h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES
fbmn
FACHBEREICH MATHEMATIK
UND NATURWISSENSCHAFTEN

h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES
fbi
FACHBEREICH INFORMATIK

Hochschule Darmstadt
Fachbereich Mathematik und Naturwissenschaften &
Fachbereich Informatik

Sentiment-Analyse politischer Reden in deutschen Gremien

Abschlussarbeit zur Erlangung des akademischen Grades
Master of Science (M.Sc.)
im Studiengang Data Science

Melanie Renate Heidel

Matrikelnummer: 1126875

5. Dezember 2025

Referentin	:	Prof. Dr. Melanie Siegel
Korreferent	:	Prof. Dr. Horst Zisgen
Ausgabedatum	:	9. Juni 2025
Abgabedatum	:	5. Dezember 2025

Erklärung

Ich versichere hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die im Literaturverzeichnis angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder noch nicht veröffentlichten Quellen entnommen sind, sind als solche kenntlich gemacht. Die Zeichnungen oder Abbildungen in dieser Arbeit sind von mir selbst erstellt worden oder mit einem entsprechenden Quellennachweis versehen. Diese Arbeit ist in gleicher oder ähnlicher Form noch bei keiner anderen Prüfungsbehörde eingereicht worden.

Während der Vorbereitung dieser Arbeit habe ich GitHub Copilot, als Unterstützung zum Programmieren und ChatGPT, um Textpassagen stilistisch und grammatikalisch anzupassen, verwendet. Nach der Nutzung dieser Tools/ Dienste habe ich den Inhalt habe ich den Inhalt überprüft und bearbeitet und übernehme die volle Verantwortung für den Inhalt der Veröffentlichung.

Darmstadt, den 5. Dezember 2025

Melanie Heidel

Zusammenfassung

Die vorliegende Arbeit untersucht Methoden aus dem Bereich von Natural Language Processing (NLP) zur Analyse des Stimmungsbildes der Debatten in deutschen politischen Gremien. Die zentrale Forschungsfrage thematisiert die Eignung von ausgewählten Methoden zur Bearbeitung aufgestellter sozialwissenschaftlicher Fragestellungen. Schwerpunkt liegt dabei auf der Durchführung der Sentiment-Analyse, die durch die Erkennung von Hassrede unterstützt wird. Im Rahmen der Arbeit werden Parlamentsdaten aus deutschen politischen Gremien in einem Zeitraum von 2002 bis 2022 analysiert. Neben dem Deutschen Bundestag umfasst die Untersuchung die Landtage von Hessen, Sachsen, Schleswig-Holstein und Thüringen.

Zur Auswahl der geeignetsten Verfahren werden verschiedene Ansätze der Sentiment-Analyse und der Erkennung von Hassrede evaluiert. Die Sentiment-Analyse wird mit einem deutschsprachigen Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA)-Modell realisiert, dessen weiteres Training primär mit Parlamentsdaten durchgeführt wurde. Die Erkennung von Hassrede erfolgt in einem zweistufigen Verfahren. Im ersten Schritt filtert ein transformer-basiertes Modell, das auf die Erkennung von Hassrede trainiert ist, die Daten vor. Die vom Modell als potenzielle Hassrede identifizierten Einträge werden durch ein Large Language Model (LLM) weiterverarbeitet. Die Klassifizierungsergebnisse der Sentiment-Analyse und Erkennung von Hassrede werden für die Datenanalyse visualisiert. Analysiert werden die Daten im Hinblick auf folgende sozialwissenschaftlichen Fragestellungen:

(a) *Sind die Debatten in den deutschen Gremien aggressiver und negativer geworden?*

(b) *Lassen sich Wendepunkte in der Grundstimmung identifizieren und fallen diese mit markanten Ereignissen zusammen?*

(c) *Lässt sich eine Veränderung des Stimmungsbildes vor und nach Wahlen feststellen?*

(d) *Lässt sich ein negativer Bias gegenüber den Regionen des ehemaligen Ostens Deutschlands feststellen?*

Aus der Datenanalyse stechen die erkennbare Änderung um die Wahltermine sowie die Entwicklung hin zu zunehmender Negativität im Sächsischen Landtag hervor. Die Untersuchung hat gezeigt, dass mithilfe der ausgewählten Methoden der Sentiment-Analyse und der Erkennung von Hassrede die Fragestellungen thematisiert werden können und sich für eine Auseinandersetzung mit ihnen eignen. Insbesondere die unterstützende Untersuchung der Debatten auf das Vorkommen von Hassrede zeigt sich als vorteilhaft. Ausnahme bildet dabei die Fragestellung (d), für diese stellen die Methoden keine ausreichende Basis dar. Gleichzeitig werden dadurch weiterführende methodische Ansätze und Forschungspunkte verdeutlicht.

Abstract

This thesis analyzes debates in German political parliaments using methods from the field of NLP. The main research question addresses the selected methods and evaluates their suitability for analysing proposed social science questions. The main focus lies on the implementation and application of selected sentiment analysis methods. In addition hate speech detection methods are applied. Within the scope of this thesis parliamentary debates from German parliaments at both the federal and state levels, are studied for a time period from 2002 to 2022. Considered are the German Bundestag as well as the federal state parliaments from Hesse, Saxony, Schleswig-Holstein and Thuringia.

An evaluation is conducted to identify the most suitable model from the chosen methods in the fields of sentiment analysis and hate speech detection. The sentiment analysis is implemented using a german ELECTRA-model, which is further trained on data primarily from parliaments. Hate speech detection is performed in a two-step process. In the first instance a transformer-based model trained on the task of hate speech detection pre-filters the data. Entries marked by this model as potentially containing hate speech are then further processed with a LLM. The classification results from the sentiment analysis and hate speech detection tasks are visualised in order to support subsequent data analysis. The analysis regards the following social science questions:

(a) *Is there a trend toward more negative or aggressive debates in German parliaments?*

(b) *Is it possible to identify turning points in the general mood, and do these coincide with selected prominent events?*

(c) *Can a change be deteced, especially before and after elections?*

(d) *Can a negative bias be recognized toward regions of the former East Germany?*

The most notable findings from the data analysis are the observable changes around elections and the trend toward increasing negativity in the parliament of Saxony. The study shows that the selected methods for sentiment analysis and hate speech detection are suitable for addressing and analysing social science questions. In particular, the additional information provided by the detection of hate speechs proves to be advantageous. One exception is question (d). For this question the proposed methods do not provide a sufficient basis to fully address the issue. This identified limitation highlights further methodological approaches and new research topics.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AfD Alternative für Deutschland.

API Application Programming Interface.

BERT Bidirectional Encoder Representations from Transformers.

BPE Byte Pair Encoding.

ELECTRA Efficiently Learning an Encoder that Classifies Token Replacements Accurately.

GPU Graphics Processor Unit.

LLaMA Large Language Model Meta AI.

LLM Large Language Model.

LOESS Locally Estimated Scatterplot Smoothing.

MLM Masked Language Modeling.

MPID Member Parliament ID.

NLP Natural Language Processing.

NSP Next Sentence Prediction.

PoS Part-of-Speech.

ReLU Rectified linear unit.

SPD Sozialdemokratische Partei Deutschlands.

STL Seasonal Trend decomposition using LOESS.

SwiGLU Swish Gated Linear Units.

XLM cross-lingual language model.

1 Einleitung

1.1 Motivation

Deutsche Politikerinnen äußern in öffentlichen Interviews die Vermutung und persönliche Einschätzung, dass politische Debatten in Parlamenten negativer und aggressiver geworden sind. Die ehemalige Bundestagspräsidentin Bärbel Bas nimmt hierzu mit Blick auf den Bundestag Stellung. Ihrer Einschätzung nach ist eine zunehmende Härte in den Debatten erkennbar. Zudem habe sich die Sprache insgesamt diskriminierender entwickelt, sodass eine Veränderung der Atmosphäre im Plenum wahrnehmbar sei. Sie beschreibt weiterhin die Tatsache, dass Ordnungsrufe als Sammlungsstücke angesehen werden (tagesschau.de, 2024).

In einem weiteren Interview erklärt die amtierende Bundestagspräsidentin Julia Klöckner, sie nimmt eine steigende Polarisierung und Spaltung wahr. Ihrer Einschätzung nach besteht eine direkte Auswirkung des rauen Umgangstons auf die Gesellschaft (Kathe, 2025). Die Beschreibungen beider Politikerinnen machen die Relevanz, Bedeutung und Aktualität für die Gesellschaft deutlich, Debatten in deutschen Gremien zu untersuchen.

(Rheault et al., 2016) und (Lehtosalo & Nerbonne, 2020) zeigen in ihren Forschungsarbeiten, dass die Daten für das englische und finnische Parlament diesem negativ vermuteten Trend widersprechen. Vor diesem Hintergrund stellt sich die Frage, inwieweit Aussagen wie die von Bärbel Bas und Julia Klöckner durch eine systematische Analyse der Stimmung in deutschen politischen Gremien bestätigt oder widerlegt werden können.

Neben dem gesellschaftlichen Interesse ist diese Untersuchung im Bereich von NLP von wissenschaftlicher Relevanz. In dem Review von (Németh, 2022) wird der Einsatz von NLP im Zusammenhang zur Polarisierung in der politischen Domäne betrachtet. Dabei wird ein Trend zu überwachten Methoden aus dem Bereich des maschinellen Lernens mit 33%, zu der Untersuchung von Twitter mit 43% und mit 59% zur Thematisierung der Vereinigten Staaten festgestellt (Németh, 2022). Daraus entwickelte sich die Motivation der vorliegenden Forschungsarbeit. Entgegen diesem Trend liegt der Fokus der Arbeit auf transformer-basierten Methoden und konzentriert sich auf die Analyse von Debatten in deutschen Gremien.

Die Meinungen von Abgeordneten in Parlamenten könnten durch den Einsatz von Sentiment-Analyse besser verständlich werden. Im Gegensatz zu anderen Forschungsbereichen fehlen in dieser Domäne ausreichend große Datensätze, zur Durchführung dieser Aufgabe (Abercrombie & Batista-Navarro, 2020b). Zudem berücksichtigen öffentlich verfügbare transformer-basierte Modelle, die auf den Aufgabenbereich der Sentiment-Analyse trainiert sind, überwiegend keine Parlamentsdaten (Antypas et al., 2022; Guhr et al., 2020). Daraus ergibt sich eine weitere Motivation parlamentarische Daten in Forschungsarbeiten zu betrachten.

Zugleich ermöglicht die Analyse parlamentarischer Daten und Debatten eine enge Verbindung zu sozialwissenschaftlichen Fragestellungen. Methoden zur Analyse von Texten finden in breitgefächerten Richtungen der Sozialwissenschaft Anwendung, hierbei auch im Bereich der Politikwissenschaften. Dabei beschleunigen Methoden des NLP die Forschung in diesen Bereichen (Hou & Huang, 2025). Diese Schnittstelle bildet die Anregung für die vorliegende Arbeit.

1.2 Ziele und Problemstellung

Im Rahmen dieser Arbeit sollen Redebeiträge aus deutschen politischen Gremien untersucht werden. Grundlage bilden dabei die Parlamentsdaten der Gremien, die mit Methoden aus dem Bereich von NLP bearbeitet werden. Betrachtet werden der Deutsche Bundestag sowie die Landtage von Hessen, Sachsen, Schleswig-Holstein und Thüringen als Vertreter der bundes- und landespolitischen Ebenen. Die Analyse basiert auf ausgewählten NLP-Methoden. Schwerpunkt bildet hierbei der Bereich der Sentiment-Analyse. Ergänzend werden Erkenntnisse aus der Erkennung von Hassrede einbezogen.

Ziel der Forschungsarbeit ist es, politische Debatten mithilfe geeigneter Methoden aus dem Bereich des NLP zu analysieren. Insbesondere Methoden, die Gegenstand aktueller Forschungen sind, sollen dabei Inhalt der praktischen Arbeit sein. Die ausgewählten Methoden werden dafür zunächst evaluiert, um das beste Modell für die Datengrundlage auszuwählen. Anschließend sollen diese Modelle zur Klassifizierung der Daten genutzt werden. Deren Ergebnisse dienen als Grundlage für die Datenanalyse.

