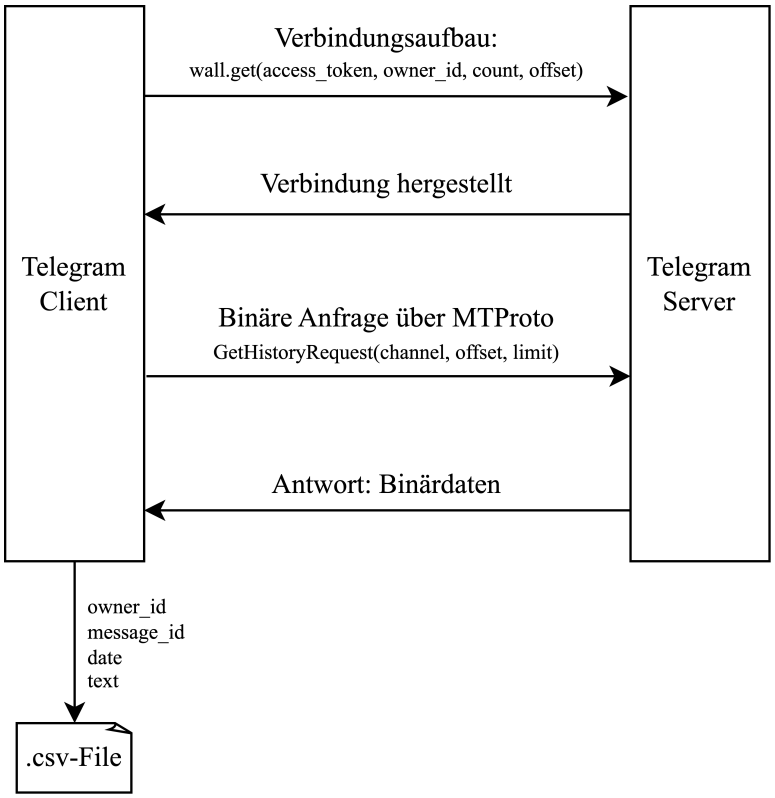


Abbildung 3.3: Ablaufdiagramm der Datenextrahierung mithilfe des Telegram API

Ein spezifischer Telegram-Client wurde für die Datensammlung entwickelt. Nach der Verbindung zum Telegram-Server über die api_id und api_hash-Parameter sendet dieser Client eine binäre MTProto-Anfrage an den Server. Der Kern dieser Anfrage ist die Methode GetHistoryRequest(), welche die Konversationshistorie eines Chats abruft [55]. Die verwendeten Parameter sind der Kanallink (channel), der Offset (offset) sowie das Limit der Nachrichten pro Anfrage (limit). Der visuelle Prozess ist in der Abbildung 3.3 abgebildet.

Auf diese Anfrage werden binäre Daten vom Server zurück bekommen. Mithilfe von Telethon-Bibliothek werden diese in Python-Objekte umgewandelt [56]. Genau wie beim VK API wurden relevante Informationen, wie ID der Community (owner_id), Dateneinheit-ID (id), Veröffentlichungsdatum (date) und Beitragstext (text), aus diesen Objekten gefiltert und in einer .csv-Datei gespeichert.

Ein Nachteil des Telegram API, welchen auch VK API hat, ist die Beschränkung auf 100 Nachrichten pro Anfrage. Daher ist eine obere Grenze für die Nachrichtenanzahl festzulegen. Diese wurde iterativ ermittelt, um Daten für 2022 von Medien wie Ria, Meduza und Mediazona zu sammeln.

3.4.2 Datenbereinigung

Während der Datensammlung und bei ersten Datenüberprüfung wurde festgestellt, dass eine umfassende Datenbereinigung vor dem Beginn der Analyse erforderlich ist. Der Prozess der Datenbereinigung wird in folgende Phasen strukturiert:

1. *Standardisierung der Formate*: Schon während des Datenabrufs aus den Informationskanälen wurde eine erste Datenbereinigung vorgenommen. Insbesondere wurde das „Datum“-Feld vom Timestamp-Format ins Date-Format konvertiert.

2. *Korrektur von Unstimmigkeiten*: Ohne die Option, Daten nach einem bestimmten Datum zu filtern, wurden Daten aus den Jahren 2021 und 2023 erfasst, die nicht im Fokus dieser Untersuchung stehen.

3. *Erkennung von Duplikaten*: Änderungen in der Serverdatenbank während des Abrufprozesses können zu Duplikationen führen, wenn zum Beispiel eine neue Nachricht hinzugefügt wurde. Diese Duplikate wurden im Nachgang entfernt.

4. *Umgang mit fehlenden Werten*: Einige Nachrichten enthalten keinen Text, sondern lediglich ein Bild oder Video. In den heruntergeladenen Daten zeigt sich dies als leerer Wert im Feld „Text“. Solche Einträge wurden entfernt.

5. *Bereinigung von Ausreißern*: Extrem kurze Nachrichten von weniger als fünf Wörtern sowie unerwünschte Zeichen, wie Zeilenumbrüche innerhalb von Texten, wurden bereinigt.

Die genannten Schritte wurden direkt nach der Datensammlung angewendet. Allerdings stellte sich die Datenbereinigung in dieser Arbeit als kontinuierlicher Prozess heraus, der sich von der Sammlung bis zur experimentellen Analyse erstreckt hat. Es wurde auch festgestellt, dass es nicht praktisch ist, alle notwendigen Bereinigungsschritte vor der Analyse vorzunehmen.

Während der NER-Analyse haben sich weitere Anpassungsbedarfe offensichtlich herausgestellt. Ein besonderes Problem stellte die unterschiedliche Bezeichnung identischer Entitäten dar. Ein gutes Beispiel davon sind die verschiedenen Namen für Länder wie „Russland“, „RF“ oder „Russische Föderation“ und „USA“, „Amerika“ oder „Vereinigte Staaten von Amerika“. Für eine konsistente Analyse ist es entscheidend, wie oft eine bestimmte Entität im Text erwähnt wird.

Eine weitere Herausforderung war die Benutzung von Emojis in Medientexten. Die NER-Analyse erkannte nicht immer korrekt, ob ein Emoji Teil des Namens einer Entität ist, was zu unterschiedlichen Schreibweisen führte. Ein praktisches Beispiel ist der Name „USA“ gefolgt von einem Emoji. Jede dieser Varianten wurde als eine separate Entität interpretiert.

Es war daher notwendig, sämtliche Varianten von Entitätsnamen zu standardisieren. Zum Beispiel wurde im Text jedes Vorkommen von „RF“ durch „Russland“ ersetzt. Dieser Schritt der Datenbereinigung war besonders zeitintensiv, da er eine hohe technische Präzision und viel manuelle Arbeit erforderte.

3.4.3 Übersetzung

Die ursprünglichen Daten lagen in russischer Sprache vor. Sowohl die Erstaufbereitung als auch die folgende Analyse wurden auf Russisch durchgeführt. Aufgrund des umfangreichen Datenmaterials war eine manuelle Übersetzung nicht praktikabel, weshalb ein automatisches Übersetzungstool zur Anwendung kam. Bei solchen Werkzeugen ist es jedoch stets eine Herausforderung, die Übersetzungsgenauigkeit für alle Texte zu gewährleisten. Weil eine präzise Übersetzung für das richtige Verständnis der Ergebnisse sehr wichtig ist, wurde die Datenbank vor der Analyse nicht ins Deutsche überzetzt. Für diese Abschlussarbeit wurden lediglich die endgültigen Analyseergebnisse übersetzt.

Der erste Übersetzungsversuch erfolgte mit dem Tool deep_translator von GoogleTranslator. Dabei bestätigte sich die Vermutung hinsichtlich der Übersetzungsqualität, und viele Ergebnisse waren von geringer Qualität (siehe Tabelle 3.5). Beispielsweise wurde das Akronym „SWO“, welches für „spezielle Kriegsoperation“ steht, falsch interpretiert und als „alles“ übersetzt. Deswegen wurde nach der automatisierten Übersetzung noch eine manuelle Überarbeitung vorgenommen.

Russisch	Automatische Übersetzung von deep_translator	Tatsächliche Bedeutung
специальная военная операция	besondere militärische Operation	spezielle Kriegsoperation
СВО	alles	SWO (spezielle Kriegsoperation)
спецоперация	Sondereinsatz	Spezialoperation

Tabelle 3.5: Beispiele für automatische Übersetzung und die tatsächliche Bedeutung der Worte

3.5 Diskussion

Im Rahmen dieser Bachelorarbeit stellt die Datenvorbereitung einen wichtigen Aspekt dar, da sie für den gesamten Forschungsprozess von Bedeutung ist. Externe Faktoren, insbesondere politische Ereignisse, können die Medienberichterstattung beeinflussen und eine konsistente Datenanalyse erschweren. Bei der Datensammlung über APIs von Plattformen wie Telegram und VK zeigen sich technische Einschränkungen, welche die Erfassung umfangreicher Datenmengen einschränken. Bei der Anwendung von NER sind Herausforderungen, vor allem bei der Identifikation von Synonymen und der Interpretation von Emojis, ersichtlich. Der Einsatz automatischer Übersetzungstools unterstreicht die Notwendigkeit menschlicher Überprüfung im wissenschaftlichen Kontext. Insgesamt bestätigt diese Untersuchung die Relevanz einer sorgfältigen Datenvorbereitung und durchgängigen Qualitätskontrolle für eine valide Datenanalyse.

4 Technischer Hintergrund

Gesammelte und bereinigte Daten dienen als Grundlage für die technische Analyse. In diesem Kapitel der Bachelorarbeit werden technische Aspekte der Analyseimplementierung erläutert. Hierbei werden die während der Arbeit eingesetzten Werkzeuge und Softwarelösungen betrachtet. Dieses Kapitel beschreibt die Funktionalitäten der verwendeten Bibliotheken bis hin zu den Verfahren zur Visualisierung der Ergebnisse mittels spezialisierter Software. Dieser Teil der Abschlussarbeit ermöglicht es, sich intensiver mit den technischen Details der Studie auseinanderzusetzen und die zentralen Schritte der Textdatenanalyse zu verstehen. Somit wird der Grundlage für die weiterführende Betrachtung der Analyseergebnisse im abschließenden Kapitel 5 gelegt.

4.1 Auswahl der Bibliotheken

In diesem Abschnitt wird eine Einführung in die Python-Bibliotheken „Pandas“ und „Natasha“ gegeben. Beide Bibliotheken wurden für dieses Projekt ausgewählt und verfügen über spezifische Eigenschaften und Funktionalitäten, die sie für die vorliegenden Anforderungen besonders hilfreich machen. Im Folgenden wird näher auf die Besonderheiten und Funktionen jeder einzelnen Bibliothek eingegangen.

Pandas-Bibliothek

In dieser Ausarbeitung nimmt die Bibliothek „Pandas“ eine wesentliche Rolle. Sie ermöglicht den Einsatz von fortgeschrittenen Datenstrukturen und Methoden, die eine effiziente und vereinfachte Datenanalyse gewährleisten [14]. Für die Analyse in dieser Arbeit wurde das DataFrame-Format von Pandas verwendet. Das Format ermöglicht die Darstellung von zweidimensionalen Datentabellen mit entsprechenden Zeilen- und Spaltenbeschriftungen.

Mithilfe von Pandas ist sowohl der Import als auch der Export von Daten im .csv-Format möglich. Ferner wurde die Konvertierung von Daten aus dem JSON-Format durchgeführt. Zudem war die Pandas-Bibliothek für eine vorläufige Datenbereinigung zuständig, wobei fehlende Werte (NaN) in der Spalte „Text“ durch leere Zeichenfolgen ersetzt wurden.

Natasha-Bibliothek

Bei der Verarbeitung der russischen Sprache besteht die Herausforderung, eine leistungsfähige Python-Bibliothek zu finden, die diese Sprache unterstützt. Zu den anerkannten Bibliotheken für NLP für russische Sprache zählen NLTK, SpaCy, DeepPavlov und Natasha. Aufgrund des umfangreichen zu verarbeitenden Datenvolumens steht die Leistung der Bibliotheken im Vordergrund der Auswahlkriterien. Natasha stellt sich hierbei als besonders schnell in der Verarbeitung heraus. Da sie speziell auf Medien-nachrichtentexten trainiert wurde, eignet sie sich optimal für die Analyse solcher Daten, was für diese Bachelorarbeit von zentraler Bedeutung ist.

Die Bibliothek Natasha spezialisiert sich auf die Analyse russischer Texte. Dabei basiert die Bibliothek auf einer Kombination aus bestimmten Regeln und Wörterbüchern. Sie erfüllt zentrale Funktionen in der Verarbeitung der russischen Sprache, zu denen die Segmentierung in Token und Sätze, die morphologische sowie syntaktische Analyse, die Lemmatisierung und die Extraktion und Normalisierung von Entitäten gehören. Es ist dabei hervorzuheben, dass die Natasha-Bibliothek nicht primär für Forschungszwecke entwickelt wurde. Die dahinterstehenden Technologien sind für praktische Anwendungen gedacht. Das Projekt umfasst neun Repositorien, welche durch die Natasha-Bibliothek unter einer gemeinsamen Schnittstelle vereint werden, wie in der Tabelle 4.1 dargestellt [8]. Zwei der Unterbibliotheken, Nerus und Navec, wurden als vortrainierte Modelle für die russische Sprache in die SpaCy integriert [9].

Repository	Beschreibung
Natasha	Sammlung von hochwertigen Open-Source-Tools zur Verarbeitung natürlicher russischer Sprache, Schnittstelle für die Low-Level-Bibliotheken des Projekts
Razdel	Segmentierung russischsprachiger Texte in Token und Sätze
Slovnet	Deep Learning-Modellierung zur Verarbeitung natürlicher russischer Sprache
Navec	Darstellung vortrainierter Wortdatensätze für russische Sprache
Nerus	Großer synthetischer Datensatz mit Markierungen für Morphologie, Syntax und benannte Entitäten
Corus	Sammlung von Links zu öffentlichen russischsprachigen Datensätzen mit Funktionen zum Herunterladen
Naeval	Quantitative Vergleichssysteme für russischsprachige NLP
Yargy-Parser	Extraktion von strukturierten Informationen aus russischsprachigen Texten mit Hilfe von Grammatiken und Wörterbüchern
Ipymarkup	Visualisierung der Markierung von benannten Entitäten und syntaktischen Verbindungen

Tabelle 4.1: Liste der Repositorien (Unterbibliotheken) des Natasha-Projekts [8]

Funktional ähnelt die Natasha der SpaCy-Bibliothek, hat jedoch eine ausführlichere Schnittstelle, wobei der Benutzer die einzelnen NLP-Schritte selbst definieren muss.

SpaCy initialisiert und ruft Modelle implizit auf. Somit gibt der Benutzer einen Text in die Funktion NLP ein und erhält als Ergebnis ein vollständig analysiertes Dokument. Wiederum übertrifft die Unterbibliothek Razdel andere Open-Source-Lösungen bei der Tokenisierung zur Verarbeitung der russischen Sprache. Wie in der Tabelle 4.2 zu erkennen ist, weist Razdel bessere Leistung sowohl bei der Geschwindigkeit als auch bei der geringeren Anzahl an Fehlern pro 1000 Token gegenüber größeren Modellen auf.

Lösungen zur Tokenisierung	Fehler pro 1000 Token	Verarbeitungszeit in Sekunden
SpaCy	17	5.4
NLTK	130	3.1
Razdel	7	2.6

Tabelle 4.2: Auflistung der Bibliotheken zur Tokenisierung und deren Eigenschaften

In der Datenanalyse dieser Bachelorarbeit wurde Natasha für die Funktionen der Tokenisierung, Segmentierung, Morphologie, Lemmatisierung, NER, Named Entity Normalization, Named Entity Parsing eingesetzt. Für alle diese Aufgaben nutzt Natasha die dazugehörigen Repositorien. Zum Beispiel, die NER-Aufgabe übernimmt das Repository Slovnet, die Tokenisierung - Razdel. Besonders zu betonen ist es, dass die Natasha-Bibliothek in dieser Arbeit für die NER-Analysen zum großen Teil eingesetzt wurde. Ein wesentlicher Vorteil dieser Bibliothek ist es, dass sowohl die Schritte des klassischen NLP-Ansatzes als auch die spezifischen NER-Analyseverfahren durch nur einen Befehl ausgeführt werden können. Durch die Anforderung, alle Schritte explizit zu definieren, lässt sich der Analyseprozess in jeder Phase effizient kontrollieren.

4.2 Verwendung von Tools

Nach der technischen Analyse ist eine Visualisierung der Daten für die Auswertung erforderlich. Während der Datenanalyse wurde festgestellt, dass die Visualisierugseigenschaften der Python IDE nicht den Anforderungen der Studie entsprechen. Mit dem zeitlichen Aufwand zur Laufzeit von über 30 Minuten bis zu einer Stunde ist für diese Studie nicht geeignet. Aus diesem Grund wurde entschieden, zusätzliche Werkzeuge wie Tableau und Gephi zu verwenden.

Tableau ist eine führende Visual-Analytics-Plattform, die innovative Ansätze zur Datenanalyse und Datennutzung bietet. Sie ermöglicht den Nutzern, komplexe Datenmengen einfach zu erkunden und zu verwalten, wodurch die Ergebnisanalyse beschleunigt wird [15]. Da Tableau zu den marktführenden Visualisierungsplattformen gehört, wurde es als bevorzugtes Visualisierungswerkzeug in dieser Bachelorarbeit ausgewählt.

Ein zusätzliches Visualisierungsmittel dieser Bachelorarbeit ist die Gephi-Software, welche für Netzwerkdarstellungen sich besonders eignet. Mittels Gephi ist im Kapitel 5 ein

Graph visualisiert. Dieser Graph stellt die Ergebnisse der NER-Analyse dar, und zwar die Zusammenhänge zwischen Personen, Organisationen und Orten aus der Nachrichtentexten. Mit der Software ist es möglich, die Darstellung der Knoten interaktiv zu verändern [5].

4.3 Pipeline

In dieser Bachelorarbeit wurde ein Datenprozess beginnend mit der Datensammlung bis hin zur Visualisierung durchgeführt. Dieser Prozess wird in der Abbildung 4.1 veranschaulicht. Die Datensätze wurden bei den Plattformen VKontakte und Telegram mittels ihrer APIs gesammelt. Nach der Datensammlung wurde eine zentrale Datenbank eingerichtet, welche in Form von einer .csv-Datei gespeichert wurde. Darauf basierend erfolgte eine Auswertung: angefangen bei einer allgemeinen Datenanalyse, über die Kontextanalyse ausgewählter Wörter, bis hin zur Netzwerkanalyse von Zusammenhänge zwischen Personen, Organisationen und Orten.

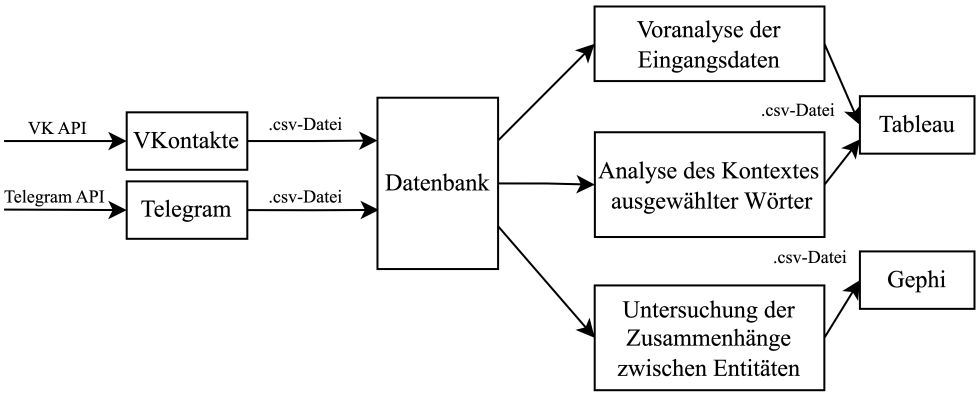


Abbildung 4.1: Prozessablauf der Daten vom Datensammlung bis zur Visualisierung

Für die Entwicklung der Datenverarbeitungsbefehle wurde die Programmiersprache Python verwendet. Abgesehen von der Datensammlung und Datenbereinigung fand die Entwicklung des Codes hauptsächlich im Jupyter Notebook statt. Die Datensammlung und Datenbereinigung wurde in Visual Studio Code-IDE entwickelt und als separate Funktionen und Skripte organisiert. Alle Python-Skripte, die während dieser Studie entwickelt wurden, sind auf der beigefügten CD zu finden.

Folgend wird jeder dieser Schritte technisch näher beleuchtet. Die ausführliche technische Erläuterung zur Datensammlung von den Plattformen VKontakte und Telegram mittels der jeweiligen APIs sowie Aspekte der Datenbereinigung und Datenübersetzung sind bereits im Kapitel 3 behandelt worden. Daher konzentriert sich dieser Abschnitt hauptsächlich auf den Analyseteil dieser Bachelorarbeit.

Die Datenanalyse gliedert sich in drei Hauptphasen:

1. Voranalyse der Eingangsdaten: hier wurde eine Übersicht über die Daten erstellt, indem die Anzahl der Nachrichten pro Tag ausgewertet und die häufigsten Wörter in den Texten identifiziert wurden.

2. Analyse des Kontextes ausgewählter Wörter: in dieser Phase stand die Untersuchung von bestimmten Wortpaaren im Mittelpunkt. Beispielhaft wurden die Wortpaare je nach ihrer Verwendung basierend auf dem Konzept des Neusprechs analysiert.

3. Untersuchung der Zusammenhänge zwischen Entitäten: hierbei wurde ein Netzwerkdiagramm mithilfe der NER-Methode erstellt, um die Verbindungen zwischen Personen, Organisationen und Orten in den Nachrichtentexten zu visualisieren.

Jeder der oben genannten Schritte hatte sowohl eine theoretische als auch eine praktische Bedeutung für das Verständnis der untersuchten Thematik. Für die Darstellung der Ergebnisse aus den ersten beiden Analysepunkten kam die Plattform Tableau zum Einsatz. Die Visualisierung des dritten Punktes wurde mit Hilfe der Software Gephi realisiert. Bevor die Daten jedoch mittels der Programme verarbeitet wurden, wurden sie entsprechend dem im Kapitel 3 beschriebenen Verfahren auf Deutsch übersetzt.

Während der ersten Phase, der Voranalyse der Eingangsdaten, wurde die Datenbank in Tableau geladen, um die Anzahl der täglich veröffentlichten Artikel zu bewerten. Für die Untersuchung der häufigsten Wörter in den Texten war eine zusätzliche Vorverarbeitung der Rohdaten notwendig. Wie es in der Abbildung 4.2 dargestellt ist, liest ein Python-Skript zuerst die Daten aus, um eine Tokenisierung und Normalisierung vorzunehmen. Anschließend wurde ein Wörterbuch generiert, das die Schlüssel normalisierter Wörter repräsentiert und die Häufigkeitswerte anzeigt. Die Gesamtergebnisse wurden in eine .csv-Datei exportiert. Zum Schluss wurden aus diesen Daten die zehn am häufigsten vorkommenden Wörter für jedes Medium separat extrahiert und analysiert.

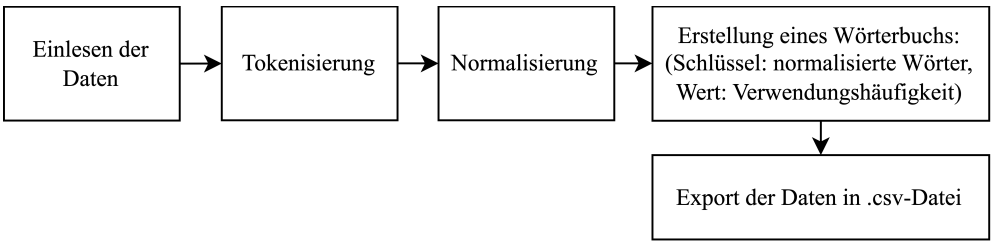


Abbildung 4.2: Ablaufdiagramm des Skripts zur Identifikation häufig wiederholter Wörter

Dieser Prozess stellt sicher, dass die relevantesten und am häufigsten wiederholten Wörter aus den analysierten Texten effektiv identifiziert und dargestellt werden können. Dadurch wird eine ausführliche Einsicht in die Hauptthemen in den untersuchten Medien ermöglicht.

Aufgrund der Ergebnisse aus der bereits durchgeführten Analyse wurden die häufigsten Wörter für alle Medien gewählt. Um die Verwendung ausgewählter Wörter in dieser Analysephase bewerten zu können, wurden zunächst Daten eingelesen und eine Überwachungsliste von Wörtern erstellt, wie es in der Abbildung 4.3 dargestellt ist. Nach der Tokenisierung und Normalisierung der Texte wurde Verwendungshäufigkeit dieser Wörter in den Texten nach Datum organisiert. Das Ergebnis wurde daraufhin in eine .csv-Datei exportiert, welche die Spalten für das jeweilige Datum und die Verwendungshäufigkeit jedes Wortes an diesem bestimmten Datum beinhaltet. Schließlich wurden diese Dateien in Tableau importiert, um sie visuell zu präsentieren.

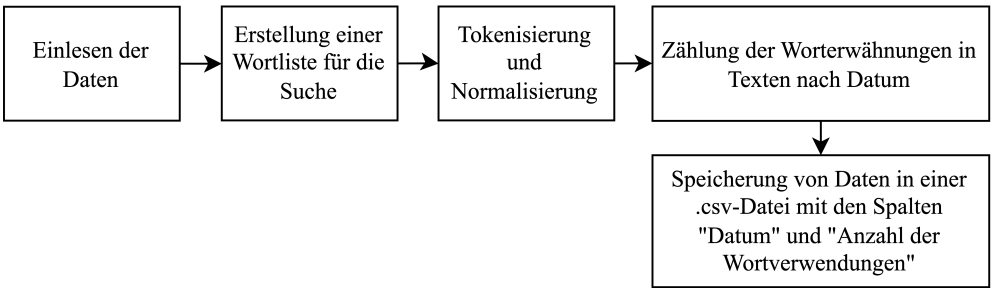


Abbildung 4.3: Ablaufdiagramm des Skripts zur Analyse ausgewählter Wörter

Durch diesen Prozess können ausgewählte Wörter und deren Verwendung in den analysierten Texten genau verfolgt werden. Das visuelle Ergebnis in Tableau gibt einen klaren Überblick darüber, wie oft und wann diese spezifischen Wörter in den Medien erwähnt wurden, wodurch die Interpretation von Trends und Themen in den Daten vereinfacht wird.

Während der zweiten Analysephase, die Kontextanalyse ausgewählter Wörter, wurde die NER-Methode eingesetzt. Wie es in der Abbildung 4.4 gezeigt wird, begann der Prozess mit dem Einlesen der Daten und dem Festlegen der Schlüsselwörter für die Analyse. Nach der Tokenisierung und Normalisierung der Texte wurde jedes Vorkommen dieser Schlüsselwörter gezählt und der betreffende Satz in einem separaten Array gespeichert. Darauf aufbauend wurden mittels der NER-Aufgabe die Entitäten wie Personen, Organisationen und Orte aus den erfassten Sätzen extrahiert und ihre Verwendungshäufigkeit ermittelt. Diese Daten wurden in eine .csv-Datei exportiert, welche die Spalten Entität, Entitätstyp und Anzahl der Entitätsverwendungen enthält. Schließlich wurden die Daten in Tableau importiert, um die Ergebnisse in Form von Wortwolken visuell aufzubereiten.

Dieser Ansatz ermöglicht es, nicht nur die Verwendungshäufigkeit ausgewählter Wörter in den analysierten Texten zu ermitteln, sondern auch den Kontext im Bezug auf bestimmte Entitäten zu vestehen. Dies enthält wichtige Informationen über die Wortbeziehungen.

Mit der dritten Analysephase wurde der Fokus auf der NER-Analyse gelegt, wobei die Zusammenhänge zwischen den Entitäten zu untersuchen waren. Die Ergebnisse dieser

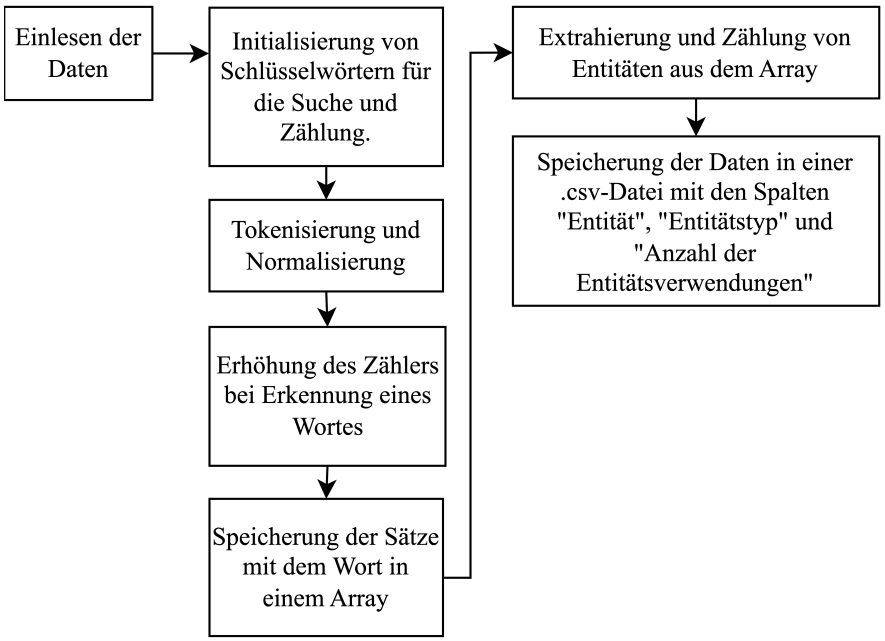


Abbildung 4.4: Ablaufdiagramm des Skripts zur Analyse des Kontexts ausgewählter Wörter

Analyse wurden mittels eines Graphen visualisiert. Wie die Abbildung 4.5 zeigt, beginnt der Prozess mit dem Einlesen der Daten. Anschließend wurden sie für die NER-Analyse aufbereitet. Diese umfasst die Satzsegmentierung, morphologische Analyse und syntaktisches Parsing. Darauf folgt die Extraktion von Personen, Organisationen und Orten aus den vorbereiteten Texten. Wenn zwei dieser Entitäten im selben Satz vorkommen, wird zwischen denen eine Verbindung in Form einer Kante im Graphen erstellt. Tritt eine solche Verbindung in den Daten mehrfach auf, wird das Gewicht dieser Kante entsprechend angepasst. Zum Abschluss wurden die gesammelten Informationen in zwei Dateien entsprechend für die Knoten und für die Kanten exportiert. Die Informationen zu den Knoten wurden in einer Liste gespeichert, während die Informationen zu den Kanten im Format „Source, Target, Weight“ festgehalten wurden. Hierbei repräsentiert „Weight“ das Gewicht der Kante innerhalb eines Datensatzes. Diese .csv-Dateien wurden als Grundlage für die Erstellung des Graphen in der Software Gephi genommen.

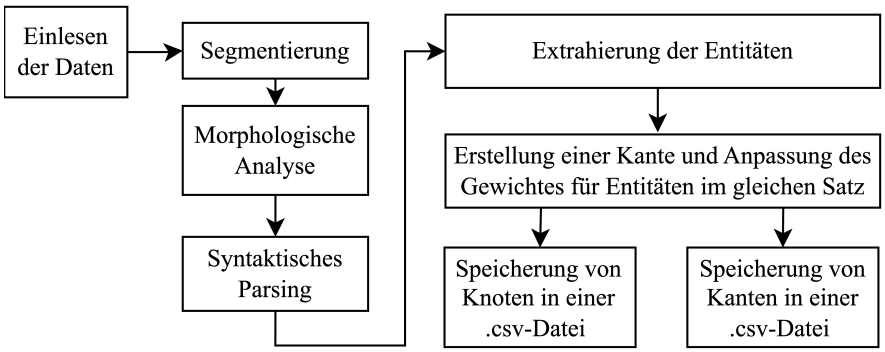


Abbildung 4.5: Ablaufdiagramm des Skripts zur Untersuchung der Zusammenhänge zwischen Entitäten

Somit verwendet jede Analysephase aus technischer Einsicht zentrale Techniken des klassischen NLP-Ansatzes und der NER-Aufgabe. Es lässt sich zusätzlich feststellen, dass sich die Intensität der NER-Analyse von einer zu anderer Analysephase erhöht. Schließlich kamen in der finalen Analysephase diverse NER-Methoden zum Einsatz, damit die Daten für die grafische Darstellung vorbereitet werden können.

4.4 Diskussion

In diesem Kapitel wurde die technische Implementierung der Analysephasen erläutert. Hierbei wurden insbesondere die Bibliotheken Pandas und Natasha vorgestellt, die essenzielle Funktionen für die Datenverarbeitung und Datenanalyse zur Verfügung stellen. Darüber hinaus wurde auf die Verwendung von Visualisierungswerkzeugen wie Tableau und Gephi eingegangen.

Zentral in diesem Kapitel war die Darstellung der technischen Vorgehensweise der Analyse. Durch diese methodische Herangehensweise war es möglich, eine tiefgreifende Analyse von Textdaten zu realisieren und die resultierenden Erkenntnisse effektiv darstellen zu können. Dies stellte sicher, dass die Ergebnisse der Arbeit nicht nur begründet, sondern auch verständlich präsentiert. Dies erleichterte die Interpretation der Ergebnisse und legte den Grundstein für ihre mögliche Verwendung in zukünftigen Studien oder Anwendungen. Eine ausführliche Betrachtung der Visualisierungsergebnisse sowie deren kritische Bewertung werden im folgenden Kapitel dargelegt.

5 Experimente

In diesem Kapitel steht die Analysebewertung erstellten Visualisierungen im Vordergrund. Zunächst wird eine allgemeine Datenbewertung durchgeführt, bei der sowohl die Anzahl der täglich veröffentlichten Artikel als auch die häufigsten Wörter in den untersuchten Texten berücksichtigt werden. Darüber hinaus erfolgt eine Untersuchung des Kontextes derjenigen Wörter, die basierend auf dem Verständnis von Neusprech ausgewählt wurden. Zum Schluss bildet sich eine Graphenanalyse, welche Zusammenhänge zwischen den Entitäten der analysierten Medien zeigt.

5.1 Voranalyse der Eingangsdaten

In diesem Abschnitt erfolgt eine allgemeine Datenanalyse, die sich auf die Ereignisse während des Beginns des russischen Angriffs auf die Ukraine konzentriert. Zuerst wird ein Blick auf die Publikationsdynamik von vier Medienressourcen geworfen, wobei insbesondere die Anzahl der täglich veröffentlichten Nachrichten im Fokus steht. Anschließend wird eine Untersuchung der in den Medientexten am häufigsten vorkommenden Wörter vorgenommen. Dies soll ein tieferes Verständnis der Schwerpunkte und Tendenzen des Informationsflusses geben.

5.1.1 Überblick über die Daten

Einige statistische Informationen wurden bereits im Kapitel 3 präsentiert, indem die Anzahl der im Jahr 2022 von jedem Medium veröffentlichten Artikeln genannt wurde. In diesem Zusammenhang wird die Anzahl der in den Medientexten verwendeten Wörter untersucht. Wie aus der Tabelle 5.1 ersichtlich ist, zeigt die Medienressource Ria für das Jahr 2022 die höchste Anzahl an Datensätzen. Allerdings hat Meduza die größte Anzahl an insgesamt verwendeten Wörtern. Besonders fällt der fast doppelt so hoher Wert auf. Das bedeutet, dass die Artikel bei Meduza im Durchschnitt länger als bei Ria sind. Bei den anderen Medien scheinen die Werte in einem zu erwarteten Bereich zu liegen.

Im Laufe der Untersuchung wurde die Einzigartigkeit der in den Medien verwendeten Wörter analysiert. Es hat sich herausgestellt, dass die Werte für Ria und Meduza vergleichsweise konsistent sind (siehe Tabelle 5.1). Bei dem Ersten Kanal und Mediazona hingegen gab es signifikante Unterschiede. Obwohl der Erste Kanal fast dreimal so viele

Quelle	Anzahl der Dantesätze	Gesamtzahl der Wörter	Anzahl der einzigartigen Wörter
Der Erste Kanal	8 486	610 386	28 124
Ria	48 601	1 576 520	47 804
Meduza	24 578	2 966 026	62 355
Mediazona	3 076	334 139	21 118

Tabelle 5.1: Aufzählung der im Jahr 2022 von jedem Medium verwendeten Wörter

Datensätze wie Mediazona aufweist und die Gesamtzahl der Wörter nahezu doppelt so hoch ist, beträgt die Differenz in der Einzigartigkeit der verwendeten Wörter lediglich etwa 7.000. Dies könnte darauf hindeuten, dass die Sprachnutzung vom Ersten Kanal tendenziell repetitiver und weniger vielfältig ist.

Zudem wurde in der Analyse darauf geachtet, prozentuale Werte anstelle von absoluten Zahlen zu verwenden. Dies wurde aufgrund des ersichtlichen Unterschieds der Datenmenge zwischen den verschiedenen Medien festgelegt (siehe Tabelle 5.1). Ein direkter Vergleich zwischen Mediazona und Ria basierend auf absoluten Zahlen wäre unangemessen, da Mediazona durchschnittlich etwa 20 Artikel pro Tag veröffentlicht, während die Zahl bei Ria über 100 liegt.

Ein weiterer Aspekt in der Datenanalyse betrifft die Verneinungen. Diese sind sowohl für die Häufigkeitsanalyse von Wörtern als auch für das Verständnis ihres jeweiligen Kontexts besonders relevant. Verneinungen können die Bedeutung eines Textes grundlegend ändern.

In der Tabelle 5.2 ist es dargestellt, wie häufig die Verneinungen „nicht“, „aber“ und „nein“ in den Artikeln der verschiedenen Medien vorkommen. Bezogen auf die Gesamtwortzahl ergeben sich für die Medien folgende prozentuale Verhältnisse: der Erste Kanal - 0,8%, Ria - 1,1%, Meduza - 0,8% und Mediazona - 1,2%.

Verneinungen	Der Erste Kanal	Ria	Meduza	Mediazona
nicht	3 217	11 410	17 097	3 368
aber	1 189	2 662	4 390	623
nein	294	1 393	1 666	190
Summe	4 700	15 465	23 153	4 181

Tabelle 5.2: Anzahl der Verneinungen in den Nachrichtentexten

Die Verwendung von Verneinungen in den Medien liegt etwa im gleichen Bereich. Es lässt sich nicht eindeutig bestimmen, ob der Schreibstil, der unsichere Ton oder das behandelte Nachrichtenthema die Verwendung von Verneinungen beeinflusst. Beispielsweise lässt sich nicht feststellen, ob ein bestimmtes Medium primär darauf ausgerichtet ist, Fakten zu widersprechen. Folglich wurde entschieden, die Verneinungen in der weiteren Analyse dieser Arbeit nicht zu berücksichtigen.

5.1.2 Anzahl der Artikel pro Tag

Nach der ersten Überprüfung der Daten bestand der nächste Schritt darin, die von jedem Medium täglich veröffentlichte Artikelanzahl zu analysieren. Es ergab sich die Vermutung, dass der Kriegsanfang das Verhalten der Medien hinsichtlich der Anzahl der täglich publizierten Artikel beeinflussen soll. Die Ergebnisse dieser Analyse sind in der Tabelle 5.3 dargestellt. Die Interpratation der Daten zeigt, dass im Vergleich zum Zeitraum zwischen dem Jahresanfang und dem 23. Februar, am Tag des Kriegsbeginns alle vier analysierten Medien ihre Publikationsrate um über 100% steigerten. Besonders ersichtlich war es, dass Mediazona und Meduza ihre Artikelanzahl sogar um mehr als 600% bzw. 800% erhöht haben. Es ist ein Trend erkennbar, dass pro-kremlische Medien (wie der Erste Kanal, Ria) ihren Output um 300-600% weniger steigerten als oppositionelle Medien. Die Abbildung zur Anzahl der Artikel pro Tag für alle Medien im Jahr 2022 ist in der Abbildung A.2 im Anhang dargestellt.

Quelle	Publikationenanzahl bis inkl. 23.02.2022 (Durchschnittswert)	Publikationenanzahl am 24.02.2022	Wachstum, %
Der Erste Kanal	16,6	37	122,9%
Ria	176,1	666	278,2%
Mediazona	4,8	44	816,7%
Meduza	33,2	246	641%

Tabelle 5.3: Anzahl der Publikationen vor und am Tag des Kriegsbeginns

Obwohl die tägliche Artikelanzahl am Tag des Angriffs stark anstieg, zeigt sich diese Tendenz im weiteren Jahresverlauf nicht bei jedem untersuchten Medium gleichbleibend. In der Tabelle 5.4 sind Daten zur durchschnittlichen Anzahl der täglich veröffentlichten Artikel sowohl vor als auch nach dem 24. Februar dargestellt. Die Tabelle zeigt auch die prozentuale Veränderung für jede Quelle individuell. Aus der Tabelle geht hervor, dass alle Medien bis auf Ria einen erhöhten Informationsfluss aufwieisen. Bei Ria hingegen verringerte sich die durchschnittliche Anzahl der täglich publizierten Artikel seit dem Kriegsbeginn in der Ukraine um 29%.

Quelle	Publikationenanzahl bis inkl. 23.02.2022 (Durchschnittswert)	Publikationenanzahl ab 24.02.2022 (Durchschnittswert)	Wachstum, %
Der Erste Kanal	16,6	24,4	47%
Ria	176,1	125,7	-29%
Mediazona	4,8	9,1	90%
Meduza	33,2	73,3	121%

Tabelle 5.4: Veränderung der durchschnittlichen Anzahl der veröffentlichten Publikationen vor und nach Kriegsbeginn

Neben dem zentralen Datum, dem 24. Februar 2022, waren dieses Jahr auch andere bedeutende Ereignisse. Die statistische Stimmungsanalyse in Russland für 2022 zeigte zwei markante Spitzen in Bezug auf die Gefühlslage „Angst“. Zwischen dem 20. und 27. Februar sowie auch zwischen dem 18. September und dem 2. Oktober [42]. Der erste Zeitraum steht offensichtlich mit dem Kriegsbeginn am 24. Februar in Verbindung. Der zweite korreliert mit der Bekanntgabe einer personellen Mobilmachung am 21. September.

Es ließ sich annehmen, dass auch der 21. September ähnlich wie am 24. Februar eine gestiegene Medienaktivität in Bezug auf diese Bekanntgabe zeigen würde. Wie es in den Abbildungen 5.1 und 5.2 zu sehen ist, bestätigt sich diese Annahme bei den oppositionellen Medien Meduza und Mediazona. Insbesondere zeigte sich hier ein markanter Anstieg der Berichterstattung am 21. September: bei Meduza stieg die Aktivität um 130% and bei Mediazona sogar um 322,3% im Vergleich zum Zeitraum vom 1. bis 20. September.

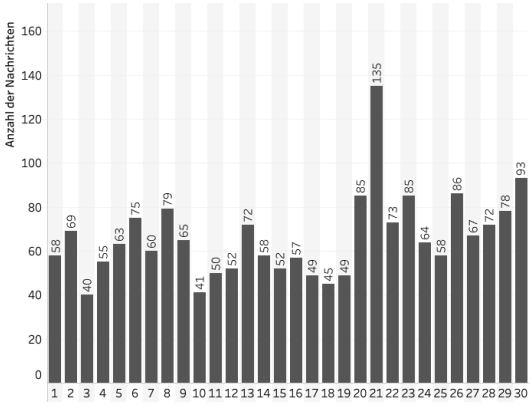


Abbildung 5.1: Anzahl der Nachrichten pro Tag bei Meduza im September 2022

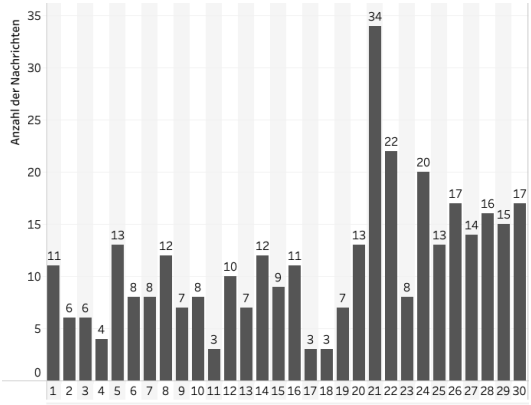


Abbildung 5.2: Anzahl der Nachrichten pro Tag bei Mediazona im September 2022

Die Abbildung 5.3 zeigt, dass auch Ria am 21. September einen Anstieg an veröffentlichten Artikeln um 85% im Vergleich zum Zeitraum vom 1. bis 20. September zu verzeichnen hatte.

Für den Ersten Kanal bestätigt sich diese Vermutung jedoch nicht. Gemäß der Abbildung 5.4 ist für den 21. September sogar ein Rückgang der Publikationen um 22,7% im Vergleich zum Zeitraum vom 1. bis 20. September festzustellen. Bei der weiteren Analyse zeigt sich, dass am 19. September, an einem Tag ohne bedeutende Ereignisse, ein Anstieg der Nachrichten um 72,9% im Vergleich zum Zeitraum vom 1. bis 18. September zu verzeichnen war. Bei der Analyse der Artikel dieses Tages ließ sich eine scharfe und aggressive Wortwahl verfolgen. Beispielsweise: „Über die Morde an Frauen und Kindern im Donbass, über ‚Säuberungen‘ unter der Zivilbevölkerung, all dies schauen amerikanische und europäische Journalisten weiterhin mit weit geschlossenen Augen an“ [13]. Oder: „In den getrennten Staaten von Amerika nimmt die Hexenjagd

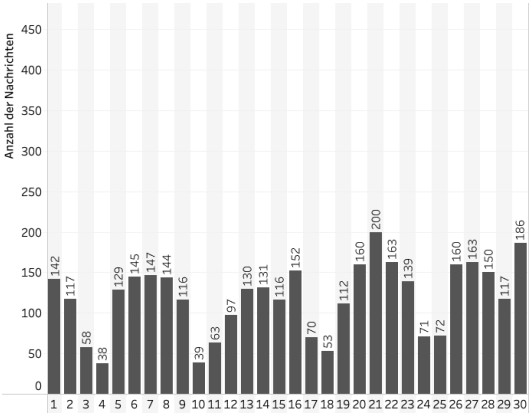


Abbildung 5.3: Anzahl der Nachrichten pro Tag bei Ria im September 2022

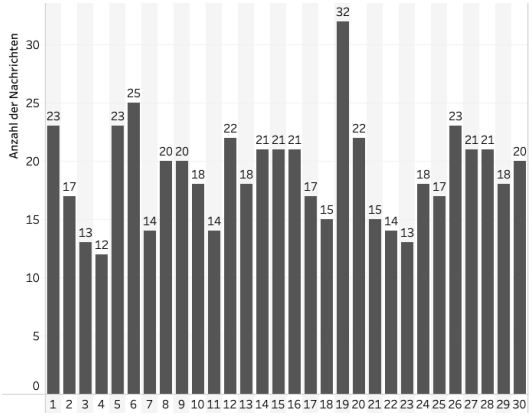


Abbildung 5.4: Anzahl der Nachrichten pro Tag beim Ersten Kanal im September 2022

mit neuer Kraft zu“ [13]. Eine solch aggressive Rhetorik könnte ein Indikator dafür sein, dass es versucht wurde, die patriotische Stimmung in der Bevölkerung zu verstärken, um den Mobilmachungsprozess zu erleichtern.

Bei einer separaten Betrachtung der Medien ohne Berücksichtigung der allgemeinen Stimmungslage in Russland im Jahr 2022 fielen zusätzliche Aktivitätsspitzen auf. Das im Anhang in der Abbildung A.2 dargestellte Diagramm zeigt, dass diese Spitzen sich je nach Mediengruppe variieren.

Beispielsweise zeigte der Erste Kanal eine erhöhte Aktivität am 9. März 2022. An diesem Tag eröffnete Russland humanitäre Korridore und es wurde über eine angebliche Provokation am Kernkraftwerk Saporischschja durch die Ukraine berichtet. Auch der 9. Mai zeigte eine signifikant gesteigerte Aktivität. Dieses Datum bezeichnet den Russlands Siegestag im Zweiten Weltkrieg und wird in den letzten Jahren häufig zu propagandistischen Zwecken genutzt. Die Aktivität von Ria verteilte sich ziemlich konstant über das Jahr 2022. Eine Ausnahme dabei ist der auffällige Anstieg am 24. Februar. Zudem waren geringfügige Schwankungen im Wochenverlauf zu beobachten, wie beispielsweise ein reduziertes Artikelvolumen an den Wochenenden.

Meduza und Mediazona wiesen wiederum zahlreiche, mäßig ausgeprägte Aktivitätsspitzen auf. Beispiele dazu waren solche Daten, wie der 14. März 2022 (Beschuss von Donezk), der 10. Oktober 2022 (Raketenangriff auf die Ukraine) oder der 15. November 2022 (erneuter Raketenangriff auf die Ukraine). Somit berichteten diese Medien nicht nur ausführlich über stimmungsrelevante Ereignisse in Russland, sondern auch über militärische Vorfälle in der Ukraine.

5.1.3 Worthäufigkeit pro Medium

Diese Analysephase beschäftigt sich mit der Untersuchung der Worthäufigkeit in den Artikeln. In der Tabelle 5.5 sind die vorläufigen Ergebnisse dargestellt, die sich aus den zehn am häufigsten vorkommenden Wörtern jedes Mediums zusammensetzen. Dies erfolgte nach der Textverarbeitung mithilfe der Natasha-Bibliothek und der Entfernung von Interpunktionen, Pronomen, Konjunktionen, Zahlen, Präpositionen und Emojis aus den Datensätzen.

1 Channel	RIA	Meduza	Mediazona
Russland	Russland	ausländisch	ausländisch
russisch	Ukraine	ausüben	russisch
militärish (das Militär)	ukrainisch	Funktion	ausüben
ukrainisch	russisch	russisch	Funktion
Ukraine	militärish (das Militär)	Agent	Agent
Mensch	Gebiet	Information	Ukraine
Land	USA	Nachricht	arbeiten
Ukrainische Streitkräfte (VSU)	Leiter	Subjekt	lesen
Post	Vladimir Putin	Mittel	Krieg
Donbass	Volksrepublik Donezk (DNR)	Material	Russland

Tabelle 5.5: Zusammenfassung der zehn am häufigsten wiederholten Wörter in jedem Medium

Es gab deutliche Übereinstimmungen zwischen den Wörtern, die in den Medien beider Gruppen am häufigsten vorkamen. Zum Beispiel zählten in dem pro-kremlischen Ria und dem Ersten Kanal „Russland“, „russisch“, „militärisch (das Militär)“, „ukrainisch“ und „Ukraine“ zu den am häufigsten wiederholten Wörtern. Die Schreibweise des Wortes „militärish (das Militär)“ ist in der russischen Sprache sowohl für ein Adjektiv als auch für ein Substantiv gleich. In den oppositionellen Medien hingegen dominieren Wörter, die mit der Kennzeichnung als ausländischer Agent in Verbindung stehen, wie es im Kapitel 3 erläutert wurde. Da diese Kennzeichnung nahezu jeden Artikel begleitete, wurde beschlossen, diese Wörter aus der Liste der häufigsten Wörter auszuschließen, um ein genaueres Ergebnis in der Analyse zu erzielen.

Demnach zeigt Tabelle 5.6 überarbeitete Wortlisten. Es ließ sich feststellen, dass viele der häufig wiederholten Wörter in den pro-kremlischen Medien auch in den oppositionellen Medien erschienen. Zudem haben sich alle diese Wörter auf politische und militärische Themen im Kontext des Konflikts zwischen Russland und der Ukraine bezogen. Dabei beschränkte sich keines der analysierten Medien offiziell auf eine politische Agenda.

1 Channel	RIA	Meduza	Mediazona
Russland	Russland	russisch	russisch
russisch	Ukraine	Russland	Ukraine
militärish (das Militär)	ukrainisch	Ukraine	arbeiten
ukrainisch	russisch	Krieg	lesen
Ukraine	militärish (das Militär)	militärish (das Militär)	Krieg
Mensch	Gebiet	ukrainisch	Russland
Land	USA	Leiter	militärish (das Militär)
Ukrainische Streitkräfte (VSU)	Leiter	Vertreter	sterben
Post	Vladimir Putin	Präsident	ukrainisch
Donbass	Volksrepublik Donezk (DNR)	Truppen	Tag

Tabelle 5.6: Zusammenfassung der zehn am häufigsten wiederholten Wörter in jedem Medium (überarbeitet)

Es ist besonders hervorzuheben, dass die oppositionellen Medien das Wort „Krieg“ als eines der häufigsten benutzt haben. Dies deutet darauf hin, dass trotz des Gesetzes gegen die Verbreitung von Falschinformationen oppositionelle Medien dieses Wort regelmäßig verwenden. Eine genauere Untersuchung des Kontexts, in dem das Wort „Krieg“ in Medienberichten verwendet wurde, wird folgend vorgestellt.

Basierend auf der Worthäufigkeitsanalyse wurden Wörter identifiziert, die in sämtlichen Medien regelmäßig vorkommen: „Russland“, „russisch“, „militärish (das Militär)“, „ukrainisch“ und „Ukraine“. Die Verwendung dieser Wörter wurde im Kontext des zeitlichen Verlaufs untersucht.

In der Abbildung 5.5 wurde im Zeitraum von Februar bis März ein besonderer Anstieg aufgezeigt. Es ist deutlich zu erkennen, dass in der Woche des Kriegsausbruchs eine signifikante Erhöhung in der Häufigkeit der Verwendung der analysierten Wörter verzeichnet wurde. Im Vergleich zum Durchschnittswert vor der Invasion, verzeichnete das Wort „militärish (das Militär)“ in der Woche des Angriffs einen Anstieg von 560,9%, „russisch“ einen Anstieg von 360,5%, „Russland“ einen Anstieg von 329,7%, „Ukraine“ einen Anstieg von 670% und „ukrainisch“ einen Anstieg von 1025%.

Wie es aus der Abbildung A.8 im Anhang ersichtlich ist, blieb die Häufigkeit nach Erreichen des Höhepunkts weiterhin auf einem erhöhten Niveau bis Ende des Jahres. Im Vergleich zur Zeitperiode vor dem Krieg stieg die Verwendung der untersuchten Wörter für den Rest des Jahres wie folgt: „militärish (das Militär)“ um 230%, „russisch“ um 99,3%, „Russland“ um 64,5%, „Ukraine“ um 143% und „ukrainisch“ um 461,9%.