Erforscht werden soll dabei, inwieweit diese Methoden geeignet sind sozialwissenschaftliche Fragestellungen zu adressieren. Der Schwerpunkt liegt in der Erforschung der Stimmungslage in den Gremien. Dabei ist zu untersuchen, ob eine gestiegene Negativität sowie Aggressivität ersichtlich wird. Ein weiteres Ziel besteht darin, Wendepunkte in der Grundstimmung zu identifizieren und zu analysieren, ob diese mit ausgewählten markanten politischen und gesellschaftlichen Ereignissen zusammenfallen. Ein besonderer Fokus liegt dabei auf den Zeiträumen vor sowie nach Wahlen. Zusätzlich ist von Interesse einen möglichen negativen Bias gegenüber den Regionen des ehemaligen Ostdeutschlands zu untersuchen. Ausgangspunkt hierfür bilden Parlamentsdaten um den Tag der Deutschen Einheit.

Daraus leitet sich die zentrale Forschungsfrage der Arbeit ab: *Welche Methoden aus dem Bereich von NLP sind geeignet, um ausgewählte sozialwissenschaftliche Fragestellungen zu untersuchen.* Aus den vorherigen Ausführungen werden folgende sozialwissenschaftliche Fragestellungen formuliert:

(a) *Sind die Debatten in den deutschen Gremien aggressiver und negativer geworden?*

(b) *Lassen sich Wendepunkte in der Grundstimmung identifizieren und fallen diese mit markanten Ereignissen zusammen?*

(c) *Lässt sich eine Veränderung des Stimmungsbildes vor und nach Wahlen feststellen?*

(d) *Lässt sich ein negativer Bias gegenüber den Regionen des ehemaligen Ostens Deutschlands feststellen?*

Aufgrund der beschriebenen Zielsetzung werden nicht nur die einzelnen Redebeiträge der Abgeordneten analysiert, sondern zusätzlich auch Zwischenfragen, Reden der Sitzungsleitung sowie Zwischenrufe und Beifallsbekundungen berücksichtigt. Dadurch soll das Stimmungsbild detaillierter abgebildet werden. Im Rahmen der vorliegenden praktischen Arbeit erfolgt daher die Analyse von Debatten in deutschen Gremien.

1.3 Aufbau der Arbeit

Die Arbeit untergliedert sich in verschiedene Schwerpunkte, die in ihren jeweiligen Funktionen zum Erreichen des Ziels beitragen:

- *Theoretische Grundlagen*: Erläutert die theoretischen Grundlagen und wichtigsten Konzepte. Dadurch wird ein tiefgreifenderes Verständnis für die verwendeten Begriffe und Methoden erreicht.
- *Stand der Forschung*: Gibt einen Überblick über aktuelle Forschungsarbeiten und verdeutlicht den Bezug auf die praktische Arbeit.
- *Datengrundlage*: Ermöglicht einen Einblick in den Aufbau der domänenspezifischen Daten, die zur Beantwortung der Forschungsfrage verwendet werden. Dadurch zeigen sich die Struktur und Besonderheiten der Daten.
- *Methoden*: Verdeutlicht in einzelnen Abschnitten das Vorgehen und die praktisch durchgeführte Arbeit. Ziel dieses Kapitels ist es, den Prozess der Bearbeitung zu veranschaulichen. Gleichzeitig ermöglicht es Einblicke in die Entscheidungsprozesse und Besonderheiten während der Arbeit. Mithilfe der Erkenntnisse aus dem Evaluationsteil dieses Kapitels werden die performantesten Methoden aus den Bereichen der Sentiment-Analyse und Erkennung von Hassrede ausgewählt.
- *Ergebnisse*: Veranschaulicht die Ergebnisse der Untersuchung. Der zentrale Punkt dieses Kapitels stellt die Thematisierung und Auseinandersetzung mit den sozialwissenschaftlichen Fragestellungen dar.
- *Schluss*: Kompakte Zusammenfassung der wichtigsten Erkenntnisse und Darlegung von Ansätzen und Inspiration für anschließende Forschungsprojekte.

2 Theoretische Grundlagen

Dieses Kapitel dient der Definition und Einordnung der für die vorliegende Arbeit zentralen Themenschwerpunkte und Fachbegriffe. Es werden die Definitionen und Hintergründe, die zur Bearbeitung relevant sind, erläutert. Die Schwerpunkte innerhalb dieses Abschnittes fokussieren sich auf die Beschreibung verschiedener Vorverarbeitungsmethoden im Bereich von NLP sowie eine Einführung in die Sentiment-Analyse und Erkennung von Hassrede. Ein weiterer Punkt ist die Erläuterung und Beschreibung relevanter Hintergründe bei transformer-basierten Modellen. Definiert werden darauffolgend die im Rahmen der Arbeit verwendeten Evaluationsmetriken. Abgeschlossen wird der theoretische Teil durch die benötigten Konzepte der Zeitreihenanalyse.

2.1 Vorverarbeitungsmethoden

In diesem Abschnitt werden die praktisch eingesetzten Vorverarbeitungsmethoden beschrieben: Tokenisierung, Kleinschreibung, Entfernen von Stoppwörtern und Satzzeichen, Stemming, Lemmatisierung und Part-of-Speech (PoS)-Tagging. Sie kommen sowohl in der wörterbuch-basierten Sentiment-Analyse als auch bei der Erkennung von Hassrede mit Wortliste zum Einsatz. Bei der Anwendung der transformer-basierten Modelle wird keine Vorverarbeitung durchgeführt, sondern die Daten direkt an die Modelle übergeben.

Ziel der Vorverarbeitung ist es, den Text in eine nachvollziehbare Form zu bringen. Von Bedeutung sind hierbei Wörter, die den Kontext eines Textes verdeutlichen. Techniken, die dafür eingesetzt werden, sind unter anderem die Tokenisierung, die Kleinschreibung, das PoS Tagging, das Entfernen von Satzzeichen und Stoppwörtern sowie Stemming und Lemmatisierung (Tabassum & Patil, 2020).

Unter dem Begriff Tokenisierung ist der Prozess zu verstehen, der einen Text in Elemente unterteilt. Diese werden als Tokens bezeichnet. Beispiele hierfür sind Wortgruppen, Symbole oder Wörter. Die weitere Verarbeitung basiert auf den Tokens (Kannan et al., 2014). Ein Token wird in diesem Kontext als atomar betrachtet. Es erfolgt keine weitere Unterteilung eines Tokens (Webster & Kit, 1992). Neben Wörtern kann ein Token ein Sonderzeichen, eine Zahl oder eine Abkürzung darstellen (Siegel & Alexa, 2020).

Eine Maschine interpretiert ein und dasselbe Wort als unterschiedlich, im Falle, dass es einmal klein und einmal groß geschrieben wird. Die Kleinschreibung aller Wörter zählt zu den Vorgehensweisen, die sich bewährt haben (Tabassum & Patil, 2020). In der deutschen Sprache erfolgt die Großschreibung von Substantiven. Es entstehen Kontexte, in denen eine Unterscheidung in Groß- und Kleinschreibung für die eindeutige Identifizierung der Stimmung von Bedeutung sind. Das Verb 'würde' und das Substantiv 'Würde' verdeutlichen dies (Fehle et al., 2021).

Als Stoppwörter sind Wörter anzusehen, die häufig benutzt werden (Kannan et al., 2014). Exemplarisch sind dafür Pronomen, Artikel oder Präpositionen zu nennen (Vijayarani et al., 2015). Die Entfernung dieser Wörter begründet sich darin, dass sie keinen Mehrwert für die Klassifikation darstellen. Sie sind für den Inhalt und Kontext nicht relevant (Kannan et al., 2014). Eine weitere Technik ist, dass Satzzeichen entfernt werden. Deren Aufkommen erzeugt Unruhe im Text. Beispiele stellen Kommata, Ausrufezeichen und Apostrophe dar. Das Auffinden sowie das Ersetzen dieser erfolgt im Allgemeinen mit regulären Ausdrücken (Tabassum & Patil, 2020).

Stemming bildet den Prozess ab, der den Stamm von Wörtern erzeugt. Der Stamm ist die allgemeine Darstellung variierender Wortformen (Kannan et al., 2014). Der Prozess erfolgt unter der Annahme, dass die Formen in einer semantischen Beziehung zum Stamm stehen (Vijayarani et al., 2015). Die Lemmatisierung verfolgt das Konzept, das ebenfalls beim Stemming angewendet wird. Unterschiedlich dabei ist, dass Lemmatisierung die Wortart betrachtet. Dabei wird für Substantive die Einzahl und für Verben die Grundform als Repräsentationen verwendet. Diese werden als Lemma bezeichnet (Egger & Gokce, 2022).

Das PoS Tagging beschreibt eine Methode, die den Wörtern Label zuweist. Als Label wird die Wortart verwendet (Manning & Schutze, 1999). Eine Wortart fasst Wörter hinsichtlich ihrer grammatikalischen Merkmale zusammen. Betrachtet werden kann zudem die Verwendungsweise der Wörter (Srinivasa-Desikan, 2018).

Das folgende Beispiel ist der Datengrundlage (Lange & Jentsch, 2023a, 2023b) soll die verschiedenen Vorverarbeitungsschritte verdeutlichen.

> **Originaltext:** "Herr Kollege Ritter, ich rede nicht in Rätseln, ich rede offensichtlich nur über Dinge, die Sie fachlich nicht ganz so sehr interessieren."
>
> **Ergebnis Vorverarbeitung:** [(' herr ', 'NOUN'), (' kollege ', 'NOUN'), (' ritter ', 'NOUN'), (' reden ', 'VERB'), (' rätsel ', 'NOUN'), (' reden ', 'VERB'), (' offensichtlich ', 'ADJ'), (' ding ', 'NOUN'), (' fachlich ', 'ADJ'), (' interessieren ', 'VERB')]

Das Ergebnis der Vorverarbeitung setzt sich aus der Verwendung folgender Methoden zusammen: Entfernung von Satzzeichen, Stoppwörtern und Zahlen sowie Tokenisierung, Lemmatisierung, Kleinschreibung und PoS-Tagging. Der Beispieltext ist der Datengrundlage (Lange & Jentsch, 2023a, 2023b) entnommen und steht repräsentativ für die untersuchte politische Domäne. Aufgebaut ist das Ergebnis für den vorliegenden Anwendungsfall aus einer Liste von Tupeln. Ein Tupel besteht dabei im ersten Teil aus dem Lemma und im zweiten Teil aus dem PoS-Tag. Die Lemmatisierung ist bei den Wörtern 'rede', 'Rätseln' und 'Dinge' aus dem originalen Text ersichtlich. Diese sind umgeformt in 'reden', 'rätsel' und 'ding'.

2.2 Grundlagen der Sentiment-Analyse

In diesem Abschnitt sind die grundlegenden Konzepte der Sentiment-Analyse definiert. Es erfolgt zunächst die Einordnung des Bereichs der Sentiment-Analyse, gefolgt von formalen Definitionen von Meinung und Sentiment. Relevant sind diese, da die Arbeit die Stimmung in deutschen parlamentarischen Gremien untersucht. Der letzte Themenschwerpunkt erläutert zudem die theoretischen Grundlagen der Methoden, die sowohl in der Sentiment-Analyse als auch in der praktischen Umsetzung der Untersuchung eingesetzt werden.

Die Aufgabe, natürliche Sprache, die in verschiedenen Formen vorliegt, einem Computer verständlich zu machen, ist das Ziel von NLP. Die Form wird beispielsweise durch Paragraphen, Sätze oder Wörter beschrieben. Das Konzept von NLP lässt sich in weitere Unterdisziplinen unterscheiden. Zum einen in Natural Language Generation und zum anderen in Natural Language Understanding. Die im Rahmen der Arbeit untersuchte Sentiment-Analyse kann dem zuletzt genannten Bereich zugeordnet werden (Egger & Gokce, 2022).