Erwartungsgemäß kam es in der Woche nach der Ankündigung der personellen Mobilmachung am 21. September 2022 zu einem ähnlichen Anstieg in der Verwendung

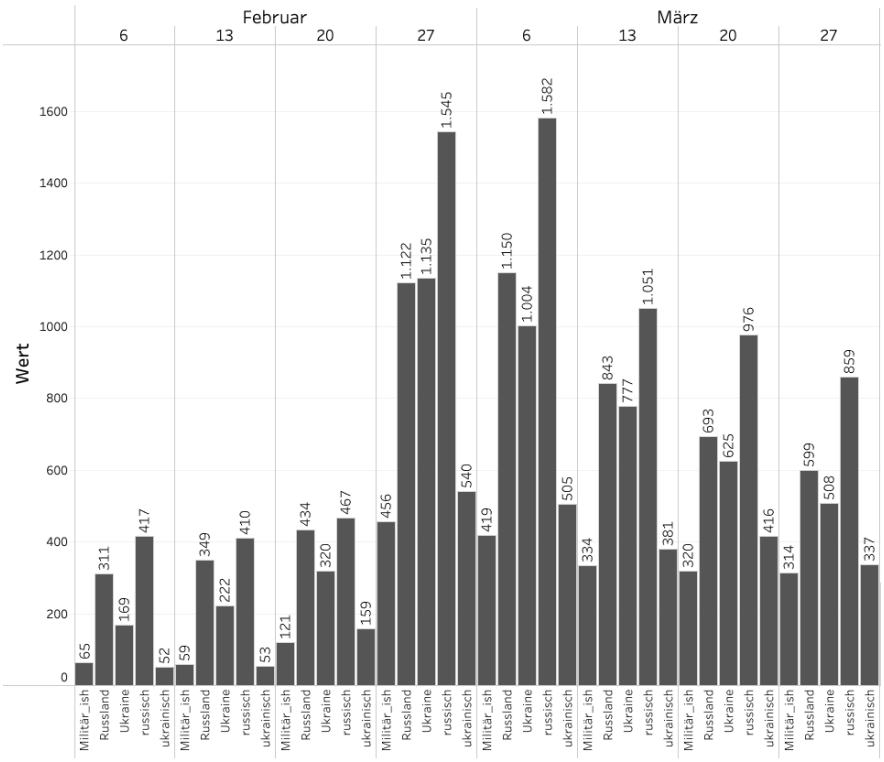


Abbildung 5.5: Diagramm der Verwendungshäufigkeit von ausgewählten Wörtern im Februar und März 2022

dieser Wörter. Allerdings zeigte sich kein ausgeprägter Anstieg, sondern lediglich eine geringfügige Erhöhung in der Häufigkeit von Wörtern wie „russisch“ und „Russland“.

Daraus lässt sich schließen, dass seit dem Kriegsbeginn in der Ukraine insbesondere die Wörter „militärish (das Militär)“ und „ukrainisch“ einen signifikanten Anstieg in ihrer Verwendung in Medienberichten erfahren haben. Diese Begriffe unterstreichen direkt die politische und militärische Thematik in den Medien zum damaligen Zeitpunkt.

5.2 Analyse des Kontextes ausgewählter Wörter

In dem Abschnitt zum Neusprech wurde das Wort „Spezialoperation“ eingeführt, welches laut der russischen Regierung die Aktivitäten von Russland in der Ukraine beschreibt. In den Medien kann die Verwendung des Wortes „Krieg“ in diesem Kontext staatliche Sanktionen zur Folge haben. Jedoch zeigt die Analyse der Worthäufigkeit, dass oppositionelle Medien das Wort „Krieg“ häufig einsetzen. Dieses Wort wurde unter den zehn am meisten gebrauchten Wörtern aufgelistet. Ein solches Muster ist bei den kremlnahen Medien nicht zu erkennen.

Im Folgenden wird eine Analyse zur Verwendung der Wortkombination „Spezialoperation - Krieg“ in den Medien durchgeführt. Weiterhin ist der Kontext mittels NER

zu analysieren, in dem diese Wörter in Bezug auf Personen, Organisationen und Orte vorkommen.

5.2.1 Krieg

Es ist ein wichtiger Schritt in der Arbeit die Verwendung des Wortes „Krieg“ zu prüfen, ob es Unterschiede in der Verwendungshäufigkeit vor und nach dem 24. Februar 2022 erkennbar sind. In der Abbildung A.11 im Anhang ließ sich feststellen, dass die Tendenz der Verwendung des Wortes in den kremlnahen Medien über das Jahr hinweg relativ konstant blieb. Eine auffällige Ausnahme bildete einen Spitzenwert am 9. Mai 2022 bei dem Ersten Kanal. Dies ist auf den Kontext des Siegestags im Zweiten Weltkrieg zurückzuführen. Durchschnittlich wurde die Verwendungshäufigkeit des Wortes „Krieg“ vor und nach dem russischen Angriff auf die Ukraine bei dem Ersten Kanal und Ria verringert. Im prozentualen Ausdruck hat sich die Verwendungshäufigkeit beim Ersten Kanal um 6,25% und bei Ria um 28,6% gesunken. Dieser Rückgang ließ vermuten, dass kremlnahe Medien nach dem Kriegsbeginn in der Ukraine dazu tendierten, das Wort „Krieg“ weniger einzusetzen.

Die oppositionellen Medien hingegen verzeichnen einen Anstieg (siehe Abbildung A.11 im Anhang). Vor dem 24. Februar wurde das Wort „Krieg“ sehr selten verwendet. Ab dem Kriegsbeginn stieg jedoch die Verwendung signifikant: bei Mediazona um 210% und bei Meduza sogar um 1406%.

Der Unterschied im Gebrauch des Wortes „Krieg“ zwischen beiden Medientypen ist ersichtlich. Doch der Kontext des Worteinsatzes war unterschiedlich. Beispielsweise konnte der Erste Kanal das Wort im Sinne eines „Wirtschaftskrieges gegen Russland“ nutzen, während Meduza eher in einem Kontext wie „Der Krieg in der Ukraine könnte ein ganzes Jahr dauern“ davon Gebrauch machte. Auch wenn beide Medien das Wort „Krieg“ verwendet haben, hieße das nicht zwangsläufig, dass sie sich auf den ukrainischen Konflikt bezogen. Und selbst dann blieb es unklar, in welchem Zusammenhang das Wort genutzt wurde.

Um den Kontext genauer zu beleuchten, wurden Wortwolken generiert. Dabei wurden mithilfe von NER-Methode jene Organisationen, Personen und Orte identifiziert, die im gleichen Satz wie „Krieg“ aufgetreten sind. Um den Überblick zu erleichtern, wurden die jeweils 20 am häufigsten auftretenden Entitäten pro Medium ausgewählt.

In den Abbildungen 5.6, 5.7, 5.8 und 5.9 sind die Wortwolken dargestellt, die im Zusammenhang mit dem Wort „Krieg“ in verschiedenen Medien erschienen. Die Analyse zeigte gemeinsame Entitäten, wie „Ukraine“, „Russland“, „Putin“, „USA“ und „Kiew“ über alle Medien. Angesichts der globalen politischen Ereignisse im Jahr 2022 war dies zu erwarten. Es gab jedoch auch spezifische Entitäten, die in kremlnahen Medien dominierten,

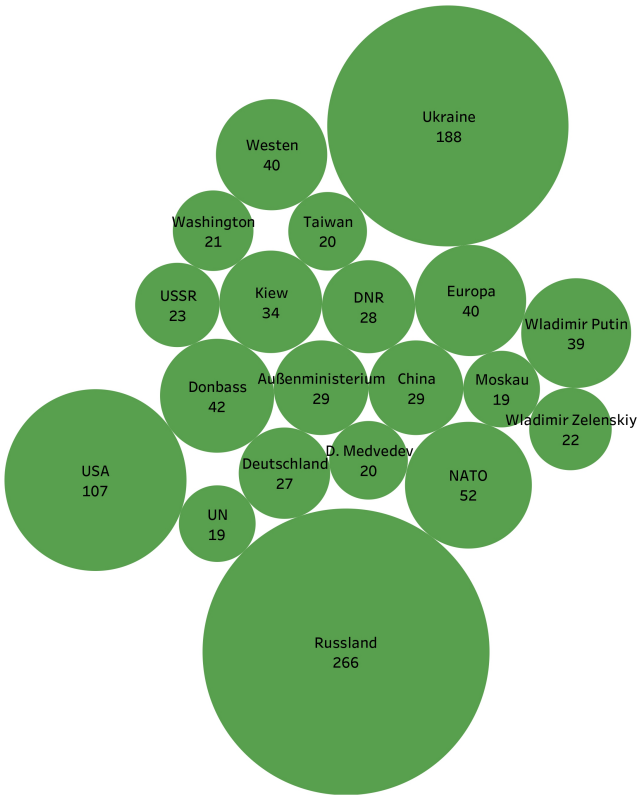


Abbildung 5.6: Wortwolke des Wortes „Krieg“ bei Ria im Jahr 2022

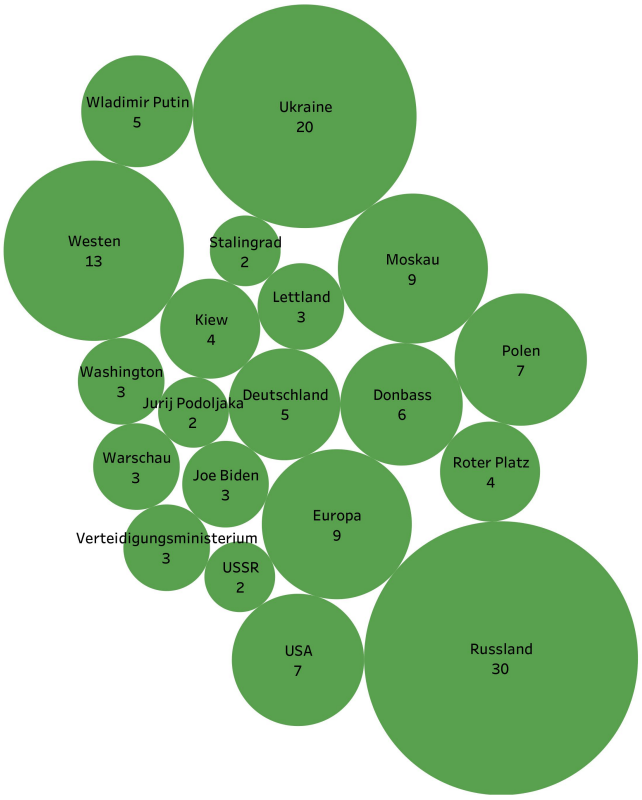


Abbildung 5.7: Wortwolke des Wortes „Krieg“ beim Ersten Kanal im Jahr 2022

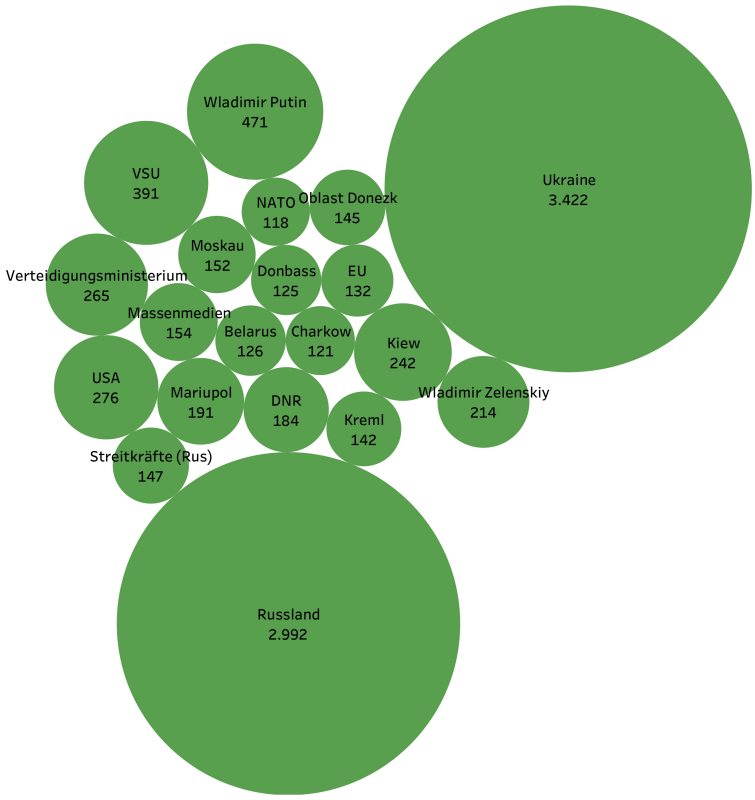


Abbildung 5.8: Wortwolke des Wortes „Krieg“ bei Meduza im Jahr 2022

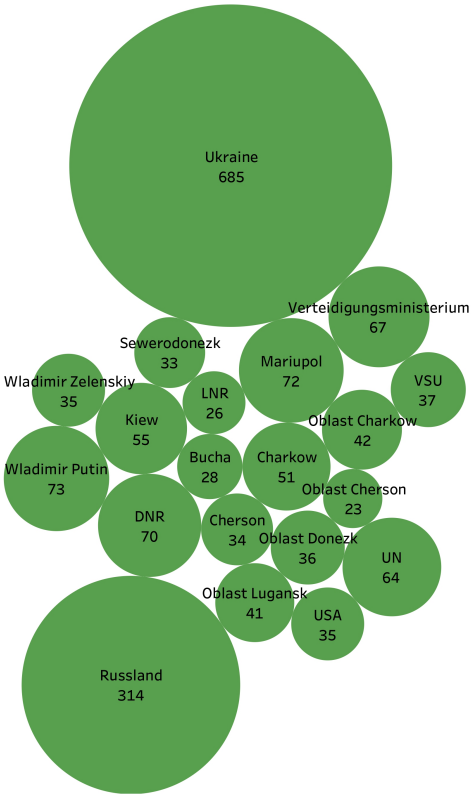


Abbildung 5.9: Wortwolke des Wortes „Krieg“ bei Mediazona im Jahr 2022

darunter: „Donbass“, „Westen“, „USSR“, „Deutschland“ und „Moskau“. Im Gegensatz dazu erwähnten oppositionelle Medien öfter Wörter wie „Ukrainische Streitkräfte (VSU)“, „Volksrepublik Donezk (DNR)“, „Charkow“ und „Mariupol“.

Es war für russische kremlnahen Medien bezeichnend, Gebiete wie Volksrepublik Donezk (DNR) und Volksrepublik Lugansk (LNR) als „Donbass“ zu bezeichnen und Europa mit USA pauschal als „Westen“ zu benennen. Auf der anderen Seite haben oppositionelle Medien den Schwerpunkt auf bestimmte Ereignisse in der Ukraine gesetzt. Aus diesem Grund waren die Entitäten bestimmter ukrainischer Städte und Regionen bei Meduza und Mediazona zu erkennen. Die Unterschiede in den Benachrichtigungen sind ersichtlich. Während oppositionelle Medien detailliert über die Vorgänge in der Ukraine informierten, betonten kremlnahe Medien die internationale Beteiligung in den Ereignissen.

In dieser Analysephase war zu erkennen, dass kremlnahe und oppositionelle Medien unterschiedliche Bezeichnungen zum Staat Belarus verwenden. Während kremlnahe Medien die traditionelle Bezeichnung „Weißrussland“ aus sowjetischen Zeiten bevorzugten, verwendeten oppositionelle Medien den neueren Begriff „Belarus“, der seit dem Ende der Sowjetunion gebräuchlich ist.

Eine wichtige Bemerkung ist es, dass Meduza und Mediazona häufig andere Publikationen darunter auch häufig aus Ria zitieren. Praktisch gilt Ria als eines der meistzitierten Medien in Russland. Dieser Fakt könnte der Grund dafür sein, dass in den Berichterstattungen der oppositionellen Medien die Wörter zu treffen sind, die typischerweise mit Propaganda assoziiert werden, wie „Donbass“ (anstelle von Volksrepublik Donezk (DNR) und Volksrepublik Lugansk (LNR)) oder „Weißrussland“ (anstelle von Belarus).

5.2.2 Spezialoperation

In Bezug auf die russischen Aktionen in der Ukraine wurde der Begriff „Spezialoperation“ erstmals am 24. Februar 2022 verwendet, als der russische Präsident Wladimir Putin von einer „speziellen Militäroperation“ sprach. Vor diesem Zeitpunkt nutzten sämtliche Medien das Wort hauptsächlich im Kontext von Anti-Terror-Einsätzen, insbesondere in Bezug auf Vorfälle in Kasachstan im Januar 2022.

Das Wort „Spezialoperation“ kam in den Medien in verschiedenen Formulierungen vor: „SWO“ (das Akronym für „spezielle Kriegsoperation“"), „spezielle Militäroperation“, „spezielle Operation“ und schlicht „Spezialoperation“. Dies erschwerte die Identifizierung von Begriffen und Phrasen. Ferner wurde das Akronym SWO oft fehlerhaft vom NER-System als eine Organisation identifiziert, was zusätzliche Bereinigungen in der Datenaufbereitung erforderte.

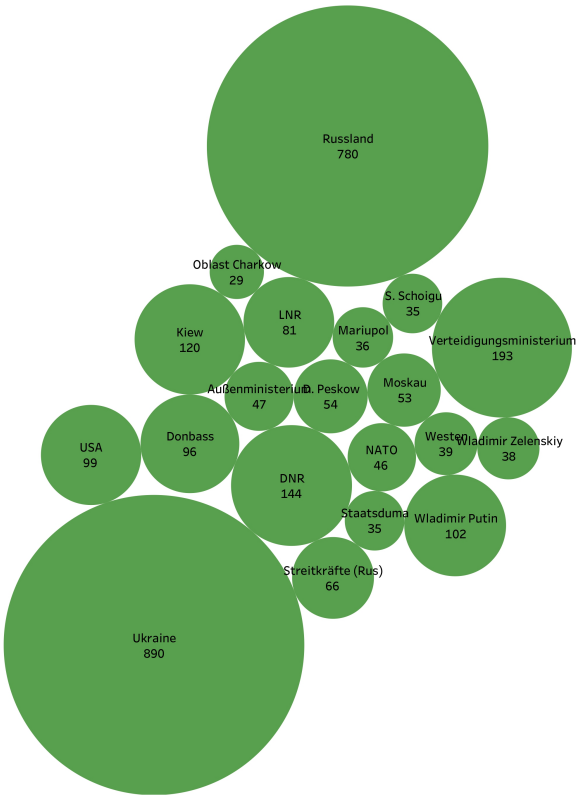


Abbildung 5.10: Wortwolke des Wortes „Spezialoperation“ bei Ria im Jahr 2022

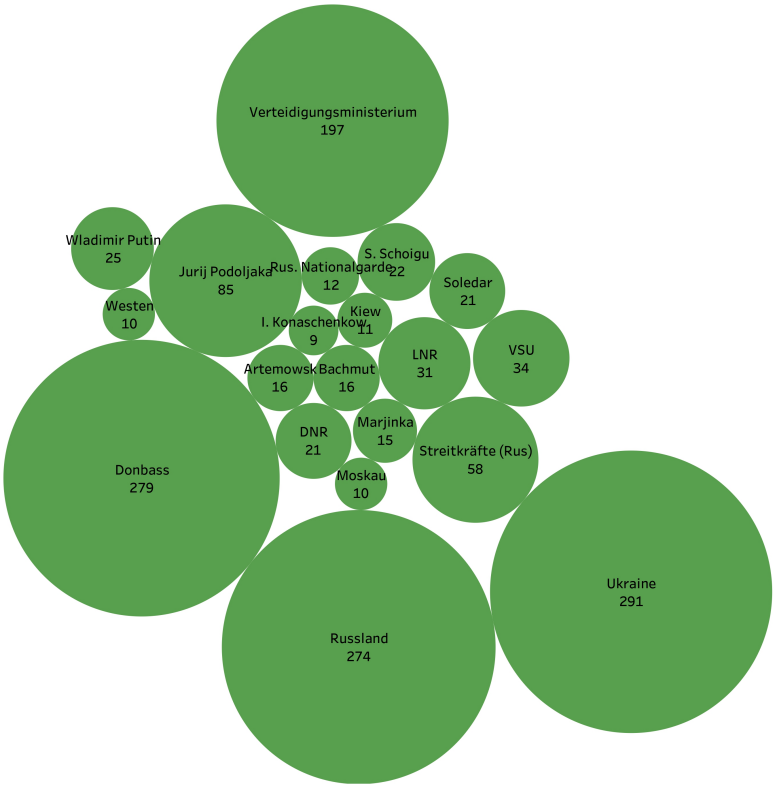


Abbildung 5.11: Wortwolke des Wortes „Spezialoperation“ beim Ersten Kanal im Jahr 2022

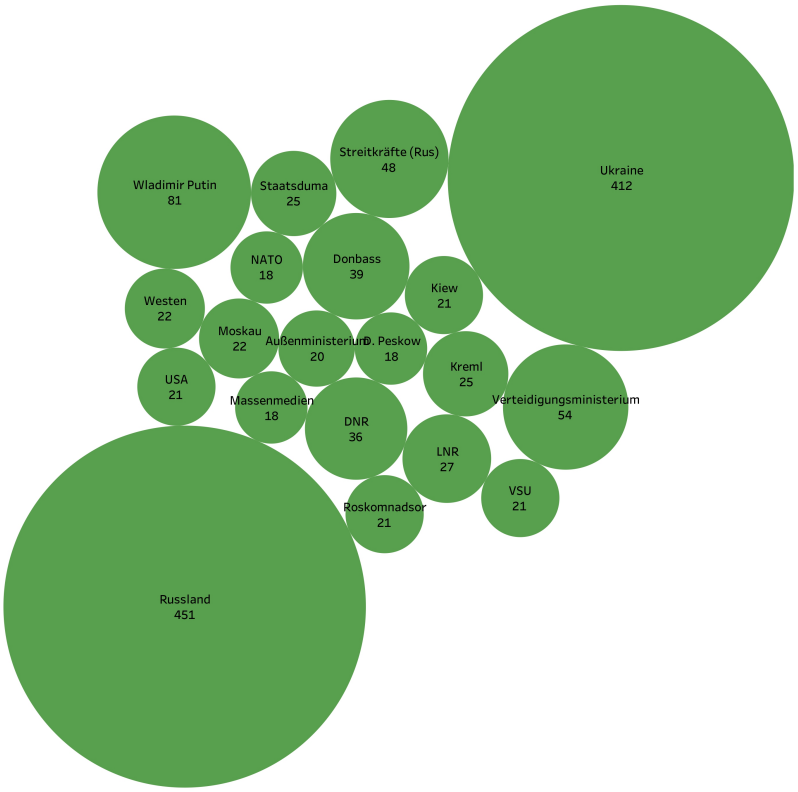


Abbildung 5.12: Wortwolke des Wortes „Spezialoperation“ bei Meduza im Jahr 2022

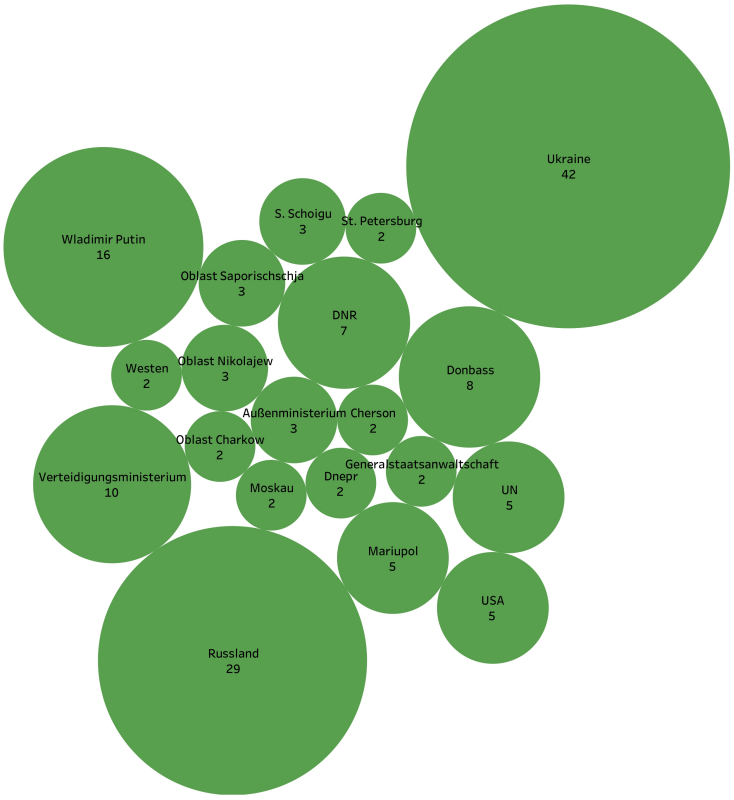


Abbildung 5.13: Wortwolke des Wortes „Spezialoperation“ bei Mediazona im Jahr 2022

Nach dem 24. Februar 2022 verzeichneten alle untersuchten Medien einen signifikanten Anstieg in der Verwendung des Wortes „Spezialoperation“ und ähnlicher Formulierungen. So hat der Erste Kanal einen Zuwachs von 120%, Ria von 213%, Mediazona von 130% und Meduza von 180% gezeigt (siehe Abbildung A.17 im Anhang). Es ließ sich jedoch nicht simpel unterscheiden, ob oppositionelle Medien das Wort in Zitaten oder in eigenständig verfassten Artikeln nutzten. Wie bereits bei der Analyse des Wortes „Krieg“ hilfte eine Betrachtung des genauen Kontextes, in dem die Wörter getroffen wurden.

Bei der Analyse des Kontexts des Wortes „Spezialoperation“ wurden Wortwolken erstellt (siehe Abbildungen 5.10, 5.11, 5.12 und 5.13). Eine NER-basierte Untersuchung ergab, dass der Kontext des Wortes „Krieg“ je nach Medienkategorie variierte. Für „Spezialoperation“ wurden jedoch erhebliche Überschneidungen zwischen den beiden Medienkategorien festgestellt. Zu den häufigst identifizierten Entitätem gehörten „Russland“, „Ukraine“, „Donbass“, „Verteidigungsministerium“, „Volksrepublik Donezk (DNR)“, „Putin“, „Westen“ und „Moskau“. Bei pro-kremlischen Medien standen „Volksrepublik Lugansk (LNR)“, „Kiew“ und „Schoigu“ im Fokus, während für oppositionelle Medien die „USA“ und das „Außenministerium“ identifiziert wurden. Die Vielzahl an gemeinsamen Entitäten deutete darauf hin, dass das Wort in beiden Medienkategorien in einem ähnlichen Kontext möglicherweise durch Zitate aus anderen Quellen verwendet wurde.

Es wurde angenommen, dass die ermittelten Entitäten auf Ereignisse in der Ukraine hingewiesen haben. Daraus lässt sich schließen, dass das Wort „Spezialoperation“ im Zusammenhang mit dem russischen Angriff auf der Ukraine verwendet wurde.

5.2.3 Diskussion

Bevor die Wortpaare „Krieg-Spezialoperation“ analysiert wurde, wurde angenommen, dass die Häufigkeit des Auftretens des Wortes „Krieg“ in oppositionellen Medien direkt mit dem Wort „Spezialoperation“ in pro-kremlischen Medien im Zusammenhang sein konnte. Eine solche Korrelation ließ sich jedoch nicht feststellen.

Es zeigte sich, dass das Wort „Krieg“ in pro-kremlischen Medien seltener bis auf überhaupt nicht auftreten wurde im Vergleich zu oppositionellen. Ein paralleler Rückgang bei Ria und dem Ersten Kanal stellt ein direkte Abhängigkeit dar. Mögliche Gründe könnten eine gestiegene Selbstzensur der Medien aufgrund möglicher staatlicher Sanktionen sein oder direktive Anweisungen von Regierung. In jedem Szenario deutet dies auf eine Form der Zensur hin.

Es war zu erwarten, dass das Wort „Spezialoperation“ nach russischem Angriff auf der Ukraine häufiger zu treffen war. Vor dem Beginn der Untersuchung wurde vermutet, dass der Anstieg der Worterwähnungen in oppositionellen Medien relativ mittelmäßig bleiben soll. Die Ergebnisse der Analyse haben jedoch diese Annahme nicht bestätigt.