Das automatisierte Untersuchen von Meinungstexten beschreibt die Kernaufgabe der Sentiment-Analyse (Siegel & Alexa, 2020). Ziel ist es, Emotionen, Meinungen, Bewertungen oder Einstellungen, die Menschen in Bezug auf Entitäten haben, zu untersuchen. Beispiele für diese Entitäten sind Personen oder Produkte sowie Ereignisse oder Themen. Auch die Eigenschaften der Entitäten sollen analysiert werden (Liu, 2012, 2015). Die persönliche Position oder Interpretation wird durch ein Sentiment vermittelt. Eine Klassifizierung in neutral, negativ oder positiv ist realisierbar und wird als Polarität bezeichnet (Siegel & Alexa, 2020).

Als Sentiment ist ein Gefühl zu verstehen, das einer Meinung zugrunde liegt. Dieses kann positiv oder negativ sein (Liu, 2015, 2017). Sentimentwörter werden als Wörter definiert, die eingesetzt werden, um Gefühle zum Ausdruck zu bringen. Sie sind ein Indikator für Sentiments. 'Wunderschön' und 'gut' repräsentieren beispielhaft positive Wörter, während 'schrecklich' und 'schlecht' negative Beispiele darstellen (Liu, 2012).

Formal wird eine Meinung nach (Liu, 2012) durch ein Quintupel definiert. Dargestellt ist dies durch $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$. Die Bezeichnung der Entität ist dabei mit e_i abgebildet. Der dazugehörige Aspekt beschreibt der Ausdruck a_{ij}. Das Sentiment in Bezug auf diesen Aspekt ist s_{ijkl}. Als weiterer Bestandteil ist h_k die Person, die das Sentiment äußert. Von h_k wird die Meinung zum Zeitpunkt t_l ausgedrückt (Liu, 2012; Siegel & Alexa, 2020). Die durchgeführten Untersuchungen innerhalb der vorliegenden Arbeit betrachten nicht den Aspekt, daher wird von der Definition als Quadruple nach (Liu, 2012, 2015) ausgegangen. Dieses wird formuliert als (g,s,h,t). Dabei stellt g den Gegenstand des Sentiments dar. Mit s wird das Sentiment ausgedrückt, während h den Meinungshalter repräsentiert. Die zeitliche Komponente ist im Quadruple t. Unter dem Gegenstand eines Sentiments g ist die Eigenschaft oder die Entität selbst zu verstehen, an die das Sentiment gerichtet ist (Liu, 2012, 2015).

Das Sentiment s besteht formal aus dem Typ, der Orientierung und der Intensität. Dabei ist die Orientierung im Rahmen dieser Arbeit von Bedeutung. Diese kann neutral, negativ oder positiv sein. Die neutrale Orientierung wird dabei über das Ausbleiben eines Sentiments ausgedrückt Polarität wird als Synonym zum Wort Orientierung verwendet (Liu, 2015, 2017). Die Zuordnung einer Polarität zu einem Text definiert die Aufgabe der Sentiment-Klassifikation (Naglik & Lango, 2025). Diese wird in der vorliegenden Arbeit angewendet.

Die Sentiment-Analyse lässt sich in verschiedene Betrachtungsebenen unterteilen. Eine der Ebenen stellt die Dokumentenebene dar. Das Ziel besteht darin, ein gesamtes Dokument zu klassifizieren. Auf dieser Ebene gilt die Annahme, dass die Meinungen des Dokuments auf eine Entität bezogen und durch eine Person geäußert sind. Weiterhin kann die Satzebene betrachtet werden. Dabei erfolgt die Klassifizierung in neutral, negativ oder positiv für jeden Satz. Es besteht die Annahme, dass ausschließlich eine Meinung im Satz enthalten ist. Weiterhin gilt, dass diese von einer Person ausgedrückt wird. Die dritte Ebene bildet die Aspektebene. Hierbei liegt der Fokus nicht auf sprachlichen Elementen, sondern auf dem Gegenstand der Meinung sowie der Meinung an sich (Liu, 2012, 2015). In der vorliegenden Arbeit erfolgt die Sentiment-Analyse auf der Dokumentenebene, da für jeden Redebeitrag die Polarität bestimmt werden soll. Ein Redebeitrag kann dabei aus mehreren Sätzen aufgebaut sein.

Ein Sentiment-Wörterbuch besteht aus Wörtern, für die ihre Sentiment-Orientierung sowie ihre Intensität angegeben sind (Kirilenko et al., 2022; Liu, 2015). Ausgelesen wird der Sentiment-Score über den Abgleich zwischen Wörterbuch und einem Wort. Das Vorkommen einer Negation verändert den Wert in dessen Inversion. In Bezug auf ein Dokument kann dessen Sentiment auf zwei Arten dargestellt werden. Zum einen können die negativen und positiven Werte einzeln aufsummiert werden, zum anderen wird die Gesamtsumme gebildet (Kirilenko et al., 2022).

Die Berechnung des Sentiment-Scores, die für diese Arbeit betrachtet wird, ist in Gleichung 2.2.1 dargestellt. Die Gleichung ist aus der Arbeit von (Rauh, 2018b) entnommen und ist von ihm auf Basis der Angaben aus (Young & Soroka, 2012) aufgestellt.

$$\text{Sentiment} = \frac{\text{\# positive Begriffe - \# negative Begriffe}}{\text{\# alle Begriffe}} \tag{2.2.1}$$

Der Einsatz eines Wörterbuchs ist domänenabhängig. Grund dafür ist, dass sich der Kontext auf die Polarität auswirken kann. Um diese Abhängigkeit zu beachten, können spezifisch für diese Domäne erstellte Wörterbücher genutzt werden (Bashiri & Naderi, 2024).

Ein weiterer Ansatz besteht in der Verwendung von maschinellem Lernen. Bei diesem Ansatz wird ein gelabelter Datensatz aus einer Auswahl an Dokumenten erstellt. Mit diesen Daten wird ein Algorithmus trainiert, der anschließend die Klassifikation der nicht gelabelten Daten übernimmt. Das Sentiment ist dabei während des Trainings erlernt. Dieser Ansatz benötigt kein Sentiment-Wörterbuch (Kirilenko et al., 2022).

Die Entwicklung von Transformern ermöglicht einen weiteren Ansatz für die Verarbeitung von Daten. Der darin verwendete *self-attention*-Mechanismus realisiert die Gewichtung der Wörter durch die Modelle. Über längere Abstände lassen sich somit kontextuelle Zusammenhänge berücksichtigen. Mit dem Transformer Ansatz wird die präzisere Sentiment-Analyse möglich (Bashiri & Naderi, 2024).

Nach (Naglik & Lango, 2025) erfolgt in Ansätzen aus dem Bereich Deep Learning das Fine-Tuning vortrainierter transformer-basierter Modelle. Dabei wird ein gelabelter Datensatz verwendet (Naglik & Lango, 2025). Dieses Verfahren ist die Grundlage der im Rahmen der Arbeit betrachteten transformer-basierten Modelle. Unterschieden werden diese in den nachfolgenden Ausführungen in öffentlich verfügbare auf Sentiment-Analyse trainierte Modelle und ein eigenständiges Fine-Tuning-Modell.

2.3 Grundlagen der Erkennung von Hassrede

Dieses Grundlagenkapitel thematisiert die Erkennung von Hassrede. Zunächst soll eine Einordnung des Begriffs für das Verständnis von Hassrede dienen. Anschließend wird ein Überblick über die verschiedenen Ansätze zur Erkennung gegeben. Dieses Kapitel bildet die Grundlage für die Untersuchung der Datengrundlage auf das Vorkommen von Hassrede.

Die Erkennung von Hassrede hat sich aufgrund der Vielschichtigkeit des Begriffs für Maschine und Mensch als herausfordernd dargestellt. Problematisch ist, neben der Komplexität den Begriff zu definieren, zudem die Unklarheit im Gebrauch verwandter Ausdrücke. Zwischen diesen kommt es zu Überschneidungen und einer subjektiven Auffassung (Poletto et al., 2020).

Hassrede wird von (Parihar et al., 2021) als Inhalt definiert, dessen Ziel es ist, zu provozieren sowie Beleidigungen oder Bedrohungen zu äußern. Angriffspunkte stellen u.a. Religion, ethnische Gruppen oder die sexuelle Orientierung dar (Parihar et al., 2021).

Die Vereinten Nationen haben in ihrem Strategieplan zum Thema Hassrede eine Definition aufgestellt. Demnach ist unter Hassrede eine Form der Kommunikation zu verstehen, die Gruppen oder Individuen in Bezug auf deren Identität abwertet oder diskriminiert. Zu den Formen der Kommunikation zählen Schrift, Wort und das Verhalten. Faktoren der Identität sind beispielsweise die Nationalität, die Religion, das Geschlecht, die ethnische Zugehörigkeit sowie die Abstammung (United Nations, n. d.).

Das folgende Beispiel ist der Datengrundlage (Lange & Jentsch, 2023a, 2023b) entnommen und veranschaulicht eine Darstellung für Hassrede in der politischen Domäne.

> **Beispiel:** Heinrich-Wilhelm Ronsöhr [CDU/CSU]: Sie haben mit Schröder den grössten Penner!

Ein Forschungsgebiet in NLP, mit gestiegenem Interesse, stellt die automatisierte Identifikation von Hassrede dar (Fortuna et al., 2022). Nach (A. Schmidt & Wiegand, 2017) besteht eine enge Verbindung zwischen der Sentiment-Analyse und der Erkennung von Hassrede. Weiterhin gehen sie davon aus, dass ein Zusammenhang zwischen Hassrede und negativer Stimmung existiert (A. Schmidt & Wiegand, 2017).

Eine Methode, Hassrede zu erkennen, liegt in der Zuhilfenahme eines Wörterbuches oder einer Ontologie. Diese ermöglichen die Erkennung von Texten mit vermeintlich als Hassrede verstandenen Schlagwörtern (MacAvaney et al., 2019). In (Malik et al., 2024) unterteilen sich die verschiedenen Methoden, die für die Erkennung von Hassrede eingesetzt werden, in zwei Kategorien. Zum einen die klassischen Methoden der Klassifizierung und zum anderen Ansätze mit Deep Learning. Dabei differenzieren sie die Deep Learning Ansätze weiter in Methoden, basierend auf Word Embeddings und auf Transformern (Malik et al., 2024). Das Aufkommen von LLMs bietet weitere Möglichkeiten, die Erkennung von Hassrede zu fördern. Dabei zeigt sich eine Verbesserung des Kontextverständnisses und der Genauigkeit (Albladi et al., 2025). Im Rahmen dieser Arbeit ist der Ansatz unter Verwendung eines Wörterbuches sowie transformer-basierter Methoden und LLMs von Interesse. Verschiedene Formen zur Erkennung von Hassrede existieren: binär, mehrere Klassen, mehrsprachig, textuell, und multimodal. Bei der binären Klassifikation wird unterschieden, ob es sich um Hassrede handelt

oder nicht (Gandhi et al., 2024). Diese Form der Erkennung wird in der vorliegenden Arbeit angewendet.

(Weissenbacher & Kruschwitz, 2024) gehen in ihrer Forschungsarbeit von dem Ausdruck *Hass und beleidigende Sprache* aus. Sie verstehen diesen als eine breite Bezeichnung. Ihre Entscheidung begründen die Verfassenden mit (A. Schmidt & Wiegand, 2017) (Weissenbacher & Kruschwitz, 2024). Dem Ansatz von (Weissenbacher & Kruschwitz, 2024) folgend bezieht sich der Ausdruck *Erkennung von Hassrede* im Rahmen dieser Arbeit auf Hassrede und beleidigende Sprache.

2.4 Grundlagen transformerbasierter-Modelle

Der folgende Abschnitt führt zunächst in die allgemeine Architektur von Transformern ein und schafft damit die Grundlage für die weiteren Ausführungen. Anschließend werden die technischen Hintergründe zu BERT, ELECTRA und LLaMA erläutert. Diese bilden die Basis der im Methodenteil ausgewählten transformer-basierten Modelle. Die Darstellung soll ein Verständnis für die Funktionsweise der verwendeten Modelle ermöglichen.

2.4.1 Architektur

Transformer sind eine Architektur, die auf dem Mechanismus *Attention* basieren. Der *self-attention*-Mechanismus beschreibt das Verfahren, indem bei einer einzelnen Abfolge mehrere Positionen gleichzeitig miteinander in Verbindung gebracht werden (Vaswani et al., 2017).

Die Struktur eines Transformers ist aus den Komponenten Encoder und Decoder aufgebaut. Mit Hilfe der im Encoder erstellten Repräsentationen der Eingabe, erzeugt der Decoder die Ausgabe. Das Modell arbeitet auto-regressiv. Dabei wird als zusätzliche Information die bereits erzeugte Ausgabe mit einbezogen (Vaswani et al., 2017).

Im Modell werden sowohl für den Decoder als auch für den Encoder die Komponenten punktweise, vollständig vernetzter Schichten sowie der *self-attention*-Mechanismus verwendet. Diese Komponenten werden in jeweils mehreren Schichten im Encoder und Decoder aufeinanderfolgend angeordnet. Der Decoder wird ergänzt, indem die Ausgabe des Encoders in einer weiteren *Multi-Head Attention*-Schicht verarbeitet wird. Weiterhin stellt die Maskierung der *self-attention* im Decoder sowie die Verschiebung in den Embeddings der Ausgabe sicher, dass ausschließlich die Informationen vorheriger Ausgaben in die Betrachtung einfließen. Nach jeder Komponente in den Schichten von Encoder und Decoder, ist *residual connection* implementiert (Vaswani et al., 2017). Es erfolgt die Verbindung der Ausgabe einer Komponente mit der Ausgabe der Identitätsabbildung der Eingabe (K. He et al., 2016). Anschließend erfolgt eine Schichtnormalisierung (Vaswani et al., 2017). Dabei werden Varianz und Mittelwert aus allen Eingaben einer Schicht bestimmt. Die Eingaben werden zuvor aufsummiert und gehören zu einem Trainingsbeispiel (Ba et al., 2016).

Ein weiterer Bestandteil der Transformer ist die *Attention Funktion*. Diese bildet die Zuordnung von Schlüssel-Wert Paaren sowie einer Anfrage auf eine Ausgabe ab. Dabei sind alle Komponenten als Vektoren dargestellt. Eingesetzt wird eine *Attention Funktion*, die auf der Basis des Kreuzprodukts arbeitet. Für den Umgang mit großen Dimensionen der Schlüssel

und Anfragen wird das Kreuzprodukt skaliert. Die Schlüssel, Werte und Anfragen werden als Matrizen dargestellt. Berechnet wird die *Attention Funktion* entsprechend der Gleichung 2.4.1. Die Variablen Q, K und V bilden dabei die Matrizen der Anfragen, Schlüssel und Werte ab. Mit d_k wird die Dimension der Schlüssel k dargestellt (Vaswani et al., 2017).

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.4.1}$$

In der Multi-Head Attention-Schicht erfolgt die parallelisierte Berechnung der Attention Funktion. Genutzt wird dabei die lineare Projektion der Werte, Schlüssel und Anfragen. Das Ergebnis entsteht, indem die Ausgaben vereinigt werden. Danach wird erneut eine lineare Projektion angewendet (Vaswani et al., 2017). Die Gleichung 2.4.2 drückt dies mathematisch aus.

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \\ \text{where} \quad \text{head}_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{2.4.2}$$

Die Matrizen $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}, W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}, W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$und $W^O \in \mathbb{R}^{hd_v} \times d_{\text{model}}$ stellen dabei die linearen Projektionen dar. Der Parameter *head* repräsentiert einen Kopf, in dem die Berechnung der Attention Funktion durchgeführt wird (Vaswani et al., 2017). Ein Feedforward Neuronales Netz mit vollständig verbundenen Schichten ist Bestandteil der Decoder und Encoder Schichten. Die Aktivierung erfolgt mit der Rectified linear unit (ReLU) Funktion. Jedes Element wird einzeln durch das Neuronale Netz verarbeitet. Informationen über die Reihenfolge innerhalb einer Sequenz werden mit *Positional Encodings* abgebildet. Diese werden in den Encoder und Decoder Schichten integriert. Die Modellarchitektur ist in Abbildung 2.1 skizziert (Vaswani et al., 2017).

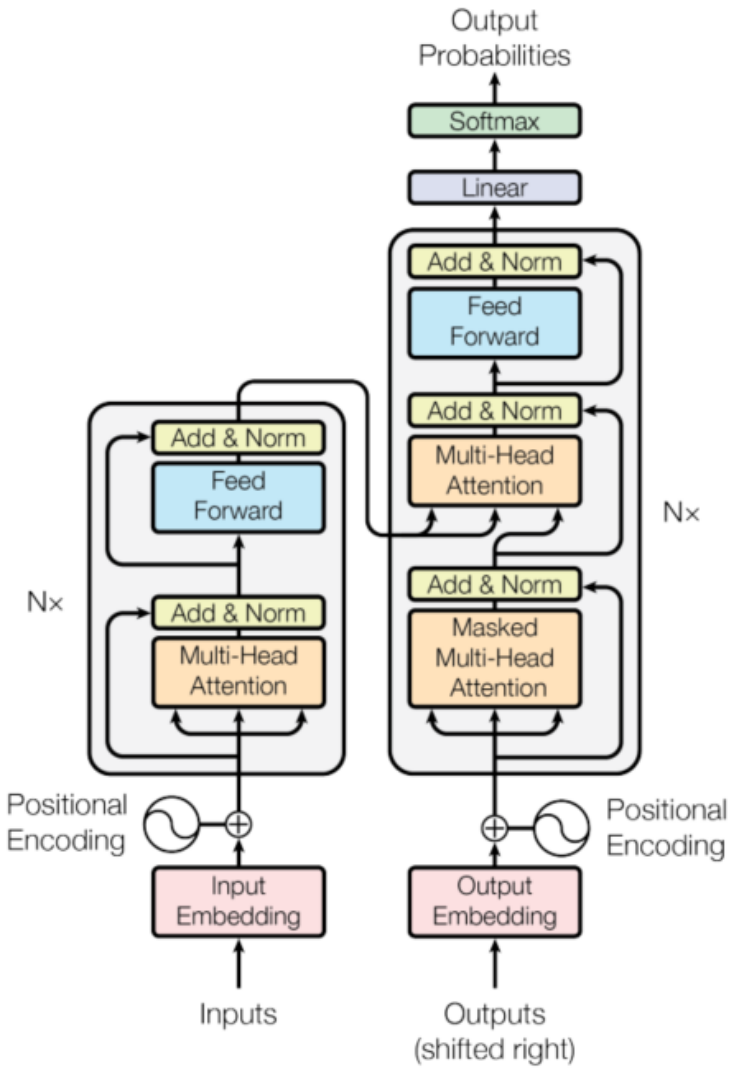


Abbildung 2.1: Transformer Architektur entnommen aus (Vaswani et al., 2017)

2.4.2 Grundlagen von BERT

Die theoretischen Hintergründe zu BERT sind notwendig, um die Arbeitsweise der ausgewählten Modelle in der durchgeführten Untersuchung zu veranschaulichen.

BERT steht für Bidirectional Encoder Representations from Transformers. Die Entwicklung von BERT ist durch die Schritte Vortrainieren und Finetunen realisiert. In der Phase des Vortrainierens erfolgt das Training auf verschiedene Aufgaben. Die Datengrundlage bilden dabei nicht annotierte Daten. Die in dieser Phase erzeugten Parameter werden für die Initialisierung des Modells im Fine-Tuning Prozess genutzt. Die Parameter werden in dieser Phase optimiert. Hierfür werden Trainingsdaten, die für die spezifische Aufgabe geeignet sind, eingesetzt. Ein Modell wird für jede spezifische Aufgabe erstellt. Die Initialisierungsparameter aus dem Prozess des Vortrainierens sind bei allen Modellen identisch (Devlin et al., 2019).

Ein Transformer Encoder, der mehrschichtig und bidirektional arbeitet, wird in der Modell-Architektur angewendet. Der Aufbau der verwendeten Transformer folgt den Ausführungen von (Vaswani et al., 2017). Der self-attention Mechanismus in den Transformern des BERT Modells ist bidirektional. Bidirektionalität ist durch Masked Language Modeling (MLM) gegeben, indem rechter und linker Kontext kombiniert dargestellt werden können. Eine Abfolge in BERT wird als Token Sequenz bezeichnet. Diese kann aus mehreren Sätzen aufgebaut sein. Der Start einer Abfolge bildet ein *special classification token*. Dieser wird als *[CLS]* repräsentiert. Um die einzelnen Sätze innerhalb der Abfolge unterscheiden zu können, wird der *[SEP]* Token benutzt. Zusätzlich erfolgt die Zuweisung der Zugehörigkeit jedes Tokens über ein erlerntes Embedding. Daraus ergibt sich die Darstellung für jeden Token als Zusammensetzung aus den Embeddings für sich selbst, dem Segment sowie der Position (Devlin et al., 2019). Unterstützend veranschaulicht ist dies in Abbildung 2.2.

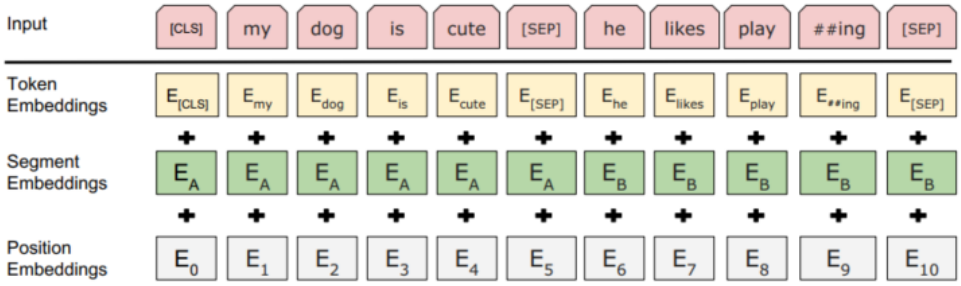


Abbildung 2.2: Darstellung der Eingabe in BERT entnommen aus (Devlin et al., 2019)

Vortrainiert wird BERT auf die beiden Aufgaben: MLM sowie Next Sentence Prediction (NSP). Für die erste Aufgabe werden Token maskiert. Die Auswahl erfolgt dabei randomisiert. Der insgesamt zu maskierende Anteil einer Abfolge liegt bei 15 %. Ausschließlich diese verschleierten Token sollen vorhergesagt werden. Der dafür eingesetzte Token *[MASK]* wird nicht im Finetuning Prozess genutzt. Dadurch entsteht zwischen den beiden Phasen eine Diskrepanz. Umgangen wird das Problem, in dem 80% der zu maskierenden Wörter mit dem Token *[MASK]* ersetzt werden. Ansonsten wird zu je 10% entweder keine Veränderung vorgenommen oder ein zufälliger Token verwendet (Devlin et al., 2019).

Der zweite Aufgabenbereich im Prozess des Vortrainierens ist NSP. Ziel ist es dabei, ein Modell, mit Verständnis für die Beziehungen zwischen Sätzen zu erhalten. Das Training

erfolgt binär, indem der Satz B zu 50% der korrekte Nachfolger von Satz A ist (Devlin et al., 2019).

Für die Auswahl der Trainingsdaten ist es wichtig, dass der Datensatz auf der Dokumentenebene basiert. Im Fine-Tuning Prozess erhält das Modell die Ein- und Ausgaben für die gewünschte Aufgabe. Danach erfolgt das Fine-Tuning der Parameter. Bei diesem Vorgehen wird dem vortrainierten Modell eine Schicht zur Klassifikation hinzugefügt (Devlin et al., 2019).