Eine Erklärung könnte hierfür das wiederholte Zitieren aus pro-kremlischen Quellen sein.

Basierend auf den Ergebnissen und dem Verständnis der Neusprech-Praktiken in pro-kremlischen Medien wurde festgestellt, dass das Wort „Krieg“ in diesen Medien gezielt durch „Spezialoperation“ ersetzt wurde. In oppositionellen Medien stiegen die Erwähnungen beider Wörter seit Kriegsbeginn in der Ukraine wesentlich an. Dabei nahm die Verwendung des Wortes „Krieg“ um das Zwei- bis Vierzehnfache (in Mediazona bzw. Meduza) zu, während die Steigerung des Wortes „Spezialoperation“ sich lediglich auf das 1,5-Fache erhöht hat.

5.3 Untersuchung der Zusammenhänge zwischen Entitäten

In der abschließenden Analysephase wurden Graphen generiert, welche die Zusammenhänge zwischen Personen, Organisationen und Orten darstellen. Wenn Entitäten im gleichen Kontext oder Satz erscheinen, wird angenommen, dass sie in einer gewissen Form miteinander verbunden sind. Ein beispielhafter Satz wäre: „Im Westen bevorzugt man es, nicht über die Verbrechen Kiews zu sprechen, aber die Wahrheit bricht dennoch durch“ [13]. Im Kontext dieses Satzes identifiziert das NER-Verfahren „Westen“ und „Kiew“ als geografische Entitäten. Dementsprechend werden sie im Graphen als Knoten repräsentiert, wobei ihre Verbindung durch eine Kante dargestellt wird.

5.3.1 Graphenvorbereitung

Wie bereits im Kapitel 4 beschrieben, wurden die Ergebnisse in zwei Dateien, nämlich „Knoten“ und „Kanten“ übertragen. Diese Dateien wurden dann in die Gephi-Software importiert. Das erste Visualisierungsergebnis zeigte eine Punktwolke, wie in der Abbildung 5.14 zu sehen ist. Nach Aktivierung der Kanten in der Software verdichtete sich diese Darstellung zu einem nahezu undurchsichtigen Quadrat, wie es in der Abbildung 5.15 dargestellt ist.

Zur Optimierung der visuellen Darstellung des Graphen wurde der Layout-Algorithmus Force Atlas 2 eingesetzt. Dieser Algorithmus optimierte den Graph, sodass die nicht verbundenen Knoten voneinander weit entfernt und verbundene Knoten angezogen wurden. Zusätzlich wurde eine Überlappung der Knoten minimiert. Das optimierte Datengraph nach der Anwendung dieses Algorithmus ist in der Abbildung 5.16 gezeigt.

Mit dem nächsten Schritt erfolgte eine farbliche Rangordnung der Knoten in Abhängigkeit vom Grad: hellere Knoten weisen einen niedrigeren Grad auf, dunklere Knoten repräsentieren einen höheren Grad. Wie in der Abbildung 5.17 zu sehen ist, fallen zwei

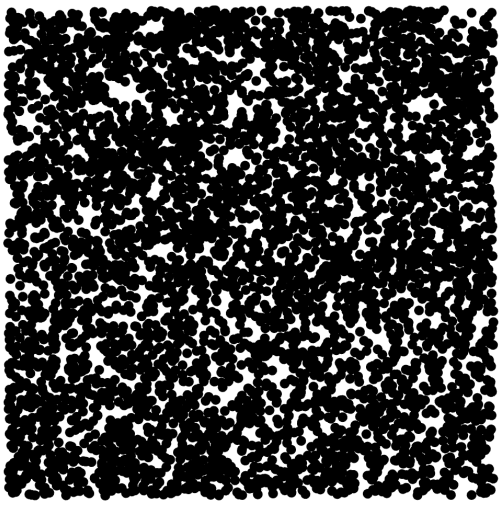


Abbildung 5.14: Datengraph ohne Kanten, Darstellung in Gephi

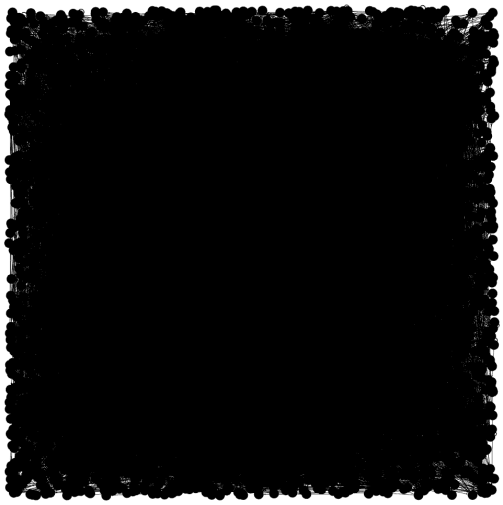


Abbildung 5.15: Datengraph mit Kanten, Darstellung in Gephi

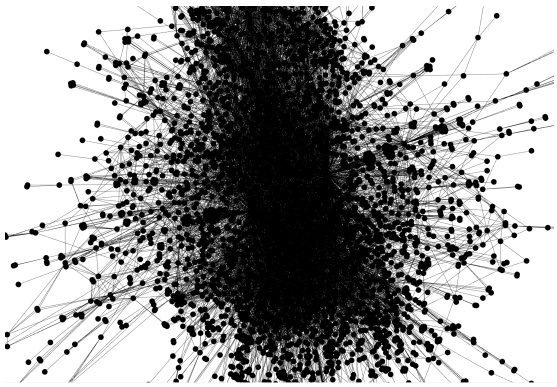


Abbildung 5.16: Datengraph nach der Anwendung von Force Atlas-2 Layout Algorithmus

Knoten besonders auf, während andere eine mittlere Intensität in ihrer Färbung zeigen.

In der generierten Darstellung des Graphen ist es jedoch wie zuvor erschwert, die wesentliche Information zu interpretieren. Aus diesem Grund wurde eine Größenrangordnung der Knoten vorgenommen. Dabei wurden Knoten mit einem höheren Grad durch eine vergrößerte Darstellung ersetzt, während solche mit einem niedrigen Grad entsprechend kleiner dargestellt wurden. Die visualisierten Ergebnisse sind in der Abbildung 5.18 abgebildet.

Obwohl der Datengraph verbessert wurde, ist er jedoch nicht in seiner optimalen Darstellungsform. Bei einer starken Vergrößerung lassen sich einzelne Knoten, die sogenannten Blattknoten, am Rand erkennen. Für die durchgeführte Analyse sind diese Knoten jedoch nicht relevant, ebenso wenig wie Kanten mit geringem Gewicht. Ein zentraler Schritt zur weiteren Optimierung ist daher die Filterung der Knoten nach ihrem Grad. Dieser Filter erlaubt es, die Knoten mit dem geringsten Grad in der Darstellung zu elimi-

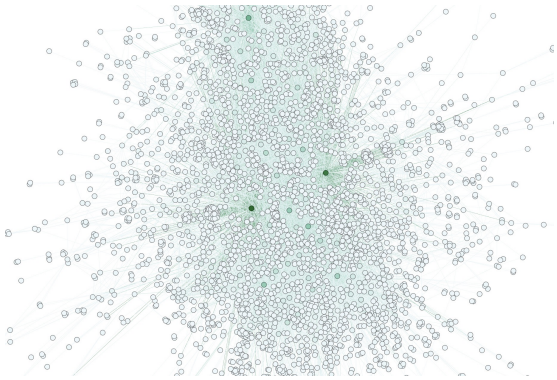


Abbildung 5.17: Datengraph mit der farblichen Rangordnung

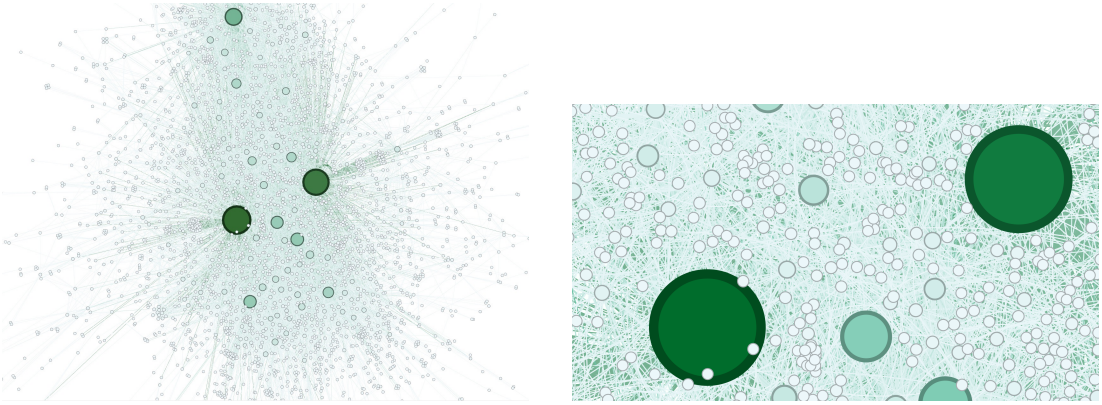


Abbildung 5.18: Datengraph mit der farblichen Größenranganordnung

nieren. So existierten im Datensatz zahlreiche Kanten mit einem Gewicht von lediglich eins oder zwei. Danach war der Graph übersichtlicher und ohne die überflüssigen Knoten (siehe Abbildung 5.19). Für die abschliessende Visualisierung des Graphen wurden Knotenbeschriftungen hinzugefügt, die im Abschnitt zur Graphenanalyse näher betrachtet werden.

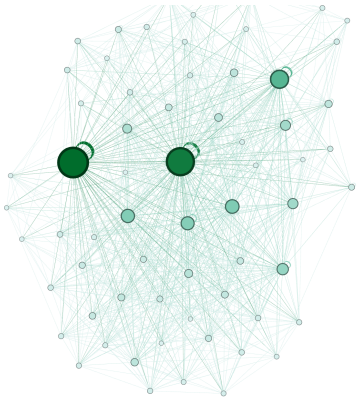


Abbildung 5.19: Datengraph nach der Anwendung vom Grad-Filter

5.3.2 Graphenanalyse

Auf Basis der analysierten Mediendaten wurde ein Graph mit 39315 Knoten und 300999 Kanten erstellt. Angesichts dieser grossen Datenmenge ist es unmöglich, alle Daten praktisch zu visualisieren. Daher war eine erweiterte Filterung der Knoten notwendig. Nach dem Filterprozess verbleiben lediglich 0,13% der ursprünglichen Knoten, was praktisch 51 Knoten entspricht. Zudem wurde die Anzahl der Kanten auf 1283 reduziert, was lediglich 0,43% der ursprünglichen Gesamtmenge darstellt, wie in der Abbildung 5.20 zu sehen ist.

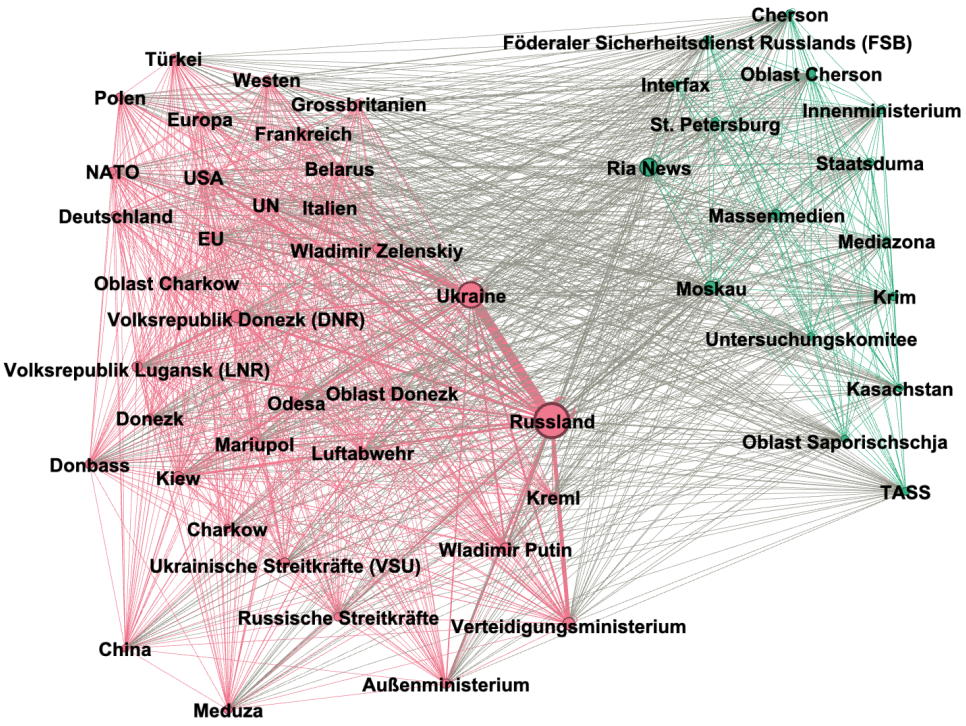


Abbildung 5.20: Datengraph auf Basis der Daten aller vier Medien

Bei der genaueren Analyse des Graphen ließ sich ein bestimmtes Muster erkennen. Sämtliche dargestellten Entitäten beziehen sich direkt oder indirekt auf die Ereignisse in der Ukraine im Jahr 2022. Wie erwartet, bilden „Russland“ und „Ukraine“ die Hauptknotenpunkte. Die Kante zwischen diesen beiden Knoten zeigt ebenfalls das stärkste Gewicht.

Die Ergebnisse der Clusterbildung sind in diesem Fall nicht optimal ausgeglichen. Von den gesamten Daten wurden 99,97% einem Cluster zugeordnet, während die restlichen 0,03% einem anderen zugehörig waren. In der Abbildung 5.20 sind diese Cluster dementsprechend durch die Farben Rot und Grün repräsentiert.

Bei der ersten Analyse des Graphen fällt eine klare inhaltliche Unterteilung der Cluster nicht sofort auf. Doch bei einer genauen Betrachtung wird ersichtlich, dass der größere

Cluster eindeutig auf die russisch-ukrainischen Beziehungen sich konzentriert. Er beinhaltet auch die Entitäten, welche die Bezeichnungen verschiedener Länder und internationaler Organisationen abbilden. In dem kleineren Cluster ist die logische Bedeutung nicht ersichtlich.

Bei der Analyse ist es wichtig, die Differenzen in den Graphen unterschiedlicher Medienquellen isoliert zu betrachten.

Basierend auf den Daten der oppositionellen Medienquelle Meduza wurde ein Graph erstellt, welcher 24762 Knoten und 166021 Kanten umfasst. In der Abbildung 5.21 ist die Visualisierung dieses Graphen nach einer Filterung dargestellt. Diese optimierte Darstellung beinhaltet 58 Knoten, was 0,23% des Gesamtvolumens entspricht, sowie 1444 Kanten, die 0,86% des Gesamtvolumens repräsentieren. Nachdem die Clusteranalyse durchgeführt wurde, haben sich zwei Cluster ergeben. Der rot gefärbte Cluster stellt 99,87% dar, während der grün gefärbte Cluster 0,13% zeigt.

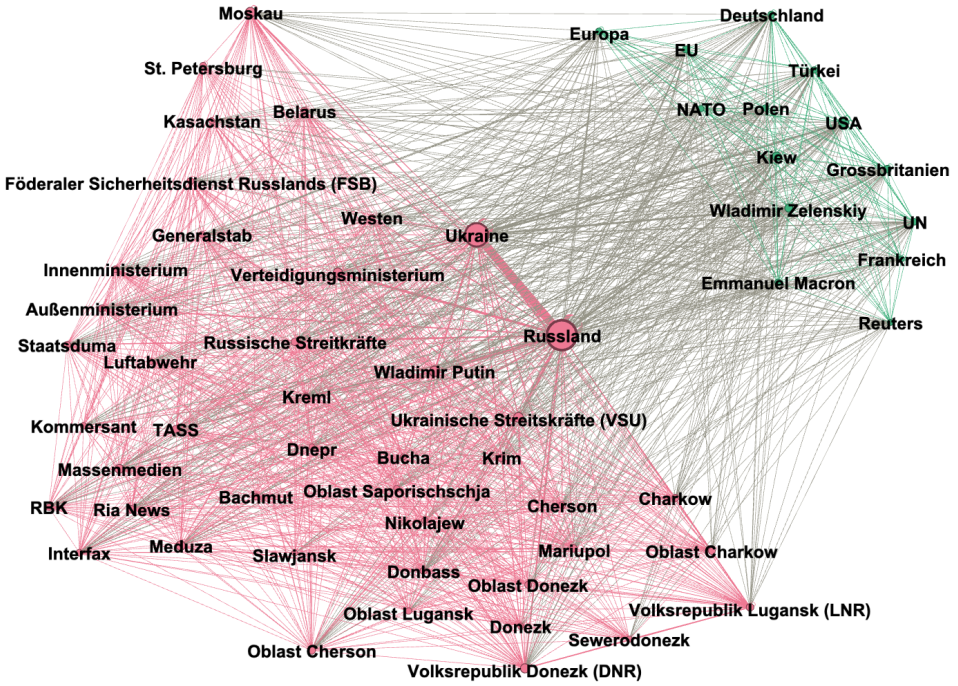


Abbildung 5.21: Datengraph auf Basis der Daten von Meduza

In dieser Visualisierung ist die logische Unterteilung zwischen den Clustern nicht sofort ersichtlich. Dennoch zeigt der grössere rote Cluster vorwiegend Entitäten, die im Kontext des Kriegsereignisses in der Ukraine im Jahr 2022 relevant sind, insbesondere Städte und Regionen, welche die Hauptorte des Kampfs waren. Der kleinere grüne Cluster hingegen beinhaltet primär Entitäten, die in Bezug zur außenpolitischen Beteiligung stehen, darunter Länder, Personen und Organisationen, welche sich zum Konflikt auf internationaler Ebene äußern.

Auffällig dabei ist, dass die Ria News unter den häufig vorkommenden Entitäten zu finden war. Dies spiegelt die im vorherigen Analysephase aufgestellte Hypothese bezüglich der häufigen Zitate von Ria durch Meduza wider.

Der nächste Graph basiert auf den Daten des oppositionellen Mediums Mediazona. Dies ist in der Abbildung 5.22 dargestellt. Ursprünglich fasste der Graph 5274 Knoten und 31709 Kanten um. Nach der Anwendung des Filters verblieben 43 Knoten (entspricht 0,82%) und 752 Kanten (entspricht 2,37%) im Graph. Die Clusteranalyse hat zwei Cluster erkannt: einen rot gefärbten Cluster mit einem Anteil von 99,54% und einen grün gefärbten Cluster mit dem 0,46%.

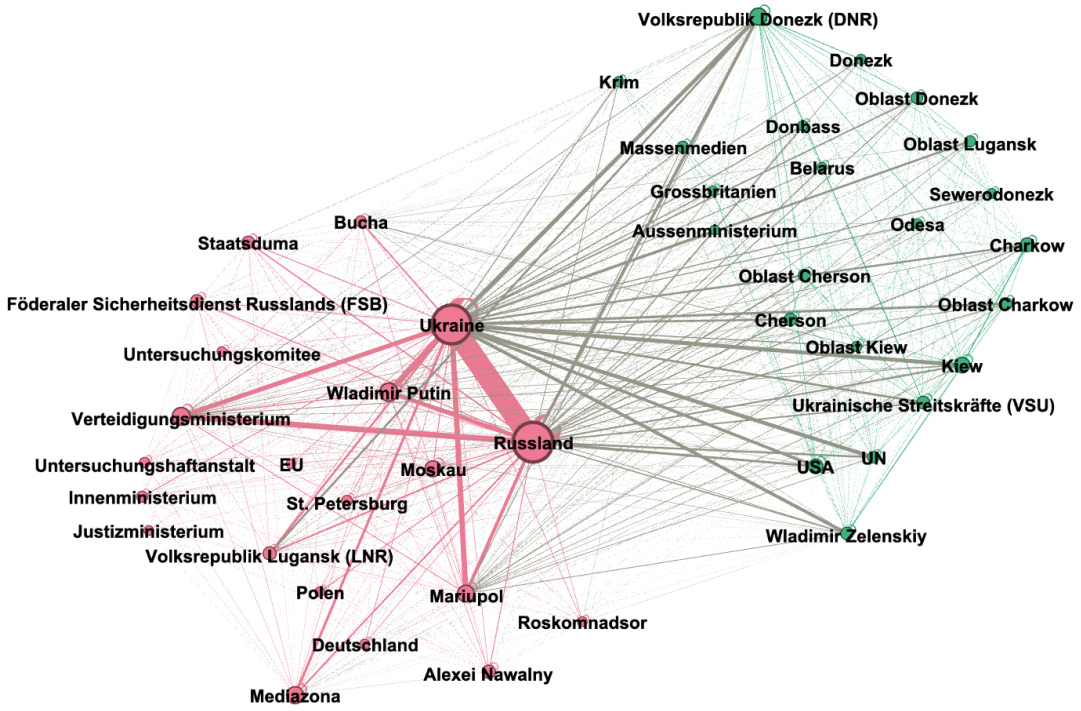


Abbildung 5.22: Datengraph auf Basis der Daten von Mediazona

In diesem Graphen lässt sich keine klare logische Unterscheidung zwischen den Clustern feststellen. Beide Cluster beinhalten Entitäten, die sowohl außenpolitischen Beziehungen als auch den militärischen Auseinandersetzungen entsprechen.

Mediazona ist das einzige untersuchte Medium, in welchem der Name des politischen Aktivisten Alexei Nawalny als Knoten vorkommt. Im Hinblick auf die erhebliche Berichterstattung über die Ukraine im Jahr 2022 ist es auffällig, dass Namen von innenpolitischen Oppositionsfiguren in den zentralen Entitäten erscheinen. Eine solche Erscheinung erfordert besondere Betonung.

Für die pro-kremlischen Medien wurden ebenfalls eigenständige Graphen generiert. Die in der Abbildung 5.23 dargestellte Visualisierung repräsentiert den Graphen der Nachrichtenagentur Ria. Ursprünglich fasste dieser Graph 17135 Knoten und 126916 Kanten

um. Nach der Filterung verbleiben 55 Knoten, was 0,32% des Ursprungswertes entspricht, und 1383 Kanten, was 1,09% des ursprünglichen Wertes abbildet. Die daraufhin durchgeführte Clusteranalyse identifizierte zwei Hauptcluster: einen rot markierter Cluster mit einem Anteil von 99,94% und einen in grün dargestellten Cluster, der 0,06% repräsentiert.

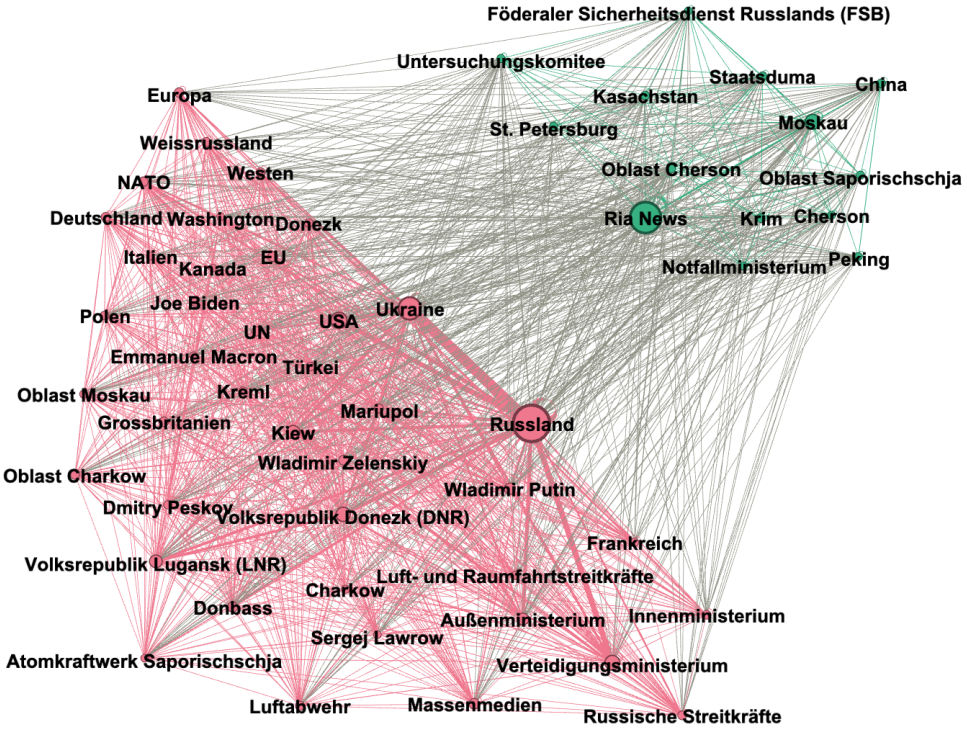


Abbildung 5.23: Datengraph auf Basis der Daten von Ria

In diesem Graphen lässt sich keine klare logische Unterscheidung zwischen den Clustern feststellen.

Im größeren rot markierten Cluster sind sowie die Länder und Organisationen, die sich über den Krieg in der Ukraine in den Medien äußern, als auch die Entitäten, die russisch-ukrainische Beziehung beschreiben. In dem kleineren Cluster ist die logische Bedeutung nicht ersichtlich.

Der abschließend analysierte Graph basiert auf den Nachrichten des Ersten Kanals. Der gefilterte Graph fasst 45 Knoten um, was 1,07% der ursprünglichen 4211 Knoten entspricht, sowie 599 Kanten, was 4,6% der ursprünglichen 13126 Kanten präsentiert. Der gefilterte Graph ist in der Abbildung 5.24 dargestellt. Im Zuge der Clusteranalyse wurden drei Cluster herausgestellt: ein lila gefärbter Cluster mit einem Anteil von 99,29%, ein orange gefärbter Cluster mit 0,52% und ein grün gefärbter Cluster mit 0,19%.

Der Erste Kanal stellt das einzige untersuchte Medium dar, in dem drei Cluster aufgefasst wurden. Bei genauer Untersuchung des größten, violett gefärbten Clusters wird es ersichtlich, dass er primär die Kriegsentwicklung in der Ukraine abbildet.

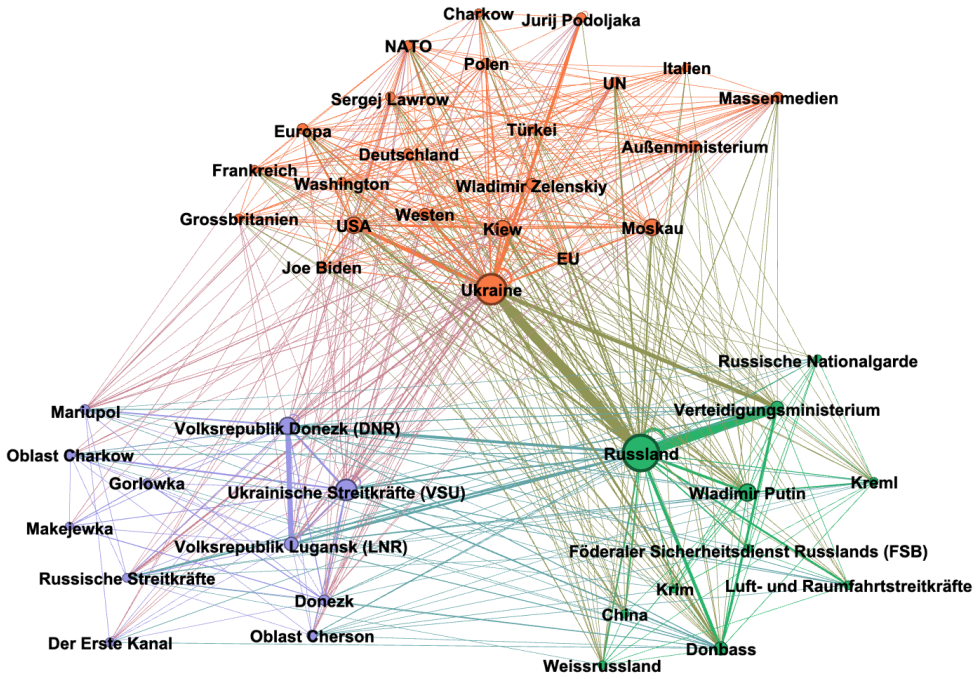


Abbildung 5.24: Datengraph auf Basis der Daten vom Ersten Kanal

Die Interpretation der orangefarbenen und violetten Cluster stellt sich als komplex heraus. Es könnte angenommen werden, dass der orangefarbene Cluster überwiegend der außenpolitischen Dimension enspricht. Hinsichtlich des kleinsten grünen Clusters, ist eine deutliche Mustererkennung nicht ersichtlich. Er besteht primär aus der russischen Behörden.

Jedoch fällt auf, dass nur in diesem Medium das Gewicht der Kante „Russland - Ukraine“ ähnlich stark repräsentiert wird, wie der Kante „Russland - Verteidigungsministerium“. Dies unterstreicht erneut die politische und militärische Thematik in den Medienartikeln.

Zum Schluss, zeigt die Abschätzung der Graphen einzelner Medien, dass sich bei der Clusterbildung lediglich ein großer Cluster heraustellt. Dies zeigt eine starke Vernetzung der Knoten und deutet auf ein Hauptthema in den betrachteten Texten hin. Es lässt sich jedoch nicht feststellen, dass die kleineren Cluster deutlich vom Hauptcluster abgegrenzt werden können. Eine Vielzahl von Verbindungen zwischen dem Hauptcluster und den kleineren Clustern ist ersichtlich. Das deutet darauf hin, dass die in den kleineren Clustern besprochenen Themen sich nicht vom zentralen Thema trennen lassen. Zudem weist dies auf eine übergreifende thematische Verknüpfung hin.

Die Untersuchung der Graphen macht die variierte Benennungen des Landes Belarus ersichtlich. In den oppositionellen Medien wird die Bezeichnung „Belarus“ verwendet, während in den kremlnahen Medien „Weissrussland“ zu finden ist. Des Weiteren bestätigen die Graphen einige der bereits in davorstehenden Abschnitten identifizierten

Beobachtungen, insbesondere die Häufigkeit, mit der Meduza die Ria-Benachrichtigungen zitiert.

Es ist jedoch zu betonen, dass die Auswertung der Graphen aus der Hinsicht der Informatik zu interpretieren war. Die politische Analyse ließ sich nur begrenzt basierend auf dem Wissenstand der Autorin ansatzweise realisieren. Für eine tiefergehende Analyse und Bewertung der dargestellten Beziehungen wäre ein umfassenderes Verständnis der Politikwissenschaft erforderlich.

5.4 Diskussion

Im Laufe des Kapitels erfolgte eine gründliche Untersuchung der Daten, sowie eine Vertiefung in die Analyse. Hierbei wurde jede weitere Analysephase genutzt, um zuvor aufgestellte Hypothesen zu überprüfen.

Die vorgenommene Analyse verdeutlicht wesentliche Unterschiede in den Artikeln von pro-kremlischen und oppositionellen Medien. Des Weiteren lassen sich Unterschiede in der Artikelmenge erkennen, insbesondere in Bezug auf die Tage vor, während und nach dem Beginn des Krieges. Durch die Analyse der häufigsten Wörter sind bestimmte Muster erkennbar, welche die beiden Medienkategorien voneinander differenzieren. Zu bemerken ist es, dass die Ergebnisse der oppositionellen Medien maßgeblich durch ihre Einstufung als „ausländischer Agent“ beeinflusst sind.

Die zweite Analysephase ermöglichte eine tiefgründige Einsicht auf die Wörter „Krieg“ und „Spezialoperation“, sowie deren spezifische Verwendung in verschiedenen Medienkontexten. Bei den pro-kremlischen Medien ließ sich ein Rückgang der Nutzung des Wortes „Krieg“ feststellen, während oppositionelle Medien diesen im Jahr 2022 vermehrt eingesetzt haben. Zu betonen ist es, dass die Verwendung des Wortes „Spezialoperation“ bei sämtlichen Medien zugenommen hat. Diese Beobachtung machte eine detaillierte Betrachtung des Kontexts in dieser Analysephase notwendig.

Im abschließenden Abschnitt der Analyse stand die Erstellung von Graphen im Fokus. Diese Graphen legten keine spezifischen Muster für die jeweiligen Mediengruppen fest. Allerdings lieferten sie vertiefte Erkenntnisse über die beschriebenen Zusammenhänge in den einzelnen Medien.

6 Fazit

6.1 Zusammenfassung

Diese Bachelorarbeit befasste sich mit der Untersuchung der Veränderungen im Verhalten der pro-kremlischen und oppositionellen Medien im Jahr 2022 in Bezug auf den Kriegsbeginn in der Ukraine zu untersuchen. In der Vorbereitungsphase wurden verschiedene Ansätze der natürlichen Sprachverarbeitung angesprochen, darunter klassische, empirische und statistische Methoden. Um ein fundiertes Verständnis von NLP zu gewinnen, wurden auf führende Sprachmodelle wie BERT und GPT inklusive ihrer jeweiligen Architekturen eingegangen. Ein wichtiges Element dieser Arbeit war die Betrachtung der Named Entity Recognition (NER), da diese Methode maßgeblich für die Analyse angewandt wurde.

In dem Daten-Abschnitt der Bachelorarbeit lag der Fokus auf der Datensammlung. Um eine möglichst optimale Auswahl der Datenquellen für die Artikelanalyse treffen zu können, war es erforderlich, die spezifischen Auswirkungen der neuen repressiven Gesetzesinitiativen zu verstehen. Insbesondere die Einstufung als „ausländischer Agent“ im Jahr 2022 sowie das Gesetz über Falschinformationen, beeinflussten die Medienverlage maßgeblich. Die Datensammlung stellte sich aufgrund der beschränkten Möglichkeiten der verfügbaren APIs als anspruchsvoll heraus. Die Sammlung, Bereinigung und Übersetzung der Daten beanspruchte einen erheblichen Zeitaufwand im Rahmen dieser Arbeit.

Die Analyse der russischen Sprache stellte sich als eine komplexe Aufgabe dar. Insbesondere hat es sich durch den Mangel an spezialisierten Python-Bibliotheken gezeigt, die effizient mit der russischen Sprache umgehen können. Mit der ausgewählten Natasha-Bibliothek konnte jedoch eine umfassende Analyse der erfassten Texte durchgeführt werden. Die Ergebnisse dieser Untersuchung dienten später als Grundlage für die Visualisierungen in den Programmen Tableau und Gephi.

Das Ziel dieser Arbeit war es, fundierte Visualisierungen zu erstellen, welche eine detaillierte Auswertung ermöglichen. Dabei zeigte es sich, dass die mehrstufige Datenauswertung für das Jahr 2022 besonders wichtig war. Jede der drei Analysephasen baute auf der vorherigen auf und eröffnete neue Perspektiven zur Datenauswertung. Es war zu erwarten, dass die Berichterstattung der pro-kremlischen sich von den oppositionellen

Medien unterscheiden würde. Die Analyse deckte verschiedene Muster auf, etwa in Bezug auf die Anzahl der veröffentlichten Artikel oder die Häufigkeit ausgewählter Wörter wie „Krieg“ und „Spezialroperation“. Besonders hervorzuheben war es, dass auch innerhalb der beiden Medienkategorien ähnliche Verhältnisse sich in der Sprachverwendung zeigten. Ein eindeutiges Beispiel dafür ist die unterschiedliche Benennung des Staates Belarus. Während die pro-kremlische Medien häufig „Weißrussland“ verwenden, nutzen oppositionelle Medien die Schreibweise „Belarus“. Das hat ebenfalls die Graphenanalyse bestätigt.

6.2 Ausblick

Im Laufe der Untersuchung zeigte sich, dass die Interpretation der Visualisierungen maßgeblich durch die Kenntnisse des Verfassers über die russische politische Situation und die militärischen Entwicklungen in der Ukraine beschränkt ist. Während für die Identifikation grundlegender Muster das allgemeine Verständnis ausreichend war, erforderte die vertiefte Analyse und die Bewertung der dargestellten Zusammenhänge ein fundierteres Wissen im Bereich der Politikwissenschaft.

In der durchgeführten Analyse haben sich weitere Entwicklungspotenziale gezeigt. Neben offensichtlichen Propagandabegriffen, wie „Donbass“, „Weißrussland“ und „Westen“ existieren subtilere sprachliche Feinheiten. Dazu gehören Ausdrücke wie „in der Ukraine“ gegenüber „auf der Ukraine“. Der Ausdruck „auf der Ukraine“ wird vorwiegend von pro-kremlischen Medien genutzt und stammt aus der USSR. In der russischen Sprache bedeutet „auf der Ukraine“ eine Peripherie, eine Art „am Rand“ und vermittelt eine Nebensächlichkeit. Im Gegensatz dazu steht „in der Ukraine“, welches als „im Land“ interpretiert wird und somit die Eigenständigkeit des Landes betont.

Eine Sentimentanalyse könnte zusätzliche Erkenntnisse in die Medienberichterstattung liefern. Während der Untersuchung fiel auf, dass der Erste Kanal vor der Ankündigung der Mobilisierung in Russland eine auffallend aggressive Rhetorik verwendete. Eine solche Sprachgebung ist für übliche journalistische Berichterstattung außerordentlich. Daher könnte es informationsreich sein, eine weiterführende Analyse in diese Richtung zu verfolgen. Insbesondere wäre es bedeutend, festzustellen, ob es signifikante Unterschiede in dem Sentiment zwischen dem pro-kremlischen und oppositionellen Medien gibt oder ob Gemeinsamkeiten in der Ausdrucksweise erkennbar sind.

Ein weiterer Ansatz wäre die Entwicklung eines Sprachmodells, das darauf spezialisiert ist, Artikel im charakteristischen Stil der vier untersuchten Medien zu generieren. Dies würde nicht nur das Verständnis für den individuellen Charakter und Stil jedes Mediums vertiefen, sondern auch die Feinheiten und Details in ihrer Berichterstattung aufdecken. Ein Modell könnte auf spezielle Redewendungen und typische Worten antrainiert werden, um Artikel im Stil der untersuchten Medien zu schreiben. Dies würde helfen zu

verstehen, wie diese Medien eine Sprache nutzen, um bestimmte Nachrichten zu senden oder Ansichten hervorzuheben.

Literaturverzeichnis

[1] : *Appellentscheidung der Zivilkammer des Moskauer Stadtsgerichts vom 20.06.2022*

[2] : *Artikel 3 Absatz 6 des Gesetzes der Russischen Föderation "Über die Massenmedien"*

[3] : *Bundesgesetz „Über die Einführung von Änderungen in das Strafgesetzbuch der Russischen Föderation und die Artikel 31 und 151 der Strafprozessordnung der Russischen Föderation“*

[4] : *Entscheidung des Bezirksgerichts Tverskoy, Moskau, vom 21.03.2022*

[5] : *Gephi: Offizielle Web-Seite.* – URL https://gephi.org/. – Zugriffsdatum: 07-09-2023

[6] : *Liste der ausländischen und internationalen Nichtregierungsorganisationen, deren Tätigkeit auf dem Territorium der Russischen Föderation als unerwünscht eingestuft wird*

[7] : *Liste der gesellschaftlichen Vereinigungen und religiösen Organisationen, gegen die ein rechtskräftiges Gerichtsurteil zur Auflösung oder zur Verbot der Tätigkeit aus den im Bundesgesetz vom 25.07.2002 Nr. 114-FZ „Über die Bekämpfung extremistischer Aktivitäten“ vorgesehenen Gründen erlassen wurde*

[8] : *Natasha Git-Repository.* – URL https://github.com/natasha/natasha. – Zugriffsdatum: 07-09-2023

[9] : *Natasha-SpaCy model proposal and acceptance.* – URL https://github.com/natasha/natasha-spacy. – Zugriffsdatum: 07-09-2023

[10] : *Offizielle Telegram-Kanal: Mediazona.* – URL https://t.me/mediazzzona. – Zugriffsdatum: 19-08-2023

[11] : *Offizielle Telegram-Kanal: Meduza.* – URL https://t.me/meduzalive. – Zugriffsdatum: 19-08-2023

[12] : *Offizielle Telegram-Kanal: RIA News.* – URL https://t.me/rian_ru. – Zugriffsdatum: 19-08-2023

[13] : *Offizielle VKontakte-Seite: Nachrichten des Ersten Kanals.* – URL https://vk.com/1tvnews. – Zugriffsdatum: 10-08-2023

[14] : *Pandas documentation.* – URL https://pandas.pydata.org/docs/. – Zugriffsdatum: 02-09-2023

[15] : *Tableau: Offizielle Web-Seite.* – URL https://www.tableau.com/de-de/why-tableau/what-is-tableau. – Zugriffsdatum: 07-09-2023

[16] Aronoff, Mark ; Fudeman, Kirsten: What is morphology? (2004), 01

[17] BBC News: *„Rain“ und „Important Stories“ gelten als ausländische Agenten. Dies stellt ein Risiko für YouTube-Blogger dar.* 20 August 2021. – URL https://www.bbc.com/russian/news-58285668. – Zugriffsdatum: 28-06-2023

[18] Beesley, Kenneth ; Karttunen, Lauri: Finite-State Morphology. In: *Bibliovault OAI Repository, the University of Chicago Press* (2003), 01

[19] Bunt, Harry ; Muskens, Reinhard: *Computational Semantics*, 01 1999

[20] Buziashvili, Eto: *What Russia reads on Telegram.* 21 Oktober 2022. – URL https://medium.com/dfrlab/what-russia-reads-on-telegram-a45d90e3bdc1. – Zugriffsdatum: 18-08-2023

[21] Charniak, E. ; Wilks, Y.: *Computational Semantics: An Introduction to Artificial Intelligence and Natural Language Comprehension.* North-Holland Publishing Company, 1976 (Fundamental studies in computer science). – URL https://books.google.de/books?id=_ZdrAAAAIAAJ

[22] Devlin, Jacob ; Chang, Ming-Wei ; Lee, Kenton ; Toutanova, Kristina: *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* 2018

[23] Echo of Moscow: *Der staatliche Fernsehsender Russia Today wird im nächsten Jahr fast 29 Milliarden Rubel aus dem Staatshaushalt erhalten.* 23 Dezember 2021. – URL https://web.archive.org/web/20220216141635/https://echo.msk.ru/news/2956368-echo.html. – Zugriffsdatum: 28-06-2023

[24] Forbes.ru: Yulia Kotova, Elena Ruzleva: *Die EU hat Sanktionen gegen "Tinkoffünd die Alpha-Bank verhängt.* 25 Februar 2023. – URL https://www.forbes.ru/society/485313-es-vvel-sankcii-protiv-tin-koff-i-al-fa-banka. – Zugriffsdatum: 17-07-2023

[25] Gaseta, Nowaja: *Offener Appell des Journalistenverbandes SSyndikat-100ßur Inakzeptabilität von Zensur in Zeiten der Pandemie.* 28 April 2020. – URL https://novayagazeta.ru/articles/2020/04/28/85124-tsena-kazhdomu-zadushennomu-izdaniyu-zhizni-lyudey. – Zugriffsdatum: 05-09-2023

[26] Gazeta.ru: *Durov hat nicht versprochen, zurückzukehren.* 22 April 2014. – URL https://www.gazeta.ru/business/2014/04/22/6002397.shtml?updated. – Zugriffsdatum: 10-08-2023

[27] Gorham, Michael S.: *After Newspeak: Language Culture and Politics in Russia from Gorbachev to Putin.* Ithaca, N.Y.: Cornell University Press, 2014

[28] Grefenstette, Gregory ; Tapanainen, Pasi: What is a word, What is a sentence? Problems of Tokenization. (1997)

[29] Indurkhya, Nitin ; Damerau, Fred J.: *Handbook of Natural Language Processing*. 2nd. Chapman & Hall/CRC, 2010

[30] interfax.ru: *Der Staat wird seinen Anteil am Ersten Kanal von 51% auf 34% reduzieren*. 22 Dezember 2020. – URL https://www.interfax.ru/russia/742516. – Zugriffsdatum: 28-06-2023

[31] Journal, Tinkoff: *Ausländische Agenten, unerwünschte Organisationen und Extremisten: Worin besteht der Unterschied, und wann kann eine normale Person unter den Artikel fallen?* 29 März 2023. – URL https://journal.tinkoff.ru/opasnye-svyazi/. – Zugriffsdatum: 20-08-2023

[32] Jurafsky, Dan ; Martin, James H.: *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, 2009

[33] Lenta.ru: Eugenija Naumova: *Der Fernsehsender „RainTV“ hat seine Arbeit eingestellt*. 03 März 2022. – URL https://lenta.ru/news/2022/03/03/dozhd_vse/. – Zugriffsdatum: 17-07-2023

[34] Manning, Christopher D. ; Schütze, Hinrich: *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999

[35] Massenmedien, Zentrum zum Schutz der Rechte der: *Alles über den Status des Äusländischen Agenten": Rechtliche Empfehlungen*. 16 Februar 2021. – URL https://mmdc.ru/services/common/vse-o-statuse-inoagenta-pravovye-rekomendaczii/. – Zugriffsdatum: 13-09-2023

[36] Mediazona: *Mediazona wurde aufgrund der Einnahmen aus Google-Anzeigen und Links zu Meduza als „ausländischer Agent“ eingestuft*. 30 Dezember 2021. – URL https://zona.media/news/2021/12/30/mz. – Zugriffsdatum: 28-06-2023

[37] Meduza: *„VKontakte“ hat die Seite von „Meduza“ blockiert*. 10 März 2022. – URL https://meduza.io/news/2022/03/10/vkontakte-zablokiruet-stranitsu-meduzy. – Zugriffsdatum: 15-08-2023

[38] Meduza: *Wie die Behörden die Russen daran hindern, sich gegen den Krieg auszusprechen, und sie dafür verfolgen. Das Wichtigste aus dem OVD-Info-Bericht*. 14 März 2022. – URL https://meduza.io/feature/2022/03/14/kak-vlasti-meshayut-rossiyanam-vyskazyvatsya-protiv-voyny-i-presleduyut-ih-za-eto. – Zugriffsdatum: 28-06-2023

[39] Meduza: *Putin unterzeichnete das Gesetz zur Bestrafung von „fake“ Handlungen des russischen Militärs*. 4 März 2022. – URL https://meduza.io/news/2022/03/04/

putin-podpisal-zakon-o-nakazanii-za-feyki-o-deystviyah-rossiyskih-voennyh. – Zugriffsdatum: 17-07-2023

[40] Meduza: Alexander Amzin: *Neusprech.* 2022. – URL https://us10.campaign-archive.com/?u=ff4a009ba1f59d865f0301f85&id=bc6d02c2b5. – Zugriffsdatum: 17-07-2023

[41] Meduza: Denis Dmitriev: *Das „*** *****“-Plakat „diskreditiert“ nun die Armee, und das Ersetzen der Preisschilder durch Informationen über die Toten sei ein „Fake“. Wie werden sie dafür bestraft?* 3 März 2022. – URL https://meduza.io/cards/gosduma-predlagaet-davat-do-15-let-za-rasprostranenie-feykov-o-deystviyah-rossiyskih-voennyh-i-chto-zhe-u-nas-schitaetsya-feykom. – Zugriffsdatum: 17-07-2023

[42] Meinung, Stiftung für die öffentliche: *Dominanten. Meinungsfeld. Ausgabe 2.* – URL https://fom.ru/Dominanty/14825. – Zugriffsdatum: 11-09-2023

[43] Mikheev, Andrei: Periods, Capitalized Words, etc. In: *Computational Linguistics* (2002), 09. – URL https://doi.org/10.1162/089120102760275992

[44] Ministerium für Justiz der Russischen Föderation: *Liste der öffentlichen Vereinigungen und religiösen Organisationen, gegen die ein rechtskräftiger Beschluss über Liquidation oder Verbot aufgrund der Bestimmungen des Bundesgesetzes Nr. 114-FZ vom 25.07.2002 "Über die Bekämpfung extremistischer Aktivitätenërlassen wurde.* – URL https://minjust.gov.ru/ru/documents/7822/. – Zugriffsdatum: 15-07-2023

[45] Orwell, George: *1984.* Secker and Warburg, 1949

[46] Paass, Gerhard ; Giesselbach, Sven: *Foundation Models for Natural Language Processing: Pre-trained Language Models Integrating Media.* 01 2023

[47] Pies, Judith: *Wandel im Journalismus autoritärer Regime: Das Beispiel Jordanien.* Bielefeld: transcript Verlag, 2015. – URL https://doi.org/10.1515/9783839431399

[48] Radford, Alec ; Narasimhan, Karthik ; Salimans, Tim ; Sutskever, Ilya: Improving language understanding by generative pre-training. (2018)

[49] Riezler, Stefan ; Hagmann, Michael: Validity, Reliability, and Significance, Empirical Methods for NLP and Data Science. (2022), 03

[50] Rossbach, Andreas: *Gazprom Media – kremlnahes Medienimperium auf Einkaufstour.* 15 Februar 2023. – URL https://www.absatzwirtschaft.de/gazprom-media-kremlnahes-medienimperium-auf-einkaufstour-229410/. – Zugriffsdatum: 15-07-2023

[51] Staatsduma. Bundesversammlung der Russischen Föderation: *Wie wird das neue Gesetz über ausländische Agenten funktionieren?* 29 Juni 2022. – URL http://duma.gov.ru/news/54760/. – Zugriffsdatum: 28-06-2023

[52] Statista: *Ranking of social media platforms in Russia Q3 2022.* 24 März 2023. – URL https://www.statista.com/statistics/867549/top-active-social-media-platforms-in-russia/. – Zugriffsdatum: 15-07-2023

[53] Statista: *Total number of active social media users in Russia 2015-2023.* 24 März 2023. – URL https://www.statista.com/statistics/278410/number-of-social-network-users-in-russia/#statisticContainer. – Zugriffsdatum: 15-07-2023

[54] Telegram: *Telegram API: access key.* – URL https://core.telegram.org/api/obtaining_api_id. – Zugriffsdatum: 18-08-2023

[55] Telegram: *Telegram API: Getting started.* – URL https://core.telegram.org/#getting-started. – Zugriffsdatum: 18-08-2023

[56] Telethon: *Telethon documentation.* – URL https://docs.telethon.dev/en/stable/. – Zugriffsdatum: 18-08-2023

[57] Time: *How Telegram Became the Digital Battlefield in the Russia-Ukraine War.* 21 März 2022. – URL https://time.com/6158437/telegram-russia-ukraine-information-war/. – Zugriffsdatum: 19-08-2023

[58] Tomita, Masaru: Efficient Parsing for Natural Language: A Fast Algorithm for Practical Systems, URL https://api.semanticscholar.org/CorpusID:60902655, 1985

[59] Traoré, Salifou: Herbst, Thomas; Schüller, Susen: Introduction to Syntactic Analysis. A Valency Approach. In: *Informationen Deutsch als Fremdsprache* (2010), 06

[60] Vaswani, Ashish ; Shazeer, Noam ; Parmar, Niki ; Uszkoreit, Jakob ; Jones, Llion ; Gomez, Aidan N. ; Kaiser, Łukasz ; Polosukhin, Illia: Attention is all you need, 2017

[61] Vedomosti: *Telegram hat zum ersten Mal WhatsApp beim Datenvolumen überholt.* 23 Januar 2023. – URL https://www.vedomosti.ru/technology/articles/2023/01/23/959995-telegram-obognal-whatsapp. – Zugriffsdatum: 18-08-2023

[62] VKontakte: *Strafverfolgung für einen Post auf VKontakte?* 23 August 2018. – URL https://vk.com/@nalogizakon-ugolovnoe-presledovanie-za-post-vo-vkontakte. – Zugriffsdatum: 10-08-2023

[63] VKontakte.com: *VK API: Access key.* – URL https://dev.vk.com/ru/api/access-token/getting-started. – Zugriffsdatum: 17-08-2023

[64] VKontakte.com: *VK API: Method wall.get.* – URL https://dev.vk.com/ru/method/wall.get. – Zugriffsdatum: 17-08-2023

[65] VKontakte.com: *VK API: Overview.* – URL https://dev.vk.com/ru/api/overview. – Zugriffsdatum: 17-08-2023

[66] Welle, Deutsche: *Welle von Vergiftungen: Journalisten aus Russland berichten Einzelheiten*. 15 August 2023. – URL https://www.dw.com/ru/volna-otravlenij-v-evrope-zurnalisty-i-aktivisty-iz-rf-rasskazali-podrobnosti/a-66542630. – Zugriffsdatum: 15-08-2023

A Anhang

A.1 Die Architektur des Transformer-Modells

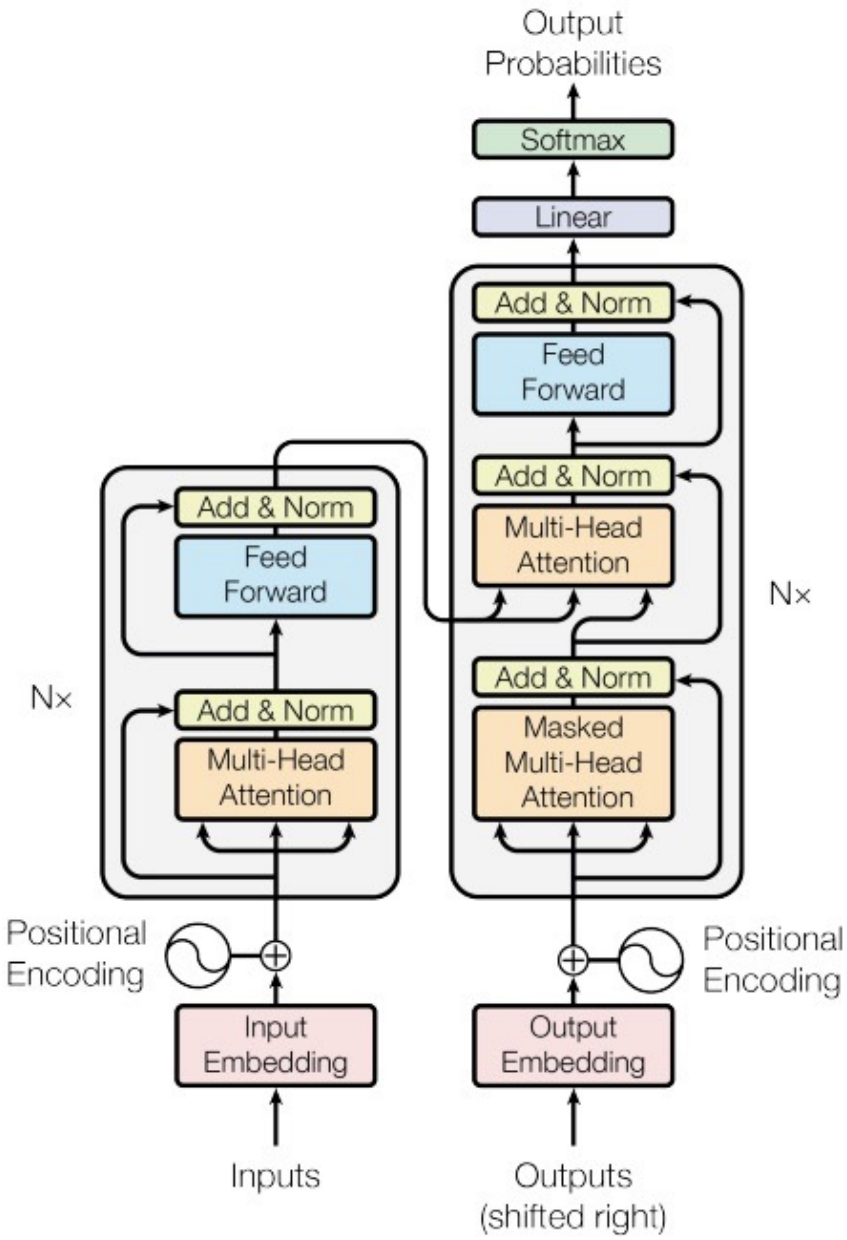


Abbildung A.1: Die Architektur des Transformer-Modells [60]