2.4.3 Grundagen von ELECTRA

Die im Folgenden dargelegten theoretischen Konzepte zu ELECTRA geben einen Einblick in den Aufbau und die Besonderheiten des Modells. Es dient zur Veranschaulichung der theoretischen Hintergründe des selbst trainierten Fine-Tuning-Modells.

ELECTRA basiert auf der Vorhersage ersetzter Token. Ziel ist es, zwischen einer generierten und der wahrhaftigen Eingabe zu differenzieren. Dabei werden Token aus der Eingabe mit einem Token, erzeugt durch eine Verteilung, ausgetauscht. Das Netz wird auf die Vorhersage der Echtheit eines Tokens vortrainiert. Es ist recheneffizient, da der Lernprozess auf den gesamten Tokens der Eingabe beruht (Clark et al., 2020).

Die Realisierung von ELECTRA erfolgt durch zwei Neuronale Netze. Trainiert wird dabei ein Netz, dass das Generieren der Tokens übernimmt und ein weiteres Netz für die Vorhersage. Der Hauptbestandteil beider Netze sind Encoder. Ein Encoder realisiert die Abbildung der Eingabe auf eine Vektordarstellung mit Kontext. Die Wahrscheinlichkeit, dass ein spezifisches Token generiert wird, erfolgt über die in Gleichung 2.4.3 dargestellte Softmax-Schicht. Zur Vorhersage über die Echtheit des Tokens wird die Sigmoid-Schicht (Gleichung 2.4.4) eingesetzt (Clark et al., 2020).

$$p_G(x_t|x) = \frac{\exp(e(x_t)^\top h_G(\mathbf{x})_t)}{\sum\limits_{x'} \exp(e(x')^\top h_G(\mathbf{x})_t)} \tag{2.4.3}$$

$$D(\mathbf{x}, t) = sigmoid(w_T h_D(\mathbf{x}_t)) \tag{2.4.4}$$

Die Variable x stellt dabei die ursprüngliche Eingabe dar. Die Indizes G und D symbolisieren das Generator Neuronale Netz (G) sowie das Neuronale Netz für die Vorhersage (D). Die genannte Vektordarstellung mit Kontext ist durch $h(x)$ mit dem jeweiligen Index des Neuronalen Netzes repräsentiert. Die spezifische Position wird durch die Variable t in den Gleichungen dargestellt. Eine Abbildung der *token embeddings* ist über die Variable e realisiert (Clark et al., 2020).

Das Neuronale Netz für die Generierung der Tokens ist ein auf MLM trainiertes Modell. Die Position der zu maskierenden Tokens werden zufällig ausgewählt. Nach dem Ersetzen der Tokens an den entsprechenden Stellen erlernt das Netz die originalen Token. Das Training erfolgt auf Basis der Maximum-Likelihood Schätzung (Clark et al., 2020).

Das zweite Netz wird darauf trainiert, zu unterscheiden, ob ein Token generiert wurde oder original ist. Es soll die Token identifizieren, die in einer veränderten Eingabe den wahren Eingaben entsprechen (Clark et al., 2020).

Die formale Definition von ELECTRA ist folgenden Gleichungen zu entnehmen:

$$m_i \sim \text{unif}(\{1, n\}), \quad \text{für } i = 1 \text{ to k} \tag{2.4.5}$$

$$x^{masked} = REPLACE(\mathbf{x}, \mathbf{m}, [MASK]) \tag{2.4.6}$$

$$\hat{x}_i \sim p_G(x_i | x^{masked}) \quad i \in m \tag{2.4.7}$$

$$x^{corrupt} = REPLACE(\mathbf{x}, \mathbf{m}, \hat{\mathbf{x}}) \tag{2.4.8}$$

Durch m werden die Positionen der zu maskierenden Tokens abgebildet. Die Eingabe mit maskierten Tokens repräsentiert x^{masked}. Durch $x^{corrupt}$ ist die Eingabe dargestellt, in der generierte Tokens ($\hat{x}$) an Positionen der Maskierten eingesetzt werden (Clark et al., 2020). Abschließend wird der Verlust beider Netze in Kombination minimiert. Nachdem Beenden des Vortrainierens, kann das Neuronale Netz, das die Vorhersage durchführt, auf spezifische Aufgaben optimiert werden (Clark et al., 2020).

2.4.4 LLaMA

Dieser Abschnitt erläutert die Besonderheiten des LLaMA Sprachmodells. Dieses bildet die Grundlage eines verwendeten Modells zur Bearbeitung der Forschungsfrage.

LLaMA wurde basierend auf Transformern trainiert. Die Datengrundlage des Trainings bilden verschiedene Datensätze aus vielfältigen Bereichen. Die Auswahl der Daten wird darauf eingeschränkt, dass ausschließlich frei verfügbare Daten einbezogen werden. Der Byte Pair Encoding (BPE) führt die Tokenisierung durch. Der ursprüngliche Algorithmus nach (Gage, 1994) für die Kompression von Daten wird dabei von (Sennrich et al., 2016) angepasst. Durch die Anpassung ist die Segmentierung von Wörtern realisierbar (Sennrich et al., 2016). Für das Training stehen danach 1.4T (Billionen) Token zur Verfügung (Touvron et al., 2023).

Die Transformer Architektur wird in drei Punkten modifiziert. Zum einen erfolgt die Normalisierung der Eingabe und nicht der Ausgabe der Schichten in dieser Struktur der Architektur. Zum anderen wird die Aktivierungsfunktion modifiziert, indem Swish Gated Linear Units (SwiGLU) (Shazeer, 2020) verwendet wird. Dabei handelt es sich um eine Aktivierungsfunktion. Die letzte Modifikation betrifft die *positional embeddings*. Dabei sollen die absoluten durch rotierende *positional embeddings* ausgetauscht werden (Touvron et al., 2023). Eingesetzt wird eine Rotationsmatrix. Zusätzlich erfolgt die Integration relativer Abhängigkeiten der Positionen in die Abbildung der self-attention (Su et al., 2024).

2.5 Evaluationsmetriken

Für die Auswahl der geeigneten Methode auf Basis der Datengrundlage werden diese evaluiert. Im Folgenden sind daher die theoretischen Grundlagen der verwendeten Evaluationsmetriken aufgeführt.

Unter Recall sind die möglichen Antwortmöglichkeiten, die wahr sind, zu verstehen. Precision beschreibt den Anteil tatsächlicher Antworten, die korrekt sind (Hinojosa Lee et al., 2024). Wird eine 100-prozentige Precision angegeben, ist mit Sicherheit davon auszugehen, dass alle positiven Beispiele tatsächlich positiv sind (Wankhade et al., 2022). Die Kombination der

beiden Metriken beschreibt die F - Metrik. Diese Metrik liegt zwischen Precision und Recall. Es erfolgt die Gewichtung von Precision und Recall unter Verwendung relativer Gewichte (Chinchor, 1992). Der mathematische Zusammenhang ist in der Gleichung 2.5.1 dargestellt.

$$F = \frac{(\beta^2 + 1.0) \cdot P \cdot R}{\beta^2 \cdot P + R} \tag{2.5.1}$$

Die Precision ist durch P und der Recall durch R abgebildet. Der Parameter β beschreibt die Gewichtung, die dem Recall gegenüber der Precision zugewiesen wird. Ein Wert von 1 für β bedeutet die Gleichgewichtung der beiden Metriken (Chinchor, 1992).

Für die F1-Metrik gibt es die Varianten micro, macro sowie gewichtet. Bei der micro F1-Metrik werden für die Berechnung die Gesamtheit der Klassen betrachtet. In der macro Variante erfolgt die Berechnung für jede Klasse individuell. Anschließend wird der Mittelwert gebildet. Jede Klasse ist zu gleichen Teilen beachtet. Die gewichtete Variante der F1-Metrik unterscheidet sich zur macro Berechnung in der Gewichtung. Es wird für jede Klasse der Anteil an wahren Ausprägungen als Gewicht betrachtet. Die mathematischen Zusammenhänge sind in den Gleichungen 2.5.2 - 2.5.4 dargestellt (Hinojosa Lee et al., 2024).

$$F1_{micro} = \frac{2 \cdot \text{Precision}_{micro} \cdot \text{Recall}_{micro}}{\text{Precision}_{micro} + \text{Recall}_{micro}} \tag{2.5.2}$$

$$F1_{macro} = \frac{1}{N} \sum_{i=1}^{N} \frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{2.5.3}$$

$$F1_{weighted} = \sum_{i=1}^{N} w_i \frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{2.5.4}$$

Durch w_i wird für das Label an Index i die relative Häufigkeit richtig klassifizierter Fälle repräsentiert (Hinojosa Lee et al., 2024).

Mit der Accuracy wird die Anzahl an korrekten Vorhersagen gemessen. Das Verhältnis von zutreffenden Vorhersagen, zu der Menge an Datenpunkten wird dadurch beschrieben. Der mathematische Zusammenhang ist in der Gleichung 2.5.5 dargestellt (Zheng, 2015).

$$\text{accuracy} = \frac{\#\text{ richtige Vorhersagen}}{\#\text{ Datenpunkte}} \tag{2.5.5}$$

Unter einer Konfusionsmatrix ist eine Matrix zu verstehen, die quadratisch ist. Die Größe ergibt sich aus der Anzahl der Rückgabeklassen. In den Zeilen werden die Häufigkeit des Vorkommens der wahren Klassen abgebildet. Die Häufigkeiten der vorhergesagten Klassen sind in den Spalten dargestellt. Dadurch ist für jede Klasse eine Aussage über die Korrektheit der Vorhersagen ermöglicht (Sathyanarayanan & Tantri, 2024).

2.6 Zeitreihenanalyse

Der vorliegende Abschnitt beschreibt zunächst die formale Definition einer Zeitreihe und widmet sich anschließend zwei zentralen Themenbereichen: der Zeitreihenzerlegung und der Erkennung von Ausreißern. Diese theoretischen Grundlagen werden im Zusammenhang mit der

Visualisierung und Analyse der Ergebnisse aus den angewendeten Methoden von Sentiment-Analyse und Erkennung von Hassrede benötigt. Darüber hinaus bilden sie eine wichtige Basis für die Untersuchung der aufgestellten sozialwissenschaftlichen Fragestellungen.

Die Gesamtheit der Beobachtungen x_t definiert eine Zeitreihe. t repräsentiert dabei den spezifischen Zeitpunkt der Beobachtung. Sind diese Zeitpunkte diskret, bezeichnet dies eine diskrete Zeitreihe (Brockwell & Davis, 2002).

Zeitreihenmuster werden in Trend, zyklisch und saisonal unterschieden. Unter einem Trend ist das Fallen und Steigen der Zeitreihe über einen längeren Zeitraum zu verstehen. Ein Wechsel zwischen An- und Abstiegsphasen wird als Trend bezeichnet, der seine Richtung verändert. Bei einem zyklischen Muster in den Daten ist ebenfalls ein An- und Absteigen erkennbar. Der Zeitraum, in dem diese Schwankungen auftreten, ist dabei nicht festgelegt. Folgen Schwankungen saisonalen Faktoren wie beispielsweise wöchentlich oder jährlich, ist dies ein saisonales Muster. Der Zeitraum ist dabei bekannt und festgelegt (Hyndman & Athanasopoulos, 2014).

Eine Zeitreihe ist durch die Komponenten 'trend-zyklisch', saisonal und den Residuen beschrieben. Es wird dabei in multiplikative und additive Zerlegung unterschieden.