A.2 Die Anzahl der Nachrichten pro Tag bei allen Medien im Jahr 2022

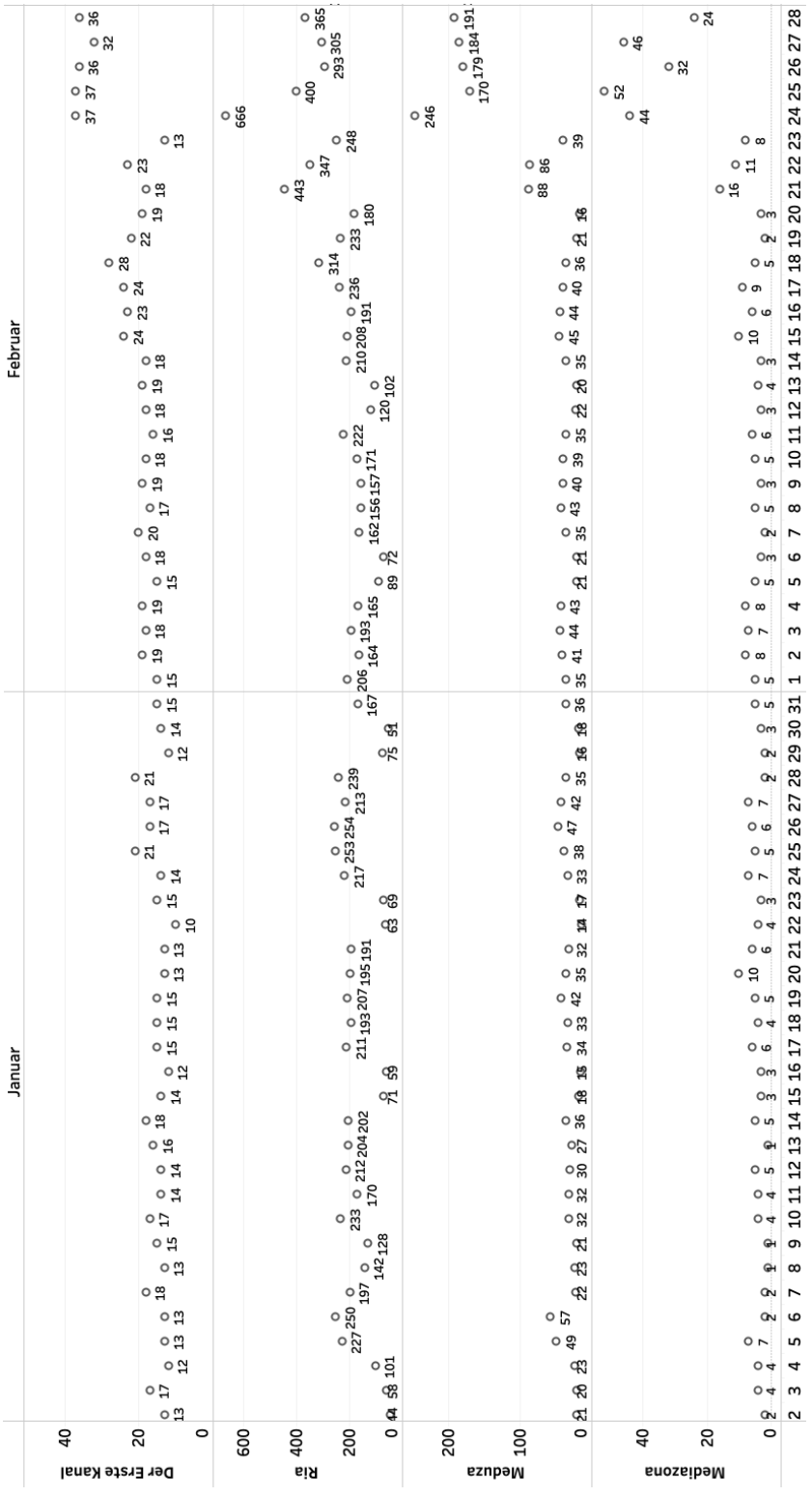


Abbildung A.2: Die Anzahl der Nachrichten pro Tag bei allen Medien im Januar-Februar 2022

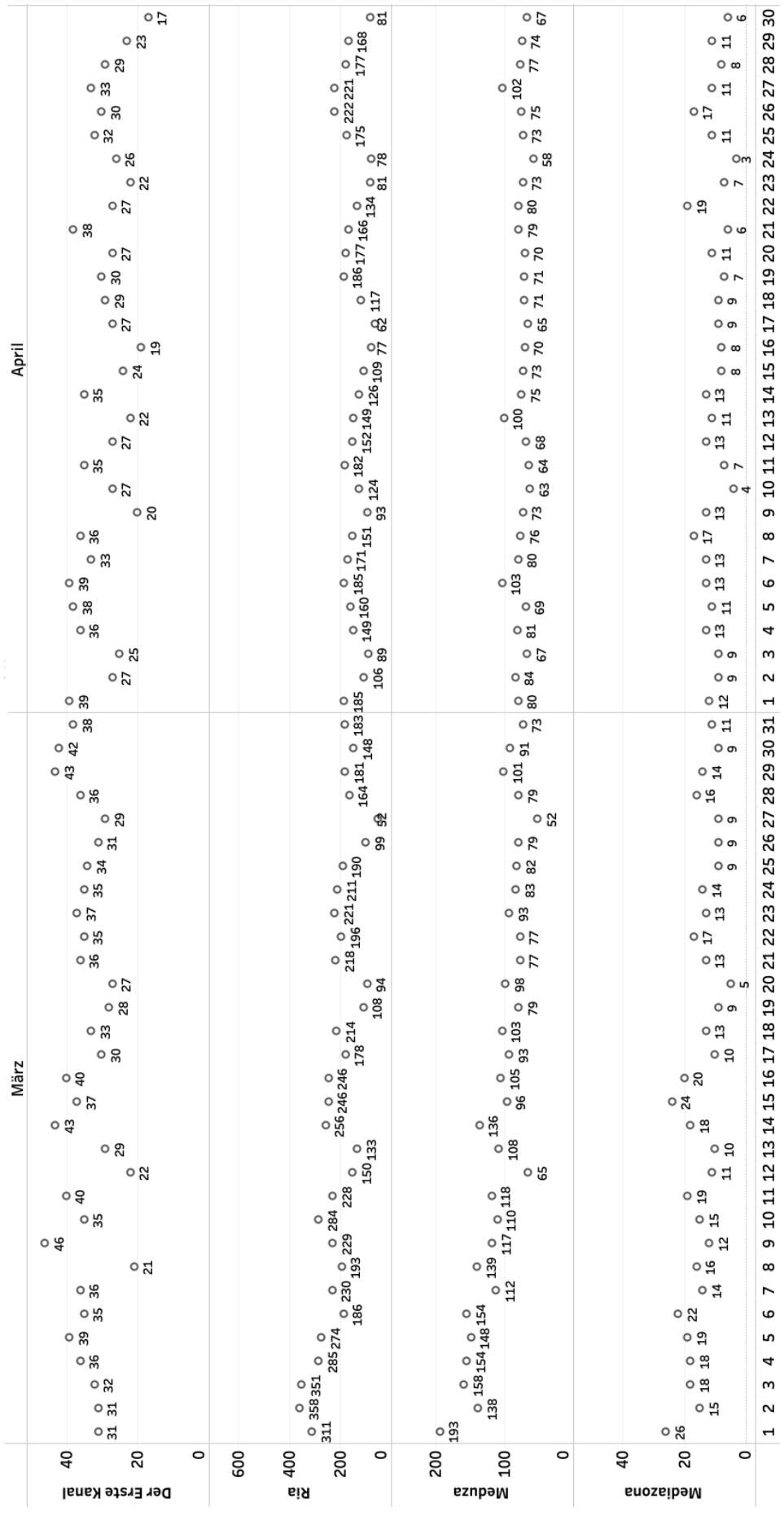


Abbildung A.3: Die Anzahl der Nachrichten pro Tag bei allen Medien im März-April 2022

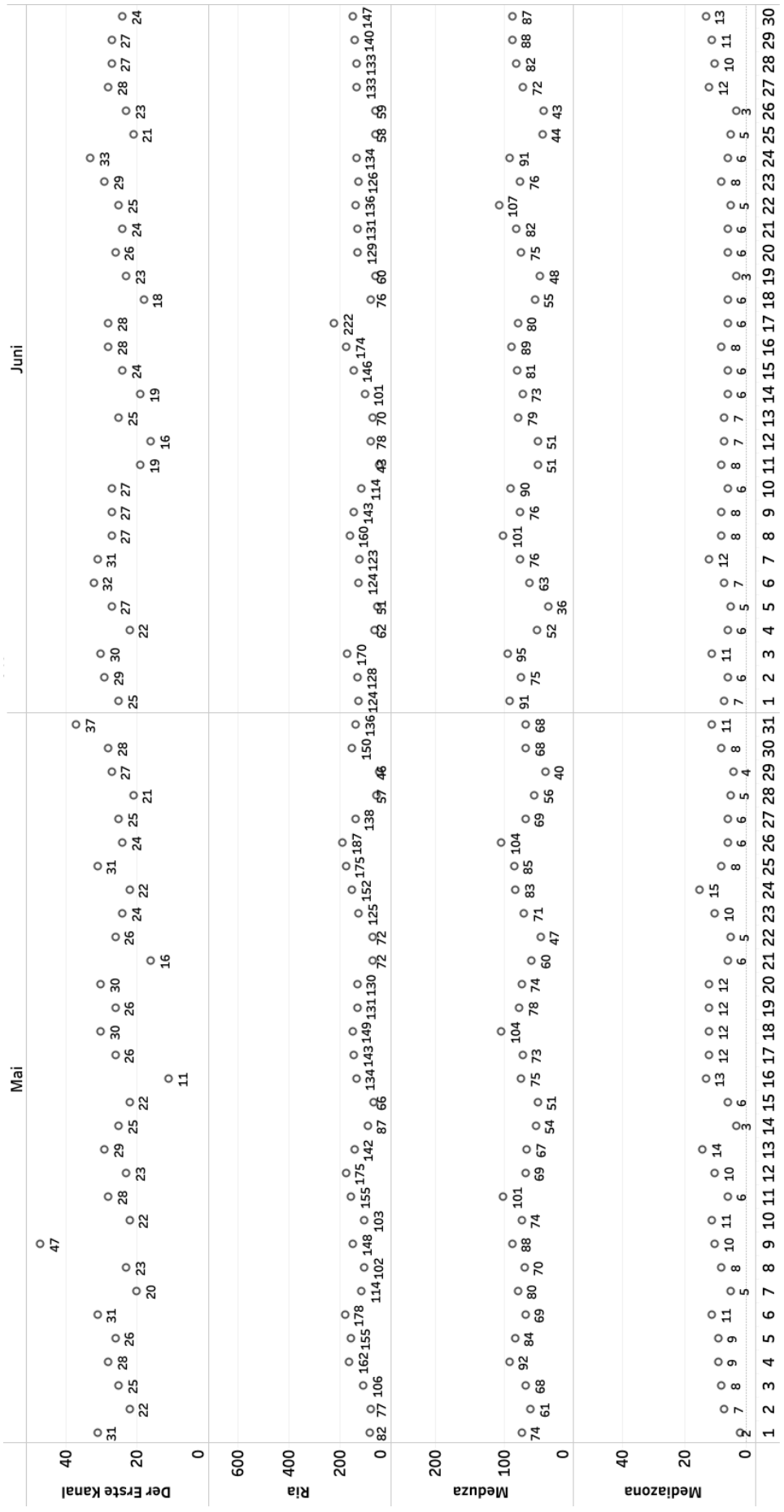


Abbildung A.4: Die Anzahl der Nachrichten pro Tag bei allen Medien im Mai-Juni 2022

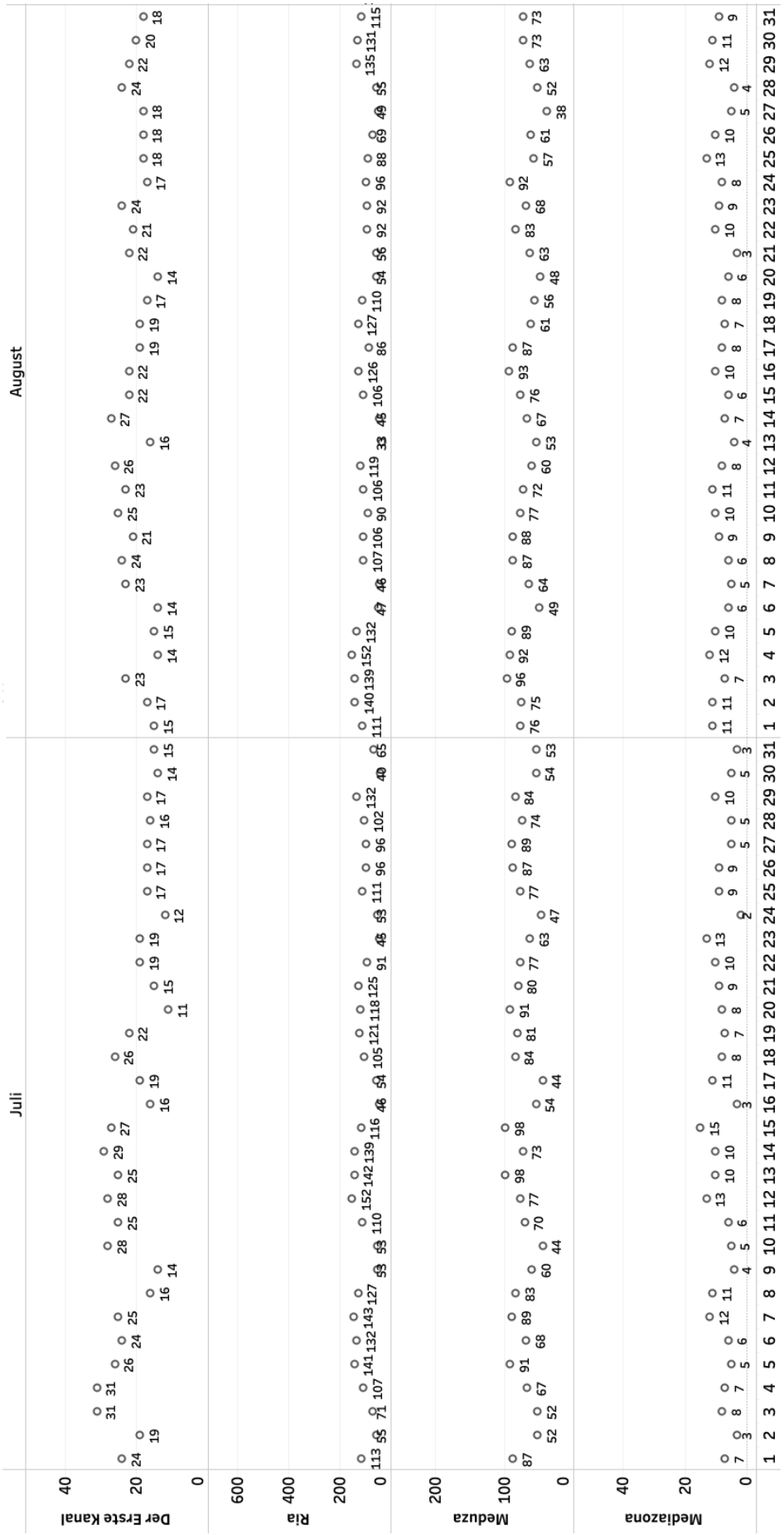


Abbildung A.5: Die Anzahl der Nachrichten pro Tag bei allen Medien im Juli-August 2022

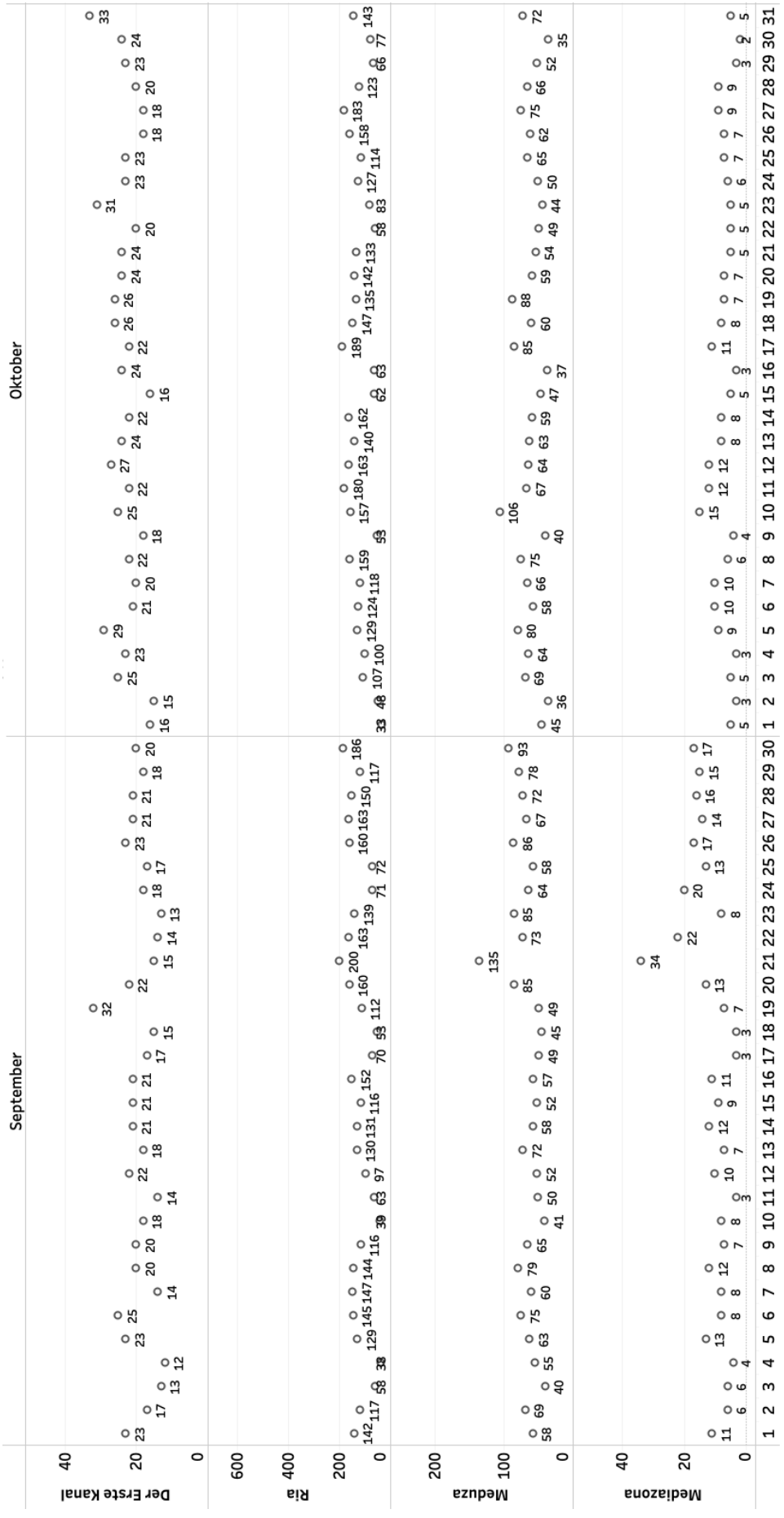


Abbildung A.6: Die Anzahl der Nachrichten pro Tag bei allen Medien im September-Oktober 2022

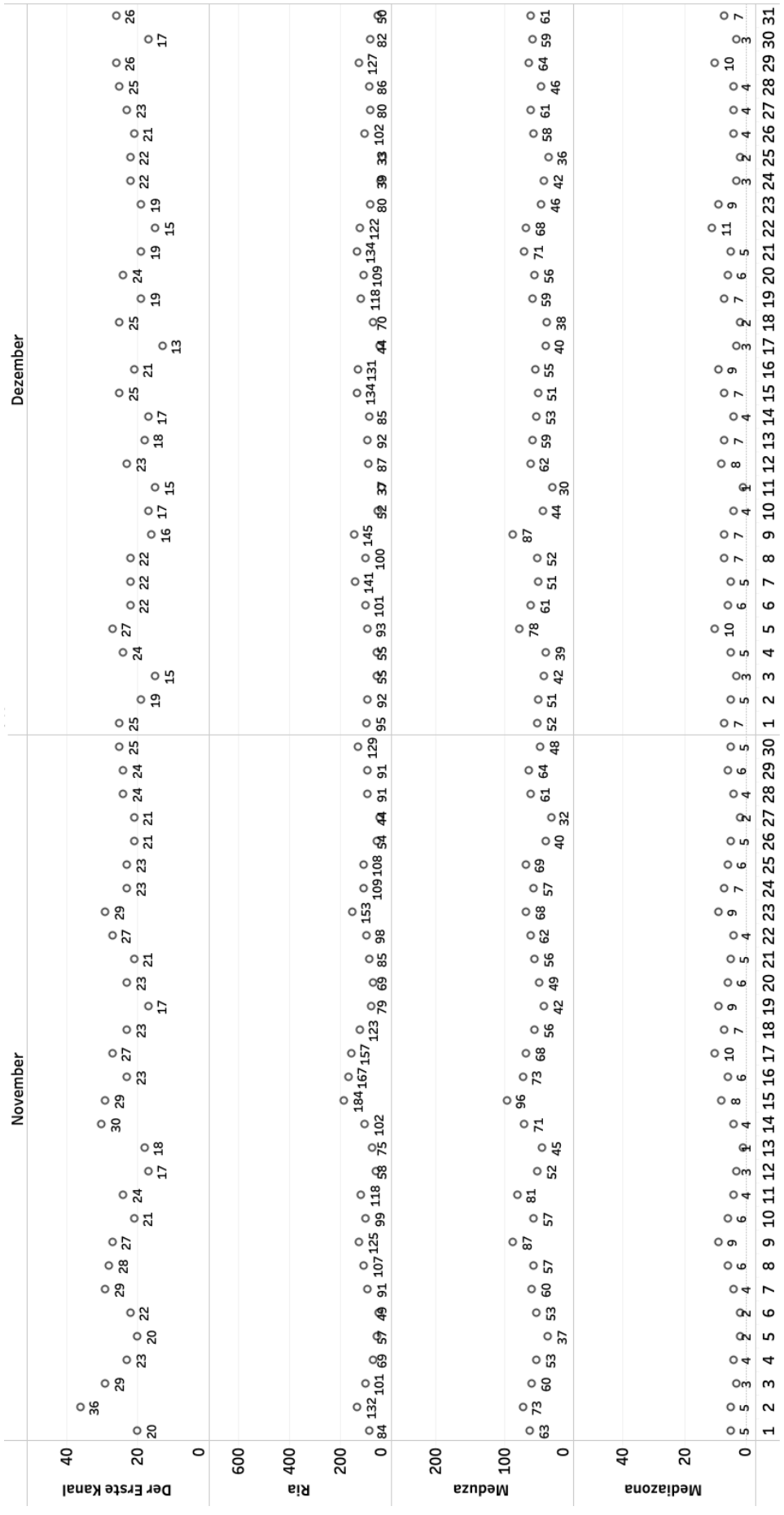


Abbildung A.7: Die Anzahl der Nachrichten pro Tag bei allen Medien im November-Dezember 2022

A.3 Worthäufigkeit der ausgewählter Wörter in allen Medien im Jahr 2022

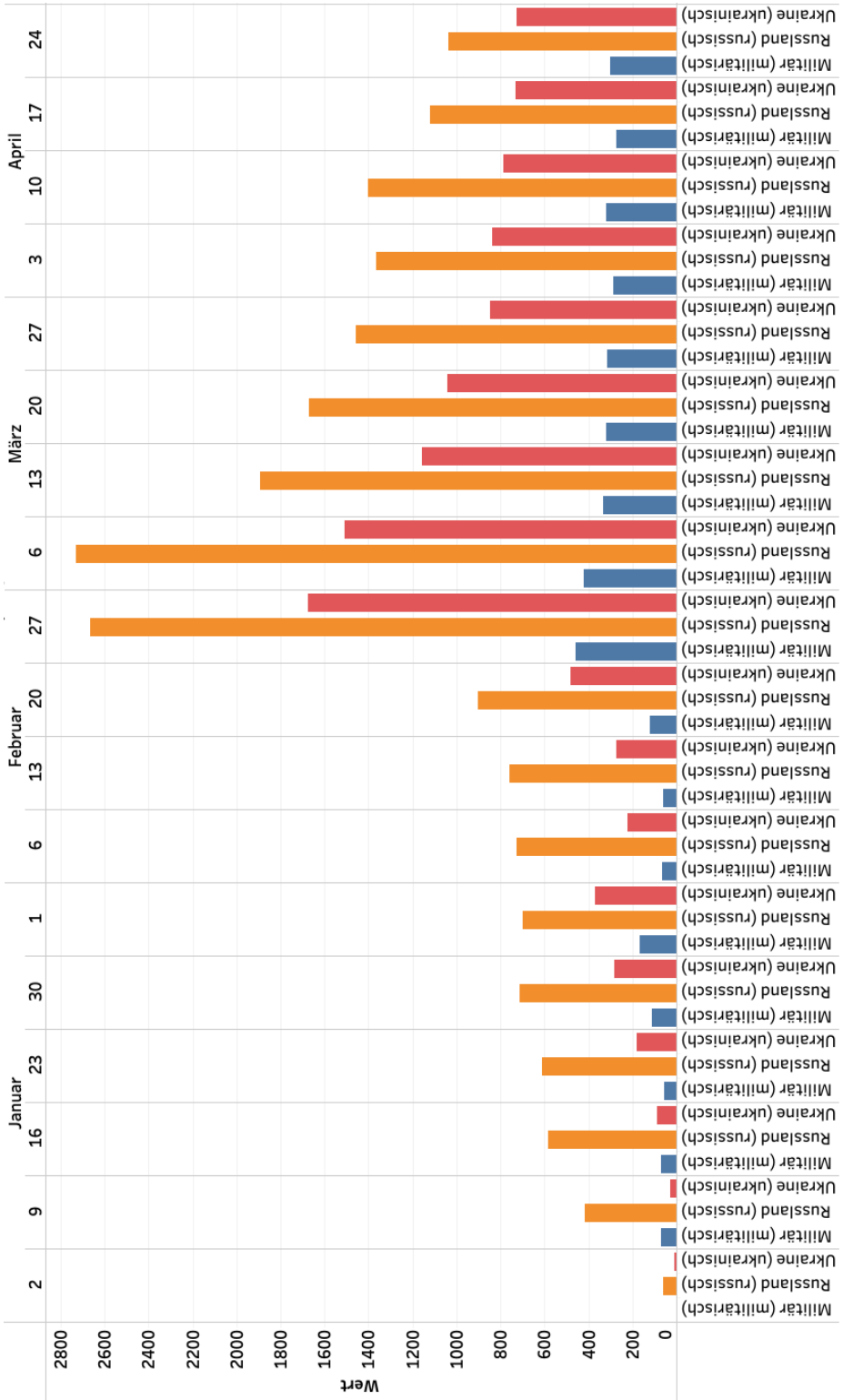


Abbildung A.8: Worthäufigkeit der ausgewählter Wörter in allen Medien im Januar-April 2022