Die additive Zerlegung kann in vier Schritten skizziert werden. Im ersten Schritt ist zunächst zu unterscheiden, ob die saisonale Komponente gerade oder ungerade ist. Davon ist die Berechnung des trend-zyklischen Anteils abhängig. Der gleitende Mittelwert wird über Daten innerhalb der festgelegten Periode berechnet. Bei einer geraden Periode erfolgt danach die erneute Berechnung des gleitenden Mittelwertes, in diesem Fall dann immer über zwei Datenpunkte. Darauf folgt der zweite Schritt, in dem die trendbereinigte Zeitreihe ermittelt wird. Dabei wird die Differenz zwischen der Zeitreihe und der im ersten Schritt berechneten trend-zyklischen Komponente ermittelt. Die saisonale Komponente wird im dritten Schritt berechnet. Gebildet wird sie aus der Aneinanderreihung der festgelegten Perioden. Im letzten Schritt werden die in den Schritten eins und drei aufgestellten Komponenten von der Zeitreihe subtrahiert. Daraus ergibt sich der Bestandteil, der die Residuen darstellt. Bei der multiplikativen Zerlegung besteht der Unterschied darin, dass die Division an den Stellen der Subtraktion angewendet wird. Für andere Modelle, die die Zeitreihenzerlegung abbilden, ist diese klassische Zerlegung die Basis (Hyndman & Athanasopoulos, 2014).

Im Rahmen dieser Arbeit sind besonders das Trendmuster und die trend-zyklische Komponente für die Analyse der aufgestellten sozialwissenschaftlichen Fragestellungen von Bedeutung. Dadurch erklären sich die Thematisierung von Mustern in Zeitreihen sowie die theoretischen Grundlagen der Zeitreihenzerlegung.

Nachfolgend werden die wichtigsten Punkte aus dem Bereich der Identifizierung von Ausreißern in Zeitreihendaten erläutert.

(Blázquez-García et al., 2021) definieren Ausreißer auf Basis der von (Hawkins, 1980) aufgestellten Beschreibung als Beobachtungen, die vom angenommenen Verhalten abweichen. In ihrer Forschungsarbeit unterscheiden sie weiterhin in drei Arten von Ausreißern: Punkt, Teilfolge und Zeitreihe (Blázquez-García et al., 2021). Im Rahmen der vorliegenden Arbeit ist der Punkt-Ausreißer von Bedeutung und wird daher eingehender betrachtet. Dieser wird als Datenpunkt definiert, dessen Verhalten im Vergleich abweicht. Dabei wird von einem konkreten Zeitpunkt ausgegangen. Der Vergleich erfolgt global oder lokal. Global werden die

gesamten Werte der Zeitreihen betrachtet, während der lokale Abgleich die Nachbarn einbezieht. Es ist zwischen multivariat und univariant zu unterscheiden. Für univariante Ausreißer sind die Methoden zur Erkennung in modellbasiert, dichtebasiert oder unter Verwendung eines Histogramms zu unterscheiden. Die im Rahmen der Arbeit relevante Methode folgt dem Konzept, das die Abweichung vom erwarteten Wert betrachtet. Übersteigt die Abweichung einen Schwellenwert wird dieser Punkt als Ausreißer angesehen. Der Wert des Schwellenwertes wird zuvor definiert. Dieses Vorgehen ist dem modellbasierten Ansatz zuzuordnen (Blázquez-García et al., 2021).

Ein robustes Streuungsmaß stellt der Quartilsabstand dar. Dieser wird als Differenz des oberem Quartils vom unteren Quartil definiert (Henze, 2013). Die Relevanz des Quartilsabstands für den vorliegenden Kontext liegt in der Erkennung von Ausreißern.

Für die Zeitreihenzerlegung wird die Seasonal Trend decomposition using LOESS (STL)-Methode durch eine Bibliothek in Python verwendet. Im Folgenden soll das grundlegende Konzept der Methode beschrieben werden. Die STL-Methode stellt ein Verfahren zur Zeitreihenzerlegung dar. Sie ist im Allgemeinen das wiederholte Durchführen der Locally Estimated Scatterplot Smoothing (LOESS) Glättung. Berechnet wird LOESS, indem zu Beginn eine Variable q festgelegt wird. Diese ist positiv und ein ganzzahliger Wert. Dieser Wert beschreibt die Anzahl an nächstgelegenen x_i, die um den Punkt x ausgewählt werden. Anschließend wird jedem ein Gewicht zugewiesen. Der Wert dessen beruht auf dem Abstand. Bezeichnet wird das Gewicht als *Nachbarschaftsgewicht*. Befindet sich x_i in einem kleinen Abstand zu x, erhält x_i einen hohen Gewichtswert. Anschließend erfolgt die Anpassung eines Polynoms unter Verwendung der berechneten Gewichte. Dessen Grad d kann ausgewählt werden. Die Regressionskurve wird glatter mit steigendem Wert von q. Ein weiteres Gewicht ρ_i ermöglicht die Umsetzung von Robustheit in der Methode. Dieses bildet die Aussagesicherheit einer Beobachtung in Bezug auf die restlichen Beobachtungen ab. Aufgebaut ist die Methode aus einer äußeren Schleife, die eine weitere innere Schleife enthält. Beide sind rekursiv. Wird die innere Schleife ausgeführt, erfolgt die Anpassung der Trendkomponente und der Komponente der Saisonalität. Der vollständige Durchlauf dieser Schleife wird durch das n-malige Ausführen des beschriebenen Prozesses abgebildet. Für die äußere Schleife werden auf den Durchlauf der inneren folgend, die Gewichte für die Robustheit berechnet. Benutzt werden die Gewichte im darauffolgenden Zyklus in der inneren Schleife. Für die erste Ausführung der äußeren Schleife sind die Gewichte der Robustheit mit dem Wert 1 initialisiert (Cleveland et al., 1990).

Mit der Hilfe von Zeitreihen lassen sich Ausreißer identifizieren. Diese können im Zusammenhang zu Ereignissen stehen, die zu Änderungen in der Stimmung führen könnten (Giachanou & Crestani, 2016). Aus dieser Erkenntnis begründet sich die Betrachtung von Ausreißern in der vorliegenden Untersuchung.

3 Stand der Forschung

Die Ausführungen zum Forschungsstand dienen dazu, die gegenwärtigen methodischen Ansätze und zentralen Erkenntnisse im Kontext der untersuchten Forschungsfrage darzustellen. Der folgende Abschnitt beginnt zunächst mit einem Einblick in den Forschungsstand in den Bereichen der Sentiment-Analyse und der Erkennung von Hassrede. Aufgeführt sind weiterhin Arbeiten, die sozialwissenschaftliche Fragestellungen in Verbindung mit politischen Daten untersuchen. Von Interesse sind insbesondere Arbeiten, die die Analyse von Parlamentsdaten thematisieren.

Mehrere wissenschaftliche Arbeiten identifizieren den transformer-basierten Ansatz als Stand der Forschung (Kawintiranon & Singh, 2022; T. Schmidt et al., 2022; Wankhade et al., 2022; Widmann & Wich, 2022). Sprachmodelle, die diesen Ansatz nutzen, sind in einigen Aufgabenbereichen von NLP als aktuell gängigste Methode aufgeführt (Kawintiranon & Singh, 2022; Widmann & Wich, 2022). In der Forschung zu NLP werden sie als grundlegender Bestandteil angesehen (Widmann & Wich, 2022). Die zugrunde liegende Architektur der Transformer wird als effektivste in der Sprachmodellierung beschrieben (P. He et al., 2020). Darüber hinaus zeigt der Einsatz von LLMs, die ebenfalls auf der Transformer-Architektur basieren, Verbesserungen im Aufgabenbereich von NLP (Raiaan et al., 2024).

Der Bereich der Sentiment-Analyse stellt ein breites Forschungsgebiet dar, in dem eine Vielzahl an Ansätzen und Modellen entwickelt werden. Die Arbeit von (Tan et al., 2023) bietet einen breitgefächerten Überblick über verschiedene Ansätze und Datensätze. Zudem sind weiterführende Forschungsrichtungen dargelegt (Tan et al., 2023). Die Arbeit von (Zhang et al., 2023) gibt einen Einblick in die Anwendung von LLMs im Bereich der Sentiment-Analyse. Der Einsatz im Bereich der Politikwissenschaft wird durch die Forschungsarbeit von (Li et al., 2024) eingehender betrachtet.

Einen breiten Überblick in die aktuelle Forschung im Bereich der Erkennung von Hassrede geben die Arbeiten von (Gandhi et al., 2024) und (Alkomah & Ma, 2022). Modelle, die auf Transformern basieren, sind die vielversprechendsten in der Erkennung beleidigender und hasserfüllter Sprache (Weissenbacher & Kruschwitz, 2024). Die Arbeit von (Albladi et al., 2025) gibt weiterhin einen umfangreichen Überblick über den Einsatz von LLMs, bei der Erkennung von Hassrede. Dabei zeigt sich, dass LLMs die Erkennung bereichert haben. Sie erreichen Erfolge bei der Steigerung von Kontextverständnis und Genauigkeit (Albladi et al., 2025).

Die nachfolgenden Ausführungen thematisieren Forschungsarbeiten, die für die praktische Durchführung relevant sind. Zu Beginn sind Arbeiten aufgeführt, die besonders für den methodischen Ansatz von Interesse sind.

Die Arbeit von (T. Schmidt et al., 2022) untersucht mit Methoden der Sentiment-Analyse Beiträge von Parteien und Politikern auf Twitter. Schwerpunkt ist dabei die Bundestagswahl 2021. Für ihre Untersuchung wählen sie neben der wörterbuch-basierten Methode Methoden

des maschinellen Lernens sowie ein deutschsprachiges BERT Modell. Eine ihrer Forschungsfragen untersucht die Eignung der Methoden in ihrem spezifischen Anwendungsfall. Mit dem besten Modell führen sie weitere Analysen durch. Die Beiträge von Twitter stammen aus der Zeitspanne von Januar bis Dezember des Jahres 2021. Der erstellte Datensatz umfasst 58.864 Beiträge. Die Evaluation der Methoden zeigt, dass das BERT-Modell die höchste Performanz erreicht. Dieses Modell wurde sowohl mit den annotierten Daten des eigens erstellten Datensatzes als auch mit zusätzlichen Daten aus einem öffentlichen Datensatz trainiert. Die durchgeführte Datenanalyse betrachtet unter anderem den Verlauf der Sentimente in Bezug auf das Wahljahr 2021. Die Analyse erfolgt differenziert für die Parteien (T. Schmidt et al., 2022). Diese Forschungsarbeit ist für die praktische Durchführung relevant, da verschiedene Methoden der Sentiment-Analyse für den spezifischen Anwendungsfall evaluiert werden. Insbesondere der Einsatz eines Wörterbuches und das Fine-tuning eines Modells sind für den Methodenteil 5 dieser Arbeit bedeutsam.

(Widmann & Wich, 2022) untersuchen verschiedene Methoden, Emotionen zu messen. Sie beziehen sich dabei auf diskrete Emotionen mit einer Datengrundlage, bestehend aus politischen Texten. Diese politischen Texte beinhalten unter anderem Parlamentsdaten. Bestandteil der Untersuchung sind neben einem transformer-basierten Modell ein selbst erstelltes Wörterbuch und ein Neuronales Netz unter Verwendung von Word Embeddings. Für das Training des transformer-basierten Modells wird das deutschsprachig trainierte ELECTRA - Modell, der *German NLP Group* ausgewählt (Widmann & Wich, 2022). Im Rahmen dieser Arbeit wird für das erstellte Fine-Tuning Modell, ebenfalls dieses Modell als Basis verwendet.

(Abercrombie & Batista-Navarro, 2020b) stellen einen annotierten Parlamentsdatensatz für die Sentiment-Analyse zur Verfügung. Dieser beinhaltet die Daten aus dem britischen Parlament. Die Intention besteht darin, für die politische Domäne Daten zur Weiterentwicklung der Sentiment-Analyse bereitzustellen. In den anschließend durchgeführten Experimenten werden die Debatten des britischen Parlaments analysiert. Dabei erfolgt der Vergleich mehrerer Ansätze. Neben der Implementierung von Methoden aus dem Bereich des maschinellen Lernens wird ein BERT Modell eingesetzt. Dieses wird mit den Parlamentsdaten weitertrainiert. (Abercrombie & Batista-Navarro, 2020b) kommen zu den Ergebnissen, dass das trainierte BERT Modell nicht zwangsläufig die besten Ergebnisse erzielt. Unter der Berücksichtigung der Dauer des Trainingsprozesses, die für die transformer-basierte Methode deutlich länger ist, sehen sie keinen Mehrwert in diesem Ansatz (Abercrombie & Batista-Navarro, 2020b). Diese Arbeit ist für den methodischen Teil der vorliegenden Untersuchung relevant, da der Ansatz getestet wird, ein transformer-basiertes Modell mit Parlamentsdaten weiterzutrainieren.