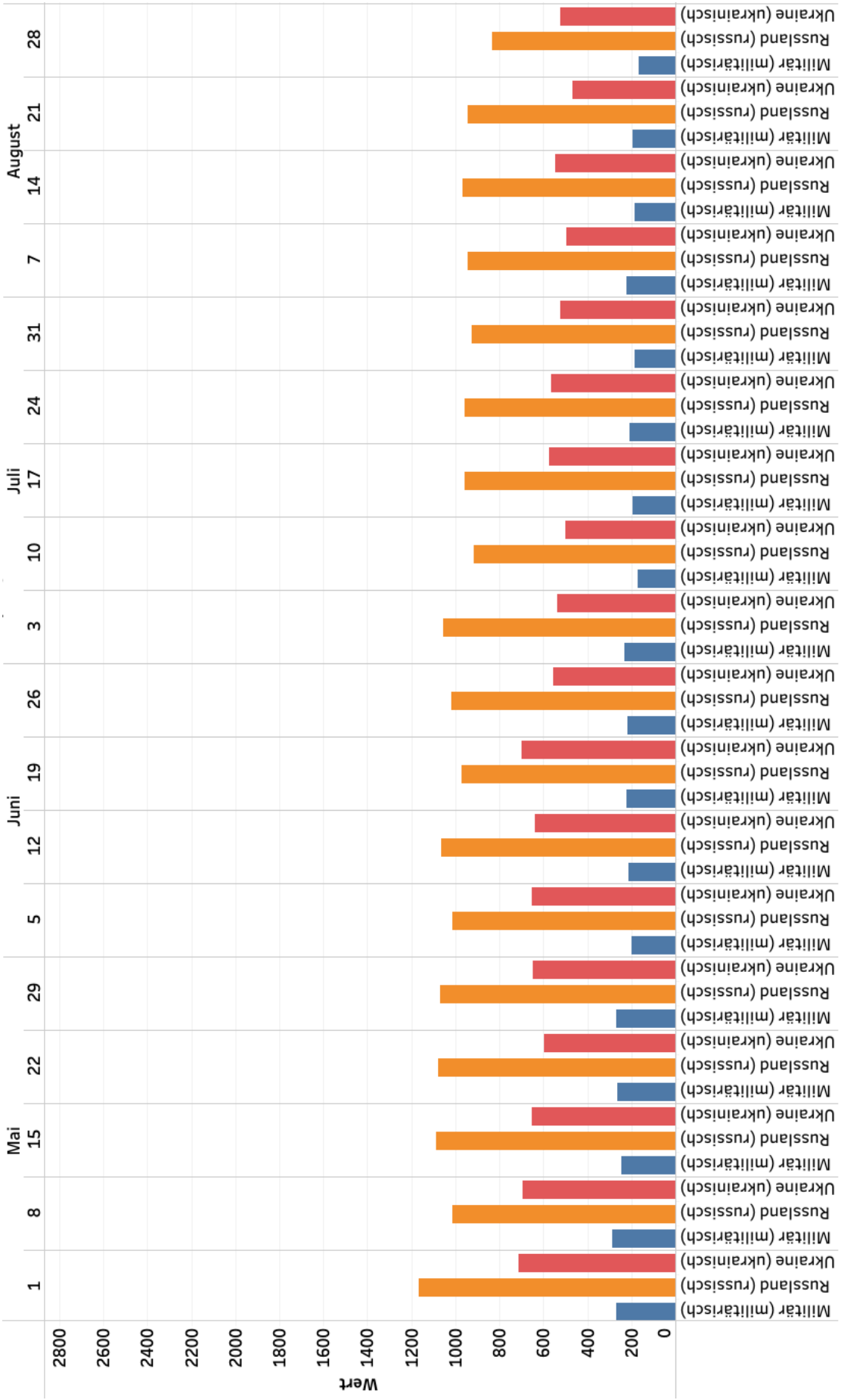


Abbildung A.9: Worthäufigkeit der ausgewählter Wörter in allen Medien im Mai-August 2022

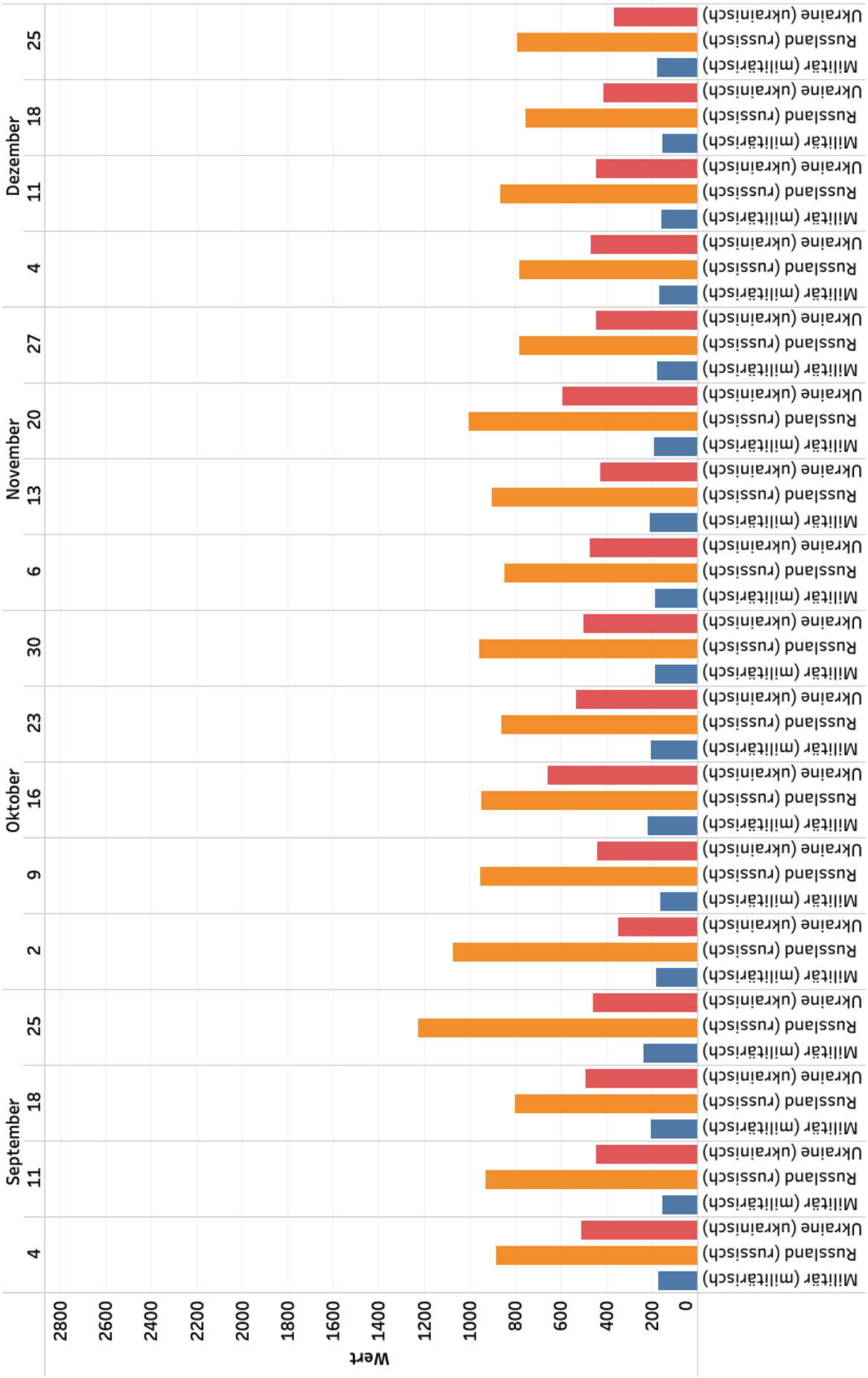


Abbildung A.10: Worthäufigkeit der ausgewählter Wörter in allen Medien im September-Dezember 2022

A.4 Verwendungshäufigkeit des Wortes „Krieg“ in allen Medien im Jahr 2022

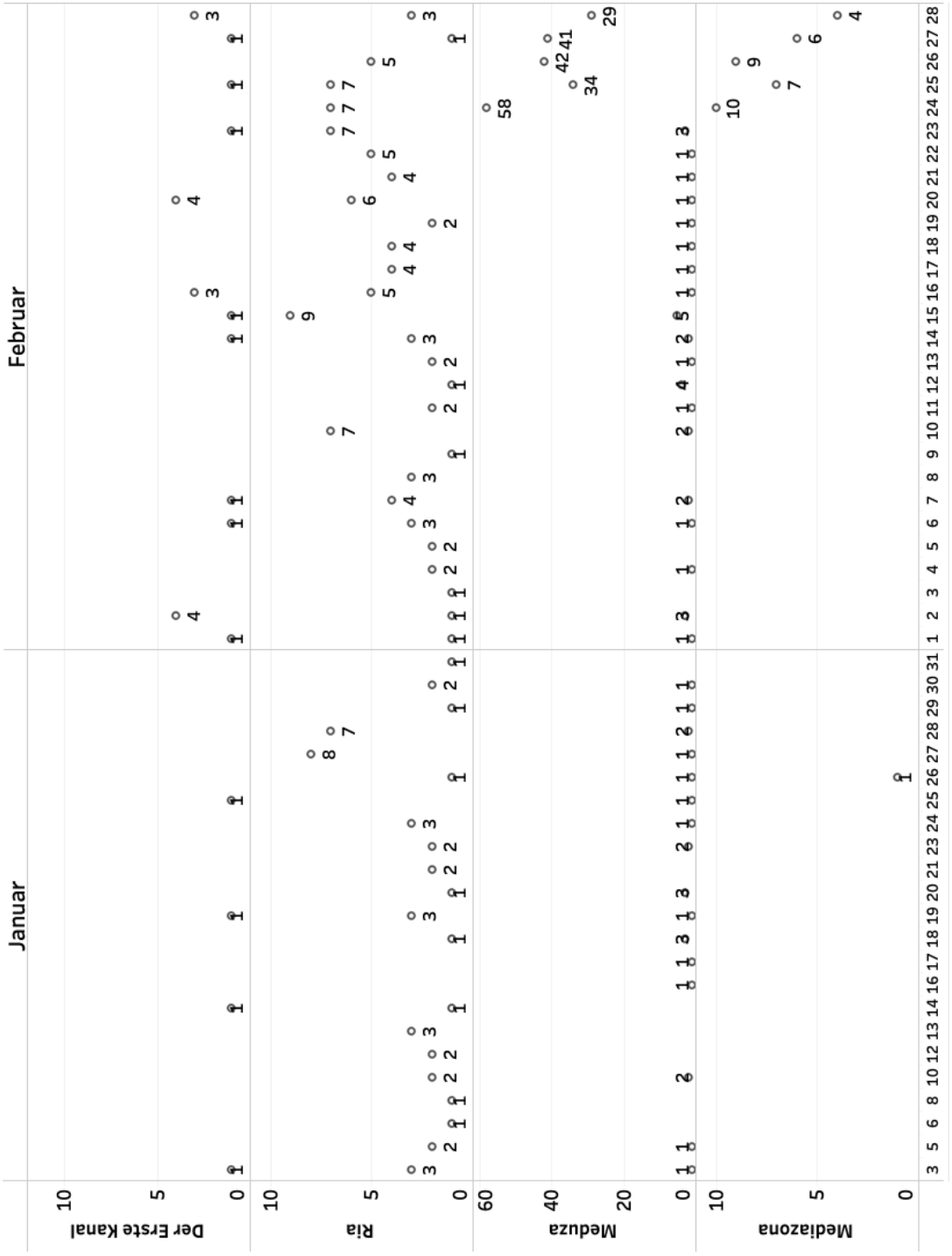


Abbildung A.11: Verwendungshäufigkeit des Wortes „Krieg“ in allen Medien im Januar-Februar 2022

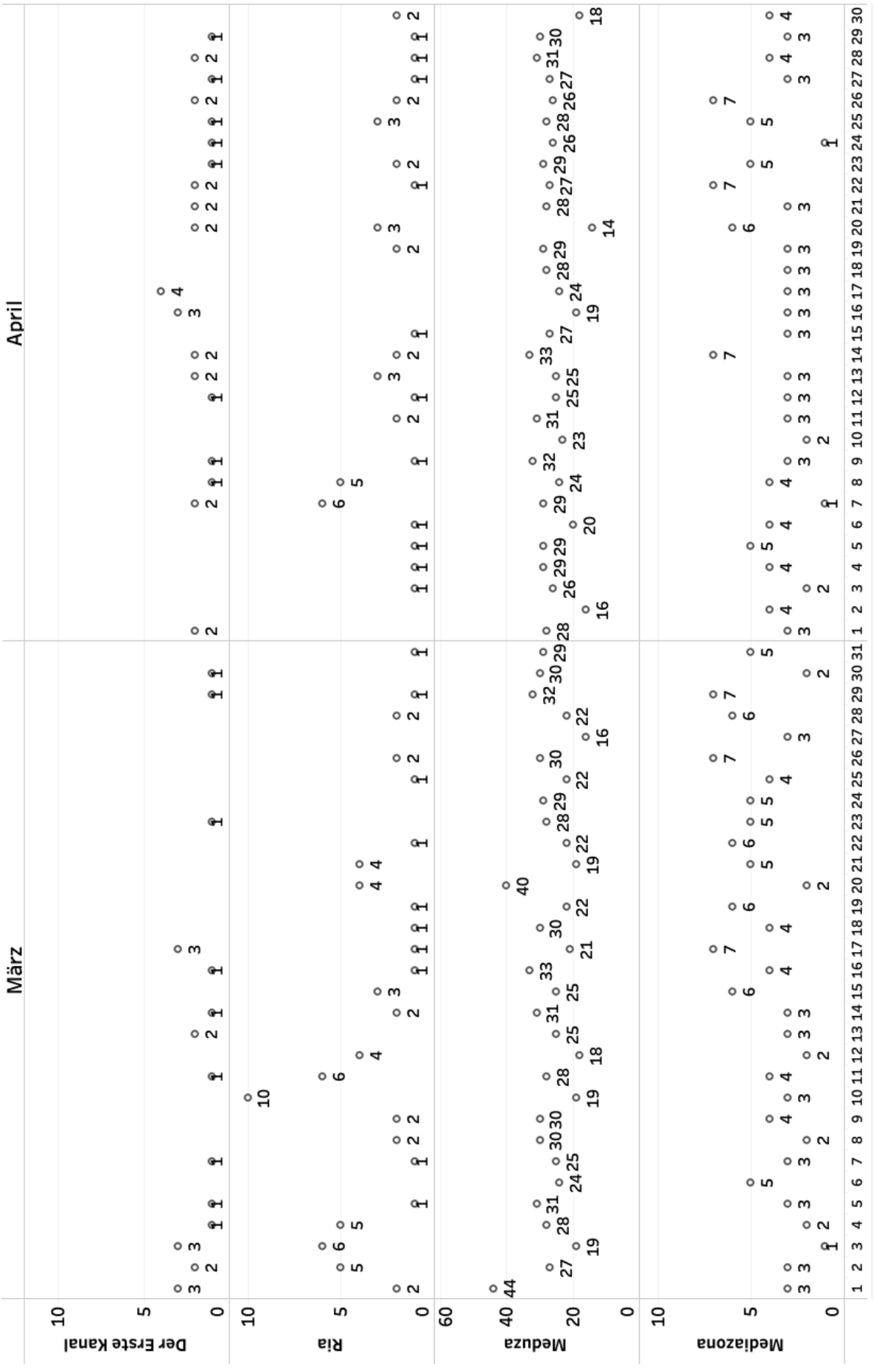


Abbildung A.12: Verwendungshäufigkeit des Wortes „Krieg“ in allen Medien im März-April 2022

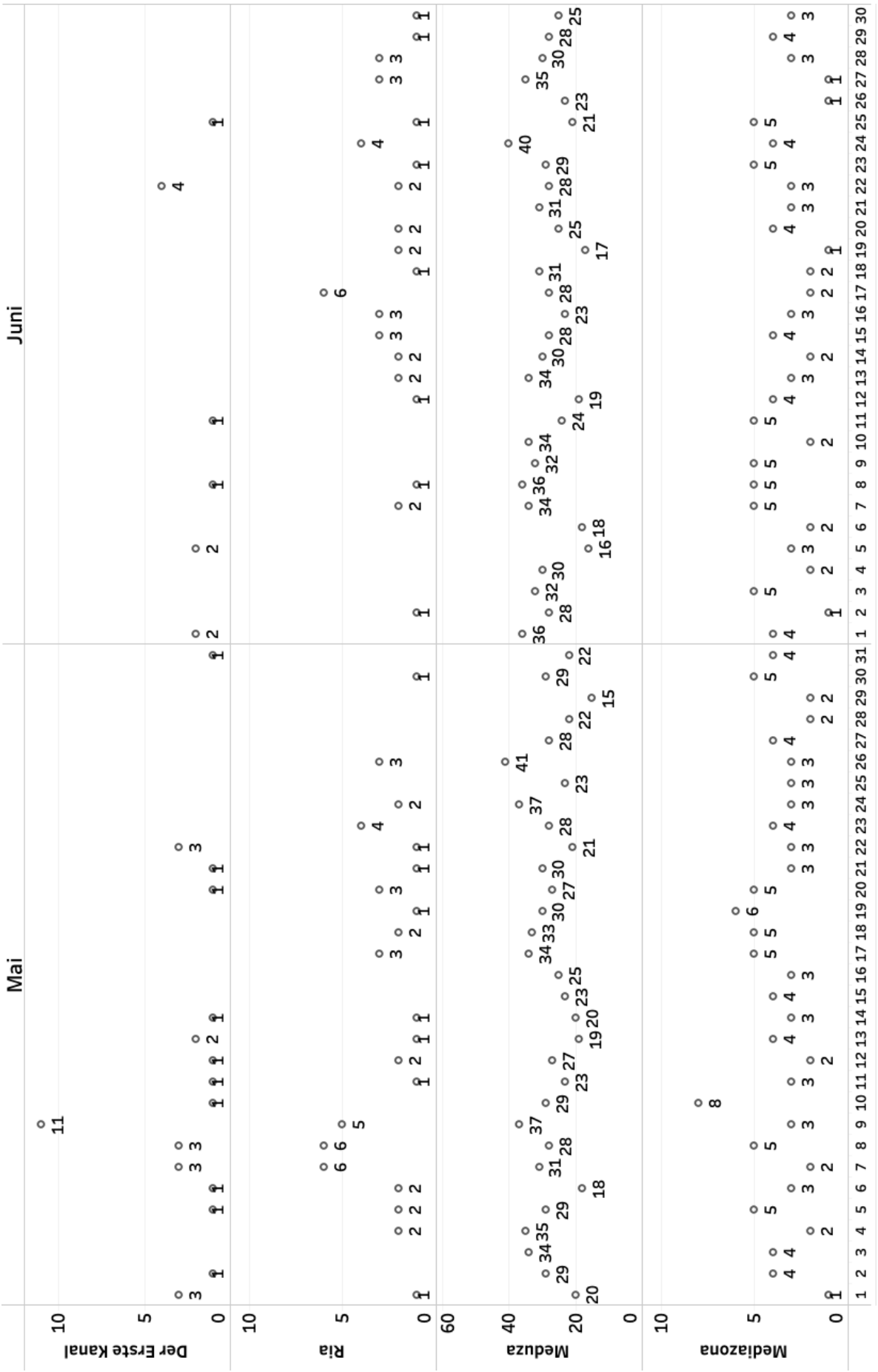


Abbildung A.13: Verwendungshäufigkeit des Wortes „Krieg“ in allen Medien im Mai-Juni 2022

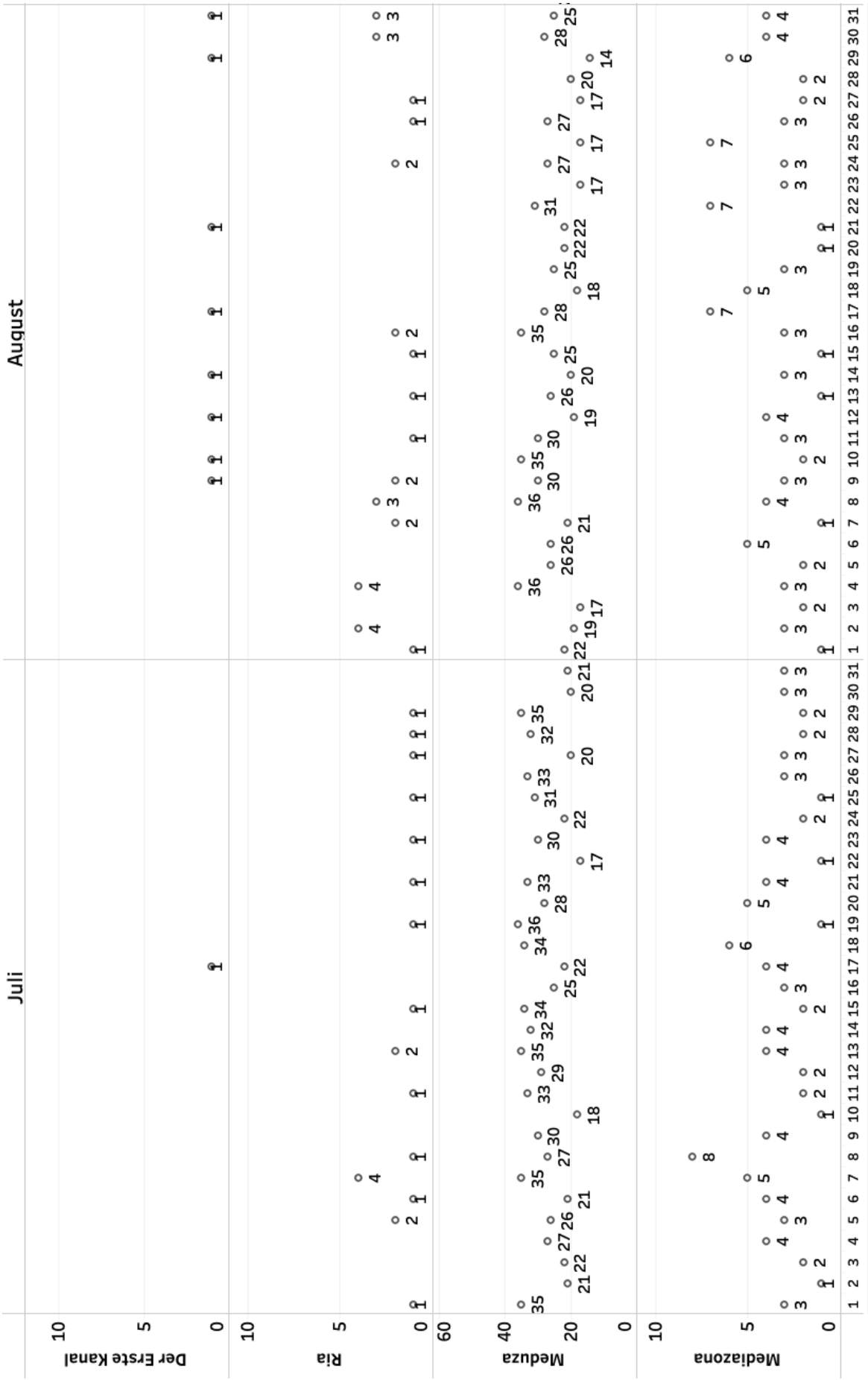


Abbildung A.14: Verwendungshäufigkeit des Wortes „Krieg“ in allen Medien im Juli-August 2022

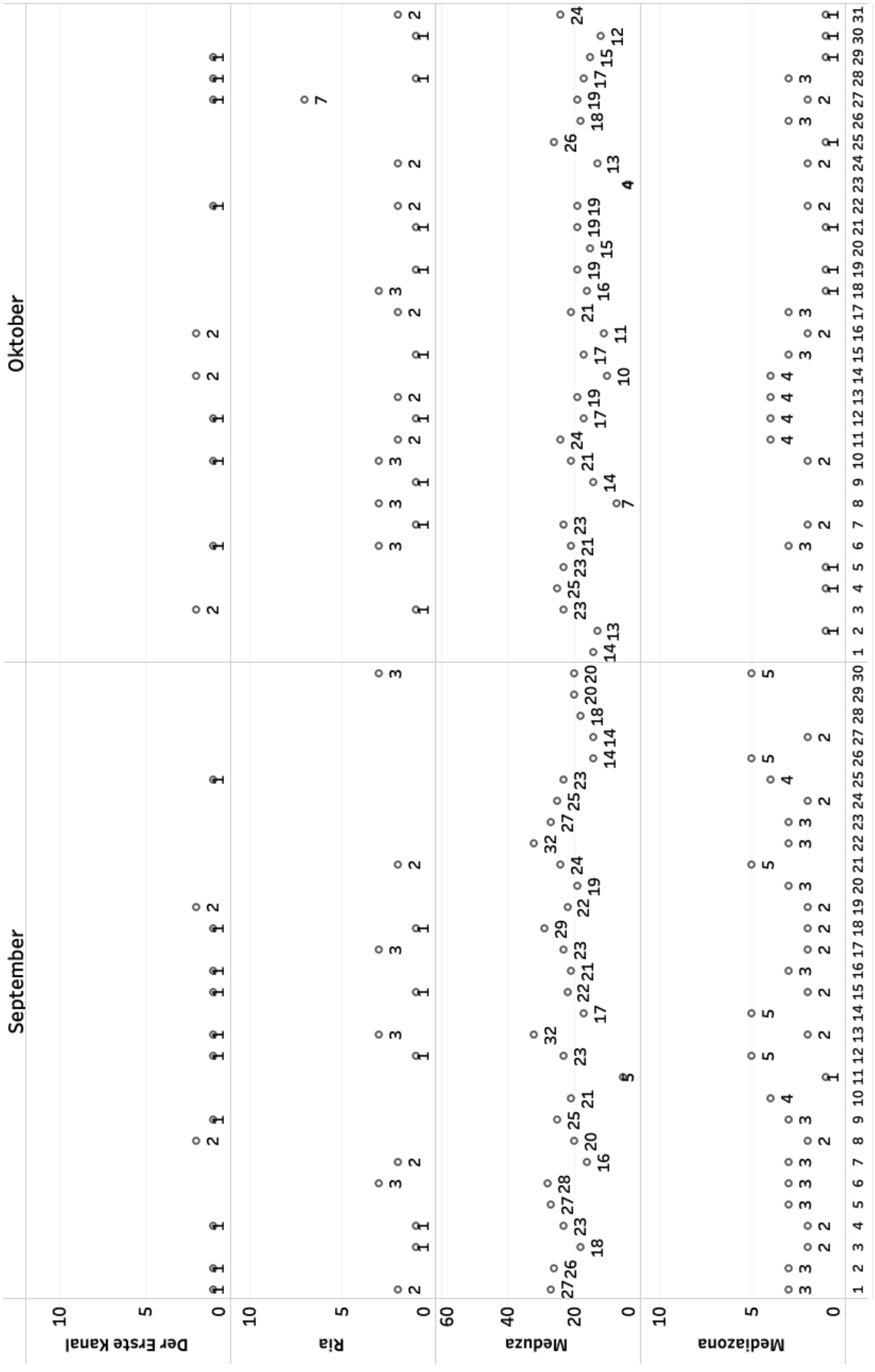


Abbildung A.15: Verwendungshäufigkeit des Wortes „Krieg“ in allen Medien im September-Oktober 2022

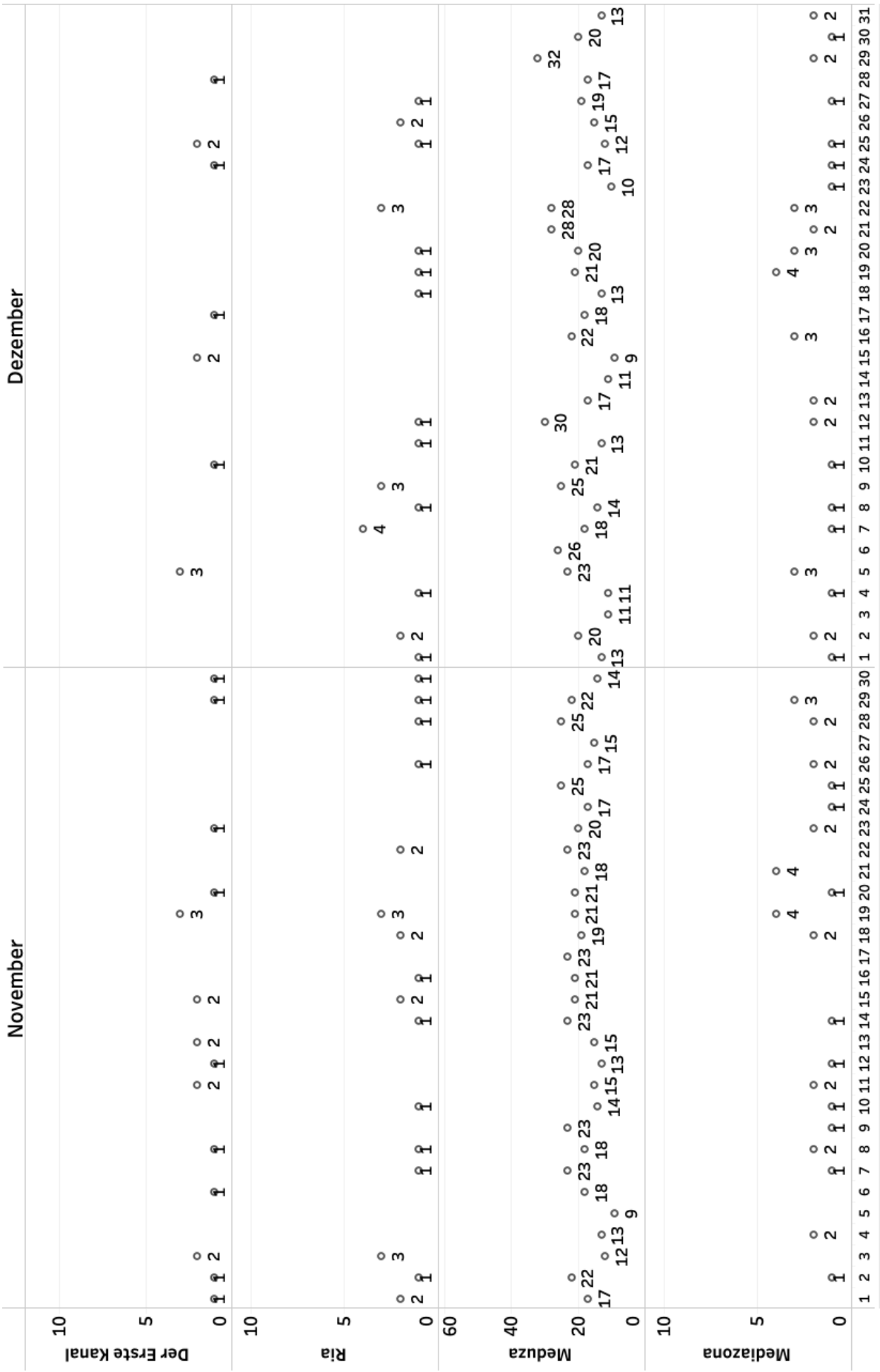


Abbildung A.16: Verwendungshäufigkeit des Wortes „Krieg“ in allen Medien im November-Dezember 2022

A.5 Verwendungshäufigkeit des Wortes „Spezialoperation“ in allen Medien im Jahr 2022

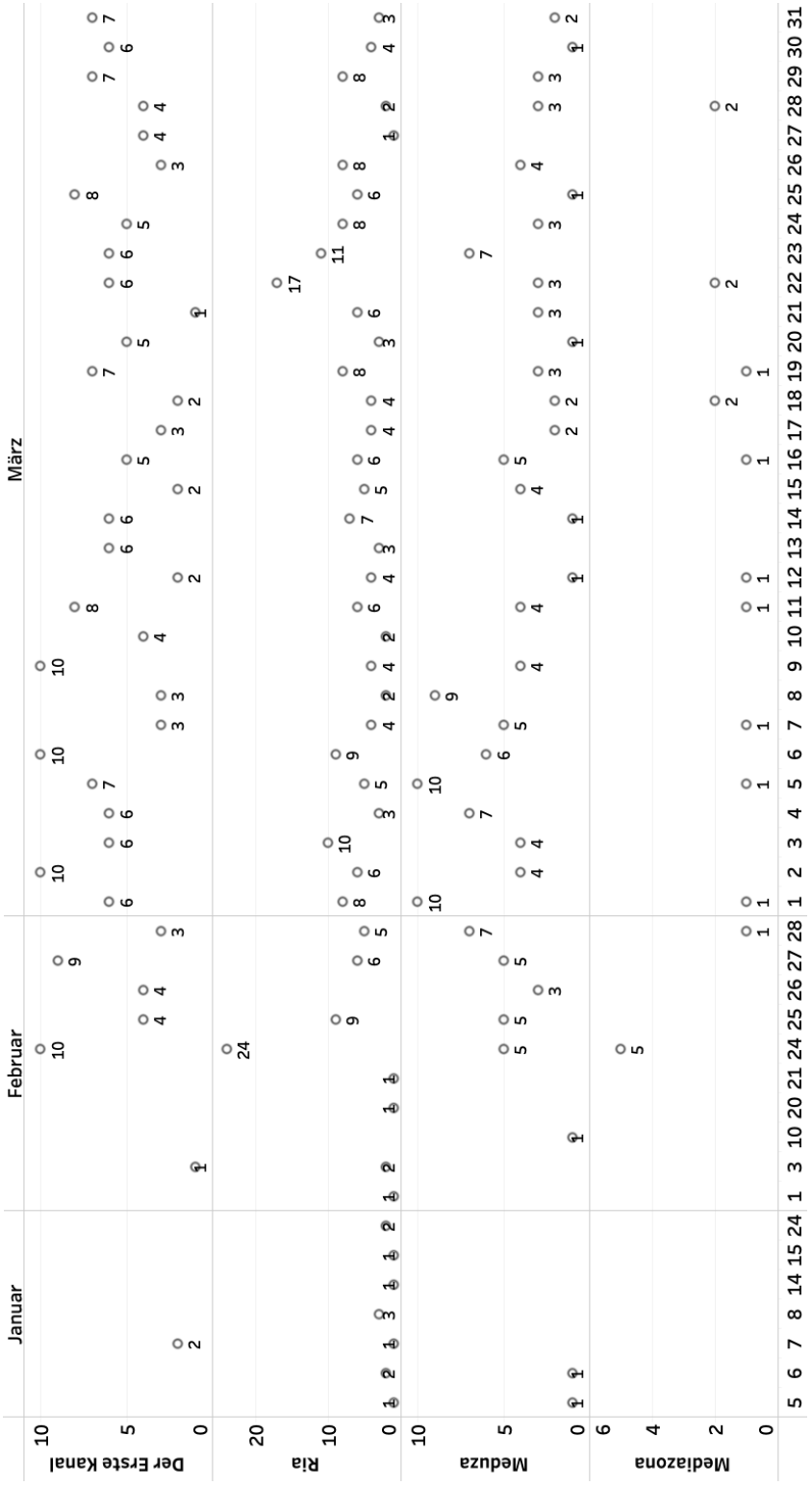


Abbildung A.17: Verwendungshäufigkeit des Wortes „Spezialoperation“ in allen Medien im Januar-März 2022

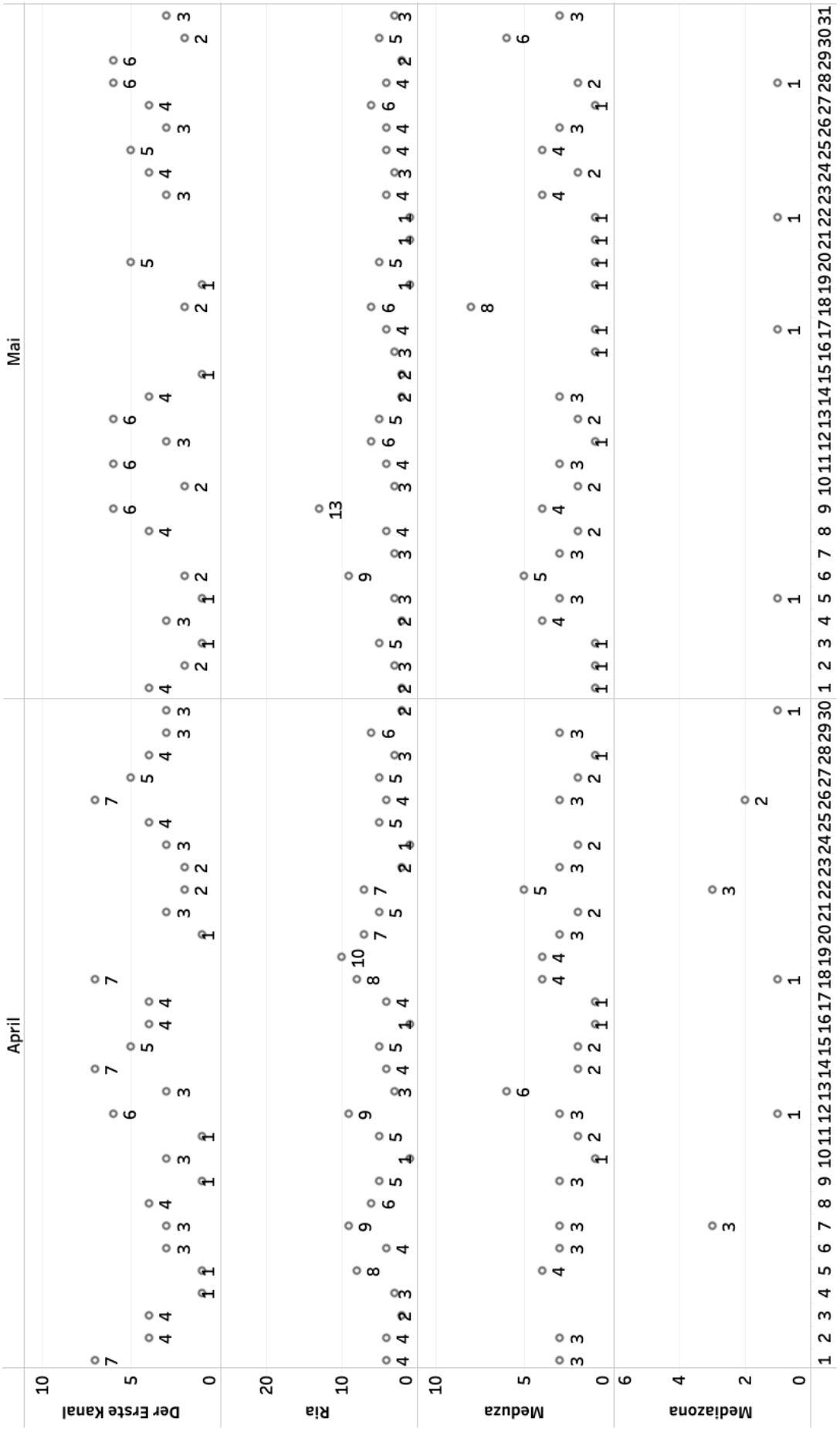


Abbildung A.18: Verwendungshäufigkeit des Wortes „Spezialoperation“ in allen Medien im April-Mai 2022

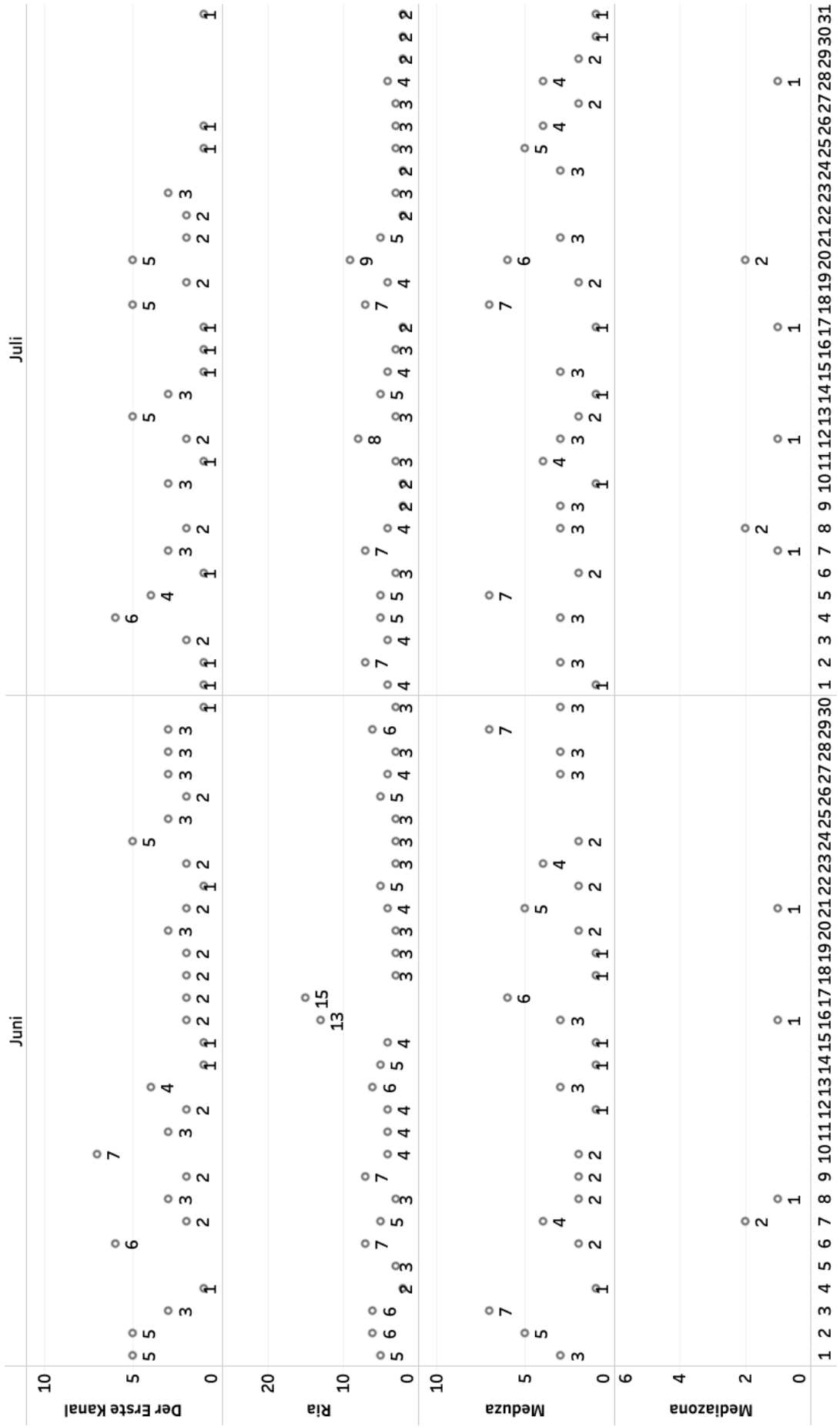


Abbildung A.19: Verwendungshäufigkeit des Wortes „Spezialoperation“ in allen Medien im Juni-Juli 2022

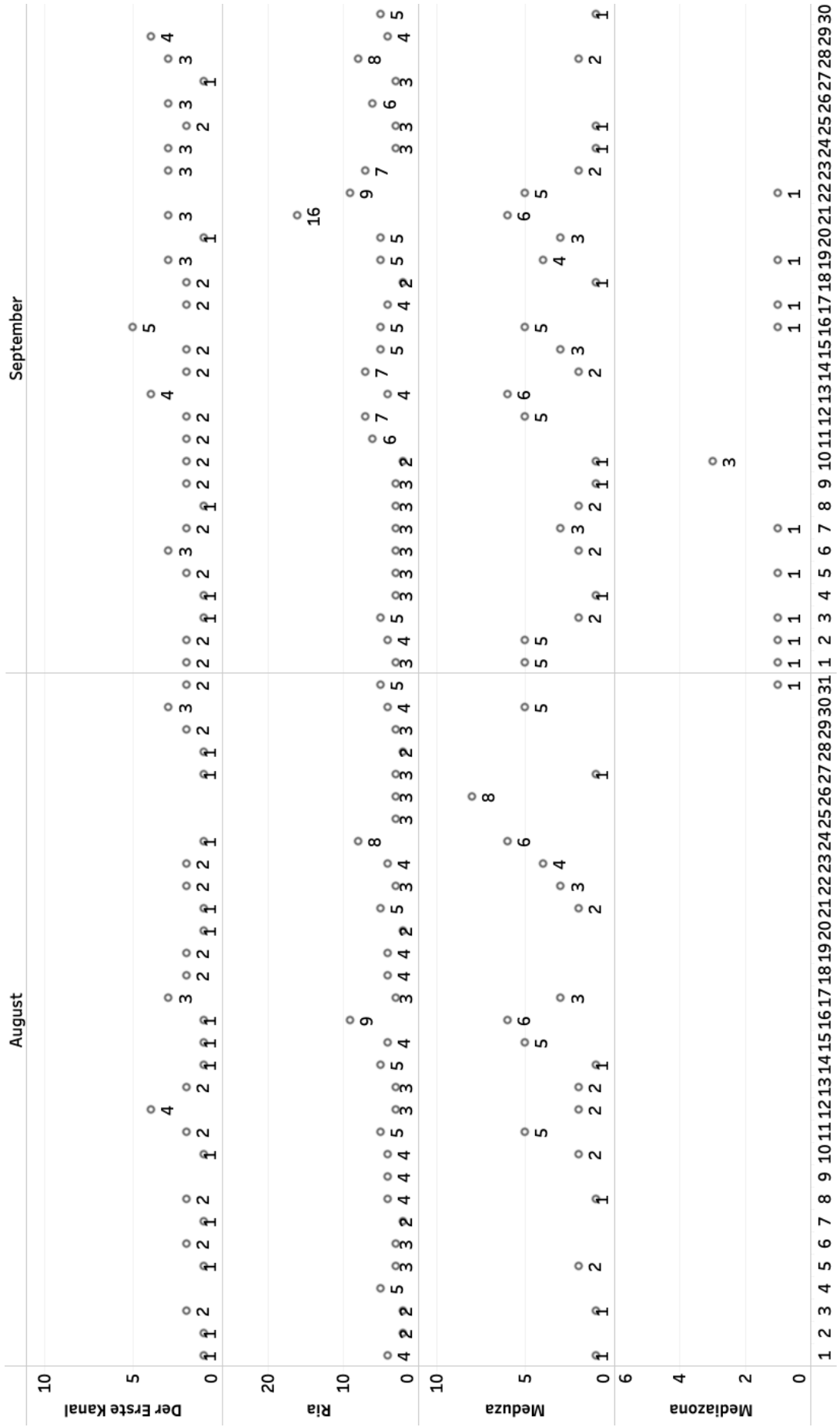


Abbildung A.20: Verwendungshäufigkeit des Wortes „Spezialoperation“ in allen Medien im August-September 2022

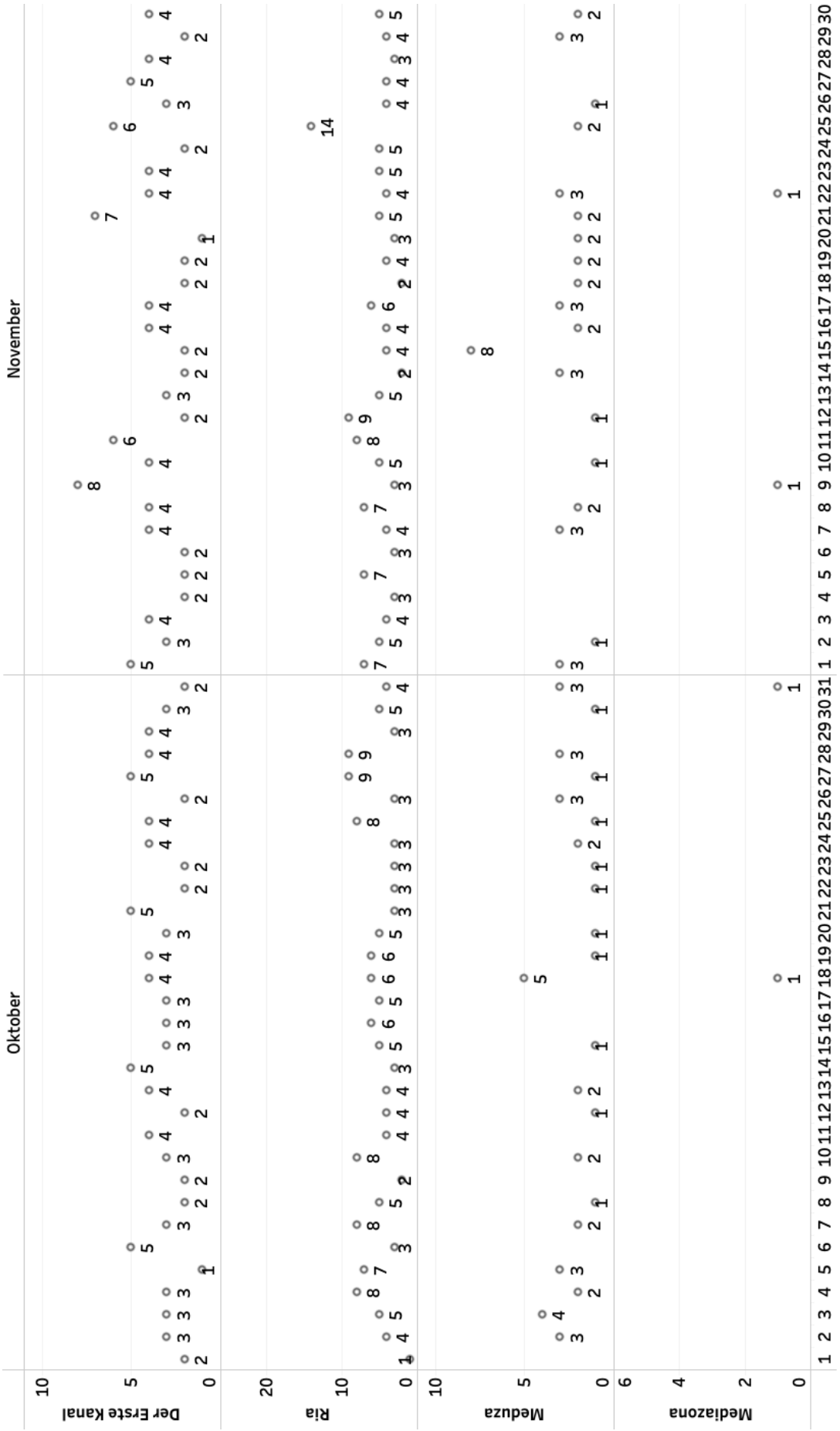


Abbildung A.21: Verwendungshäufigkeit des Wortes „Spezialoperation“ in allen Medien im Oktober-November 2022

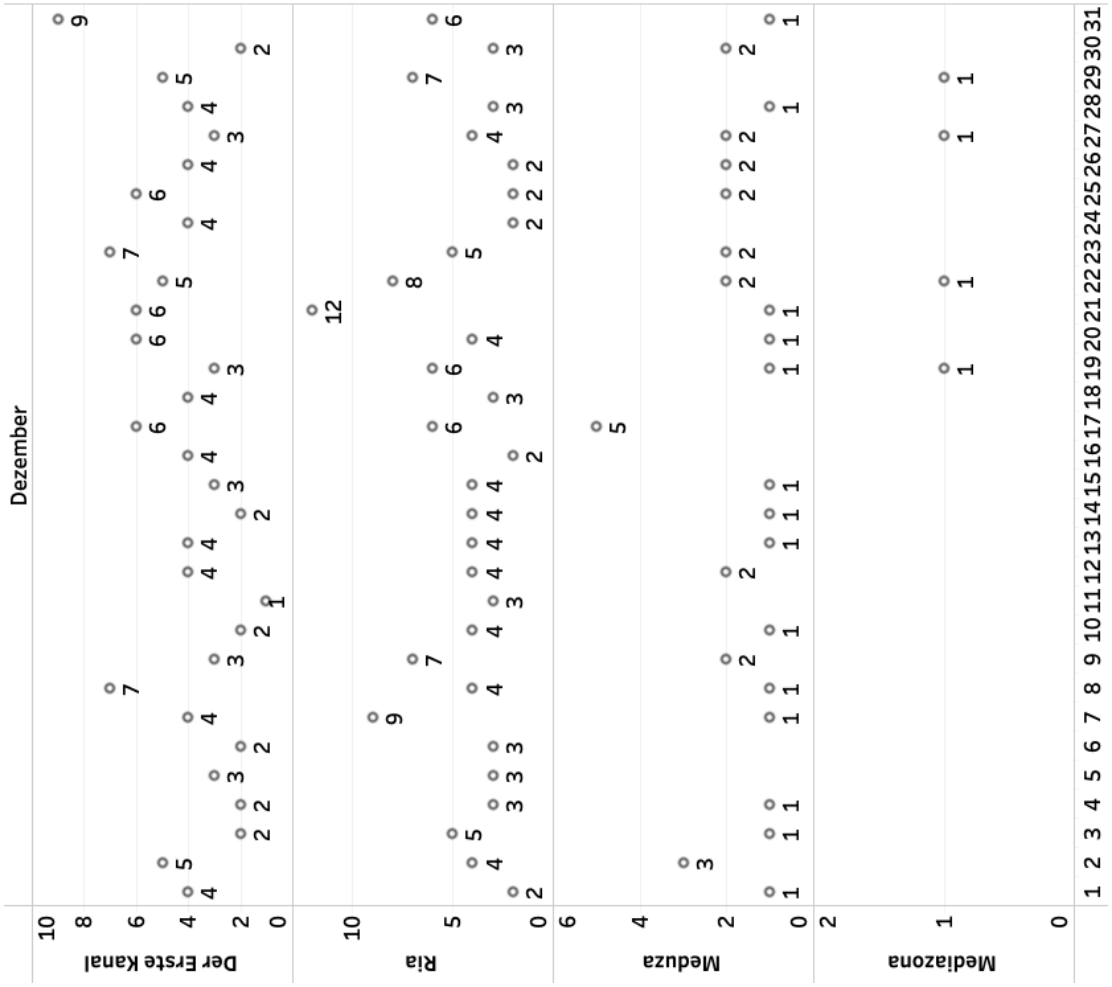


Abbildung A.22: Verwendungshäufigkeit des Wortes „Spezialoperation“ in allen Medien im Dezember 2022

Erklärung zur selbstständigen Bearbeitung

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbständig verfasst und nur die angegebenen Hilfsmittel benutzt habe. Wörtlich oder dem Sinn nach aus anderen Werken entnommene Stellen sind unter Angabe der Quellen kenntlich gemacht.

Ort	Datum	Unterschrift im Original