(Giachanou & Crestani, 2016) untersuchen in ihrer Arbeit Methoden der Zeitreihenanalyse. Ein Schwerpunkt ihrer Forschungsarbeit liegt in der Untersuchung, inwieweit diese Methoden in der Analyse von Stimmungs- und Thementrends eingesetzt werden können. Weiterhin wird die Erkennung von Ausreißern in Bezug auf die Anwendung zur Nachverfolgung der Stimmung betrachtet. Es soll gezeigt werden, ob diese Methode anwendbar ist, um Ursachen für Änderungen in den Sentimenten zu erkennen. Die Berechnung der Residuen, für die Erkennung von Ausreißern, erfolgt durch LOESS in Verbindung mit dem Interquartilsabstand. Die Datengrundlage bilden Daten aus Twitter, die über einen Zeitraum von neun Monaten erhoben wurden (Giachanou & Crestani, 2016). Sie kommen zu der Erkenntnis, dass ihre Ansätze eine Basis für weiterführende Entwicklungen sind. Von Interesse ist diese Arbeit, da die Verbindung zwischen Sentiment-Analyse und Zeitreihenanalyse dargestellt wird.

Weiterhin wird die beschriebene Methode zur Erkennung von Ausreißern im Rahmen der Arbeit eingesetzt.

Die folgenden Ausführungen thematisieren Forschungsarbeiten, die sich ebenfalls mit der Untersuchung von sozialwissenschaftlichen Fragestellungen unter Verwendung politischer Texte beschäftigen. Ein Fokus liegt zudem auf der Untersuchung von Parlamentsdaten in den Arbeiten. Einen umfangreichen Einblick in die Untersuchung von Debatten in Parlamenten gibt die Arbeit von (Abercrombie & Batista-Navarro, 2020a).

(Rauh, 2018b) erstellt ein domänenspezifisches deutsches Wörterbuch für die politische Domäne. In einem Teil der Validierung des erstellten Wörterbuches (Rauh, 2018a, 2018b) werden ebenfalls Daten aus dem deutschen Bundestag untersucht. Betrachtet werden dabei 1.500 Sätze, die aus dem Zeitraum von 1991 bis 2013 entnommen sind. (Rauh, 2018b) kommt dabei zu dem Ergebnis, dass das Wörterbuch einen Mehrwert für Forschungsarbeiten im Bereich der Politikwissenschaften bietet (Rauh, 2018b). Von Bedeutung ist diese Forschungsarbeit, da das erstellte Wörterbuch (Rauh, 2018a, 2018b) im praktischen Teil eingesetzt wird.

(Lange & Jentsch, 2023a) führen auf dem von Ihnen erstellten SpeakGer - Datensatz eine Sentiment-Analyse durch. Ziel ihrer Untersuchung ist es, zu analysieren, ob Redebeiträge von Abgeordneten bestimmter Parteien negativer oder positiver gegenüber einem Themengebiet sind. Das ausgewählte Thema ist die Corona Pandemie. Die Umsetzung der Sentiment-Analyse erfolgt mit dem Tool Lex2Sent (Lange et al., 2024). Dieses basiert auf einem wörterbuchbasierten Ansatz und verwendet dabei zusätzlich Doc2Vec (Le & Mikolov, 2014), (Lange & Jentsch, 2023a). Als Wörterbuch entschieden sich (Lange & Jentsch, 2023a) für das von (Rauh, 2018a, 2018b) entwickelte. Betrachtet werden alle Parlamente im Zeitraum von 2020 bis 2022 (Lange & Jentsch, 2023a). Die Analyse ergab ein durchschnittlich negatives Sentiment, das parteiübergreifend festgestellt werden konnte. Bei der einzelnen Betrachtung der Parteien war erkennbar, dass die Alternative für Deutschland (AfD) in 10 von 12 Fällen die negativsten Werte aufwies (Lange & Jentsch, 2023a). Diese Forschungsarbeit ist aus zwei Gründen relevant. Zum einen wird die Sentiment-Analyse auf der Datengrundlage durchgeführt, die in der vorliegenden Arbeit verwendet wird, und zum anderen wird das Wörterbuch von (Rauh, 2018a, 2018b) in der Methode angewendet.

(Rheault et al., 2016) beschäftigen sich dabei mit dem Einfluss des wirtschaftlichen Abschwungs auf Abgeordnete. Weiterhin sollen erkennbare Veränderungen in der Polarität des Parlaments betrachtet werden. Die Datengrundlage bilden dabei Daten aus dem britischen Unterhaus. Der Untersuchungszeitraum beginnt ab 1909 und endet einschließend 2013. Für die Durchführung der Sentiment-Analyse wird ein eigens erstelltes Wörterbuch für die politische Domäne verwendet. Die Analyse hat gezeigt, dass die Debatten sich zum Positiven entwickelt haben. In einer weiteren Untersuchung stellen sie die Hypothese auf, dass bei einem wirtschaftlichen Abschwung die Negativität steigt. Hierfür wird ein Zeitreihenmodell eingesetzt. Die Untersuchung ergibt, dass ein Zusammenhang zwischen Auf- und Abschwungphasen der Wirtschaft und dem Vorkommen emotional konnotierter Wörter besteht (Rheault et al., 2016).

Die von (Rheault et al., 2016) beschriebene Methode, wird von (Lehtosalo & Nerbonne, 2020) in ihrer Forschungsarbeit aufgegriffen. Die Sentiment-Analyse erfolgt dabei auf den Daten des finnischen Parlaments. Ein Ziel besteht darin, zu untersuchen, ob die Ergebnisse in anderen Sprachen reproduzierbar sind.

Zusätzlich zur Trendanalyse sowie der Analyse des Zusammenhangs zwischen Wirtschaft und Polarität soll der mögliche Einfluss von weiteren Ereignissen betrachtet werden. Die Daten wurden ab dem Jahr 1907 erhoben und umfassen alle finnischsprachigen Redebeiträge des Parlaments. Insgesamt ergeben sich 168 Millionen Token im Datenkorpus. Um die in (Rheault et al., 2016) dargelegte Methode anzuwenden, war es notwendig, ein politisches Wörterbuch der finnischen Sprache zu erstellen (Lehtosalo & Nerbonne, 2020). konnten ebenfalls einen steigenden Trend feststellen, der ab der Mitte des 20. Jahrhunderts einsetzt. Weiterhin ergibt die Forschung, dass zwischen Polarität und Inflation ein Zusammenhang ersichtlich ist. Andere untersuchte Einflussgrößen wie 'Arbeitskämpfe', 'misery-index', die Pandemie und der Beginn des russischen Angriffs auf die Ukraine führen nicht zu eindeutigen Ergebnissen (Lehtosalo & Nerbonne, 2020).

Eine weitere Forschungsarbeit auf der Basis von Redebeiträgen des Bundestages ist (Erhard et al., 2024). In ihrer Arbeit entwickeln sie ein transformer-basiertes Modell, das populistische Sprache erfasst. Den dafür benötigten Datensatz erstellen sie aus Redebeiträgen. Es wird dafür der Zeitraum von 2013 bis 2021 betrachtet. Die Klassifizierung erfolgt in: linke oder rechte Ideologie, Anti-Elitismus und Menschenzentriertheit (Erhard et al., 2024). Aufgeführt wird diese Forschungsarbeit, da Reden aus dem deutschen Parlament verwendet werden. Zudem stellt es ein Beispiel für die Untersuchung in Richtung beleidigender Sprache dar.

(Pätz et al., 2025) zeigen eine sehr aktuelle Forschungsarbeit mit Parlamentsdaten. Die Untersuchung erfolgt auf Daten aus dem Deutschen Bundestag. Eingesetzt werden dabei Methoden des maschinellen Lernens für die Klassifikation von Sentimenten und Themen. Für beide Aufgabenbereiche werden drei Modelle analysiert und das beste zur Analyse angewendet. Im Bereich der Sentiment-Analyse wird die Klassifizierung mit einem Random Forest Modell ausgeführt. Die Untersuchung von Themen wird mit Bagging umgesetzt. Ziel der durchgeführten Datenanalyse ist, die Verteilung der Stimmung und Änderungen im Themenbereich differenziert nach Parteien zu betrachten. Die Ergebnisse der Arbeit verdeutlichen Zusammenhänge bezüglich der Rolle, die die Parteien im Parlament einnehmen (Pätz et al., 2025). Diese Forschungsarbeit ist aufgeführt, da Daten aus dem Deutschen Bundestag betrachtet werden und im methodischen Vorgehen zunächst eine Evaluation durchgeführt wird.

Die Relevanz dieser Arbeit ergibt sich aus einer Kombination der im Forschungsstand dargestellten Ansätze. Demnach soll die Idee, sozialwissenschaftliche Fragestellungen mit den Methoden der Sentiment-Analyse zu betrachten, wie sie von verschiedenen Forschungsarbeiten angewendet wird, für die deutschen Gremien erweitert und mit neueren Methoden untersucht werden (Lange & Jentsch, 2023a; Lehtosalo & Nerbonne, 2020; Pätz et al., 2025; Rheault et al., 2016). Im Mittelpunkt stehen transformer-basierte Modelle und LLMs.

Darüber hinaus wird neben der Bundesebene ebenfalls die Länderebene einbezogen. Landtagsprotokolle und Debatten werden in der Forschung zur Analyse von Parlamentsdaten nur selten berücksichtigt. Im Fokus der deutschsprachigen Analyse steht hauptsächlich der Bundestag (Erhard et al., 2024; Pätz et al., 2025). Ansätze, die Daten der Landtage zur Verfügung zu stellen, sind die Datensätze *SpeakGer* (Lange & Jentsch, 2023a, 2023b) oder auch *GerParCor* (Abrami et al., 2024). Ziel ist es, dadurch Unterschiede zwischen Regionen zu betrachten (Lange & Jentsch, 2023a). Die Untersuchung umfasst daher vier ausgewählte Bundesländer, die sowohl westdeutsche als auch ostdeutsche Regionen repräsentieren.

Ergänzend wird die Erkennung von Hassrede als unterstützende Analysemethode herangezogen. Die Besonderheit liegt darin, dass Methoden der Erkennung von Hassrede auf Parlamentsdaten von deutschen Gremien angewendet werden. Die Erkennung von Hassrede in der politischen Domäne wird häufig im Rahmen der Untersuchung von politischem Diskurs in den sozialen Netzen angewendet. In Bezug auf die deutsche Sprache und Politik sei hierbei die Arbeit von (Weissenbacher & Kruschwitz, 2024) zu nennen. Weitere Forschungsarbeiten, die diese Aussage verdeutlichen, stammen beispielsweise von (Agarwal et al., 2021; De Oliveira et al., 2024) oder (Solovev & Pröllochs, 2022) Mit der vorliegenden Arbeit soll daher der Einsatz der Erkennung von Hassrede in der Untersuchung des Stimmungsbildes von Debatten in Gremien berücksichtigt werden.

4 Datengrundlage

Im folgenden Abschnitt erfolgt die Beschreibung der beiden Datensätze, die zur Durchführung der praktischen Arbeit genutzt werden. Dabei soll zunächst der Datensatz detaillierter betrachtet werden, der für die Analyse der sozialwissenschaftlichen Fragestellungen angewendet wird. Dieser bildet ebenfalls die Grundlage für die Evaluation der untersuchten Methoden. Weiterhin wird der Datensatz näher betrachtet, der als Trainingsdatensatz im Fine-Tuning Modell dient.

4.1 SpeakGer Datensatz

Eine Datengrundlage bildet der von (Lange & Jentsch, 2023a) entwickelte Datenkorpus *SpeakGer*. Veröffentlicht wurde dieser im September 2023 und liegt aktuell in der zweiten Version vor (Lange & Jentsch, 2023a, 2023b). Er besteht aus 19 einzelnen .csv Dateien. 17 dieser Dateien beinhalten die Plenarprotokolle der einzelnen Bundesländer sowie des Bundestages. Zwei weitere Dateien enthalten Metadaten-Informationen. Die Dateien der Plenarprotokolle beschreiben die Daten mit den folgenden Spalte: *Periode, Session, Date, Chair, Interjection, Member Parliament ID (MPID), Party, Constituency* und *Speech*. Über die Spalten *Session* und *Periode* wird für die Rede die jeweilige Sitzung und Wahlperiode angegeben. Auf diese beiden Spalten folgt *Date*. Sie beinhaltet das Datum der Sitzung. Die Spalte *Chair* gibt über True bzw. False an, ob der Redebeitrag von der Sitzungsleitung stammt. Aussage darüber, ob ein Redebeitrag ein Zwischenruf bzw. Anmerkung aus dem Plenum ist, gibt die Spalte *Interjection* vom Type boolean an. *MPID* repräsentiert eine individuelle Identifikationsnummer der Sprecher. Über diese Nummer können weitere Informationen aus den zusätzlichen Metadaten Dateien zugeordnet werden. *Party* ordnet dem jeweiligen Redner seine Parteizugehörigkeit zu und *Constituency* den entsprechenden Wahlkreis, den der Redner repräsentiert (Lange & Jentsch, 2023a, 2023b).

Die Daten sind nicht annotiert oder vorverarbeitet. Eine Vorverarbeitung erfolgt bewusst nicht, um eine individuell angepasste Vorverarbeitung zu ermöglichen (Lange & Jentsch, 2023a). In der Tabelle 4.1 sind für alle Dateien der Bundesländer sowie der Datei des Bundestages, das Start- und Enddatum im Datensatz aufgelistet. Dabei ist zu erkennen, dass in den Bundesländern, die den ehemaligen Osten Deutschlands bildeten, die Aufzeichnungen erst ab Oktober 1990 beginnen. Eine weitere Auffälligkeit ist für das Bundesland Niedersachsen zu erkennen, die vorhandenen Plenarprotokolle enden im Jahr 2007.

Bei der weiteren Verarbeitung der Daten ist zu beachten, dass zusammengehörige Redebeiträge aufgeteilt als eigenständige Einträge vorliegen können. Der Fall entsteht, wenn eine Unterbrechung in der Rede erfolgt. Die Fortsetzung einer unterbrochenen Rede befindet sich im Datensatz in einer neuen Zeile. Ein zusammengehörender Redebeitrag wird dadurch als voneinander getrennte Reden angesehen (Lange & Jentsch, 2023a).

Die Verfassenden (Lange & Jentsch, 2023a) gehen auf Einschränkungen und Grenzen ihres Datensatzes ein. Dazu zählt, dass Redebeiträge von Gästen nicht der Person zugeordnet werden können. Der Grund liegt in der Verwendung von Metadaten-Informationen, die bei der Zuweisung verwendet werden. Gäste sind in dieser Datei nicht aufgeführt. Weitere Schwierigkeiten resultieren vor allem aus der Bildqualität von eingescannten Protokollen. Beispielsweise konnten nicht alle Beiträge eindeutig zugeordnet werden. Außerdem lag für die Bundesländer Schleswig-Holstein, Niedersachsen sowie Berlin eine unvollständige Aufzeichnung der Datumsangaben, an denen die Plenarsitzung stattgefunden hat, vor. Mit Hilfe von Recherchearbeit konnten diese abgeschätzt werden (Lange & Jentsch, 2023a).

Der Prozess der Aufteilung in die Redebeiträge beginnt, indem Ende und Anfang des Dokuments entfernt werden. Hierbei werden reguläre Ausdrücke verwendet. Eine heuristische Methode kommt zum Einsatz, wenn kein Treffer erzielt wird. Diese besteht darin, 1000 Zeilen sowohl vom Anfang als auch vom Ende zu entfernen. Anschließend erfolgt die Aufteilung erneut mit regulären Ausdrücken sowie Metadaten-Informationen. Die Identifikation eines Redebeitrages ist über den Nachnamen von Abgeordneten, der vor einem Doppelpunkt steht, möglich. Eine Intervention ist über das Auftreten von runden oder eckigen Klammern um einen Text identifizierbar. Redebeiträge der Sitzungsleitung werden zusätzlich markiert (Lange & Jentsch, 2023a).

Name	**Startdatum**	**Enddatum**	**Anzahl Einträge**
Baden-Württemberg	unbekannt	20.04.2023	1916617
Bayern	21.12.1946	27.10.2022	1480251
Berlin	12.01.1951	29.06.2023	1178872
Brandenburg	26.10.1990	26.01.2023	454488
Bremen	08.11.1967	11.08.2022	1075950
Bundestag	07.09.1949	26.04.2023	5192475
Hamburg	13.04.1966	15.06.2022	1253089
Hessen	19.12.1946	06.12.2022	1752915
Mecklenburg-Vorpommern	26.10.1990	21.03.2023	1071801
Niedersachsen	10.07.1974	13.04.2007	1252308
Nordrhein-Westfalen	19.05.1947	02.06.2023	2078417
Rheinland-Pfalz	04.06.1947	29.03.2023	1016403
Saarland	14.10.1947	05.04.2023	540682
Sachsen-Anhalt	28.10.1990	27.04.2023	507281
Sachsen	27.10.1990	13.04.2023	621756
Schleswig-Holstein	08.05.1947	12.05.2023	1476623
Thüringen	25.10.1990	09.05.2023	622894

Tabelle 4.1: Übersicht des SpeakGer Datensatz entwickelt von (Lange & Jentsch, 2023a)

4.2 Trainingsdatensatz des Fine-Tuning Modells

In diesem Abschnitt wird der zum Training des Fine-Tuning Modells verwendete Datensatz beschrieben. Dabei soll auf dessen Besonderheiten eingegangen werden. Ziel ist es, einen Überblick über die Trainingsdaten des Modells zu erhalten.

(Haselmayer & Jenny, 2020a, 2020b) stellen einen annotierten Datensatz, für wissenschaftliche Zwecke, zur Verfügung. Dieser ist für die deutschsprachige Sentiment-Analyse in Bezug auf die politische Kommunikation erstellt. In dem Datensatz sind 125.871 Sätze aufgeführt. Diese entstammen Pressestatements österreichischer Parteien sowie dem österreichischen Parlament. Daten aus dem Parlament werden dabei randomisiert aus dem Zeitintervall von 1995 bis 2013 gezogen, während die Pressestatements sich auf einen sechs-wöchigen Zeitraum vor den Wahlen im angegebenen Zeitintervall beziehen. Für die Label wird eine Skala zwischen 0 und 4 verwendet. Der Wert '0' symbolisiert dabei die Annotation 'nicht negativ' während der Wert '4' 'sehr stark negativ' repräsentiert. Für den Prozess der Annotation sind zehn Personen zuständig. Außerdem gibt es ein weiteres Label mit dem Wert '99'. In diese Kategorie entfallen alle Sätze, die als nicht codierbar einkategorisiert werden (Haselmayer & Jenny, 2017, 2020a). Aufgebaut ist der Datensatz aus fünf Spalten, bezeichnet mit *v1, v2, v3, v4* und *v5*. Die Spalten *v1* und *v2* sowie *v5* enthalten Identifikationswerte. Von Interesse im Rahmen der Arbeit sind die Spalten *v3* und *v4*. Diese enthalten die Daten, die für das Training verwendet werden sollen. Zum einen die Sätze (v3) und zum anderen die zugewiesenen Label (v4) (Haselmayer & Jenny, 2017, 2020a).

Die Verteilung der Labels ist in Abbildung 4.1 dargestellt. Das Balkendiagramm zeigt die absolute Häufigkeit der Labels in den Trainingsdaten. Daraus wird deutlich, dass das Label mit dem Wert '3' am häufigsten vorkommt. Dieser repräsentiert Sätze, die als stark negativ eingeordnet sind (Haselmayer & Jenny, 2020a).

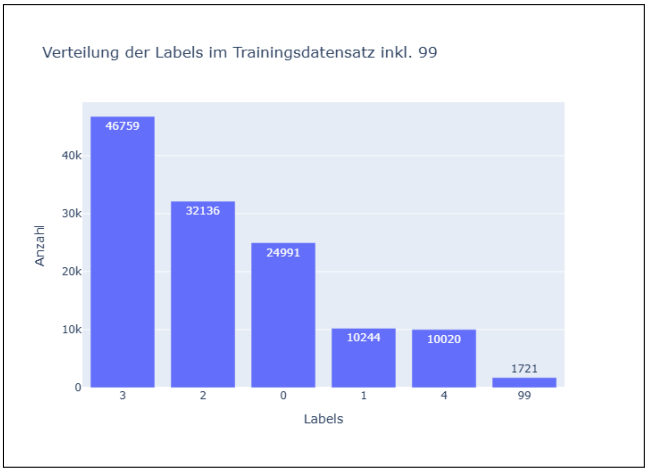


Abbildung 4.1: Balkendiagramm, das die Verteilung der Label im Trainingsdatensatz (Haselmayer & Jenny, 2020b) des Fine-Tuning Modells darstellt

5 Methoden

Dieses Kapitel beschreibt die im Zuge der praktischen Untersuchung vorgenommenen Schritte und bietet einen strukturierten Überblick über das methodische Vorgehen sowie die im Prozess getroffenen Entscheidungen. Zu Beginn werden die Untersuchungsmethode im Allgemeinen und die technischen Details der Implementierung erläutert. Anschließend gliedert sich das Kapitel in die durchgeführten Prozessschritte.

Grundlegend lässt sich die Untersuchungsmethode in folgende Prozessschritte unterteilen: Eingrenzen des Betrachtungszeitraums, Erstellung des Evaluationsdatensatzes sowie dessen Annotation, Vorverarbeitung der Daten, Anwendung der Methoden auf den Evaluationsdatensatz, Evaluation der Methoden, Anwendung der besten Methode und Analyse der Ergebnisse. Zudem ist die Untersuchungsmethode in zwei Schwerpunkte gegliedert. Im ersten Punkt erfolgt ein Vergleich verschiedener Methoden, um diejenige zu identifizieren, die für die vorliegende Datengrundlage die besten Ergebnisse liefert. Diese wird im zweiten Punkt auf die Analysedaten angewendet. Die Ergebnisse bilden die Grundlage für die anschließende Datenanalyse. Durchgeführt wird die Datenanalyse unter Berücksichtigung der sozialwissenschaftlichen Fragestellungen.

Die Implementierung erfolgt in der Entwicklungsumgebung Visual Studio Code. Genutzt wird die Programmiersprache Python. Unterstützend bei der Programmierung werden GitHub Copilot, Blogeinträge, Dokumentationen sowie Notebooks aus Lehrveranstaltungen genutzt. Die Entscheidung liegt darin begründet, dass hier Bibliotheken für NLP und Sentiment-Analyse unterstützt werden. Aufgrund der Größe des Datensatzes wird zusätzlich Google Colab verwendet. Die Durchführung der Vorverarbeitung und Klassifizierung auf den gesamten Daten sowie das Training des Fine-Tuning Modells werden dort umgesetzt. Dabei werden die in Visual Studio Code erstellten Skripte als Notebooks ausgeführt.

Während der Arbeit in Google Colab wird zum einen eine A100 und zum anderen eine L4 Graphics Processor Unit (GPU) angewendet. Die A100 GPU wird benutzt, wenn die Berechnung mit dieser deutlich schneller durchgeführt werden kann. Ansonsten wird auf die L4 GPU zurückgegriffen. Die Prozessschritte der Untersuchungsmethode sind im Diagramm 5.1 skizziert.

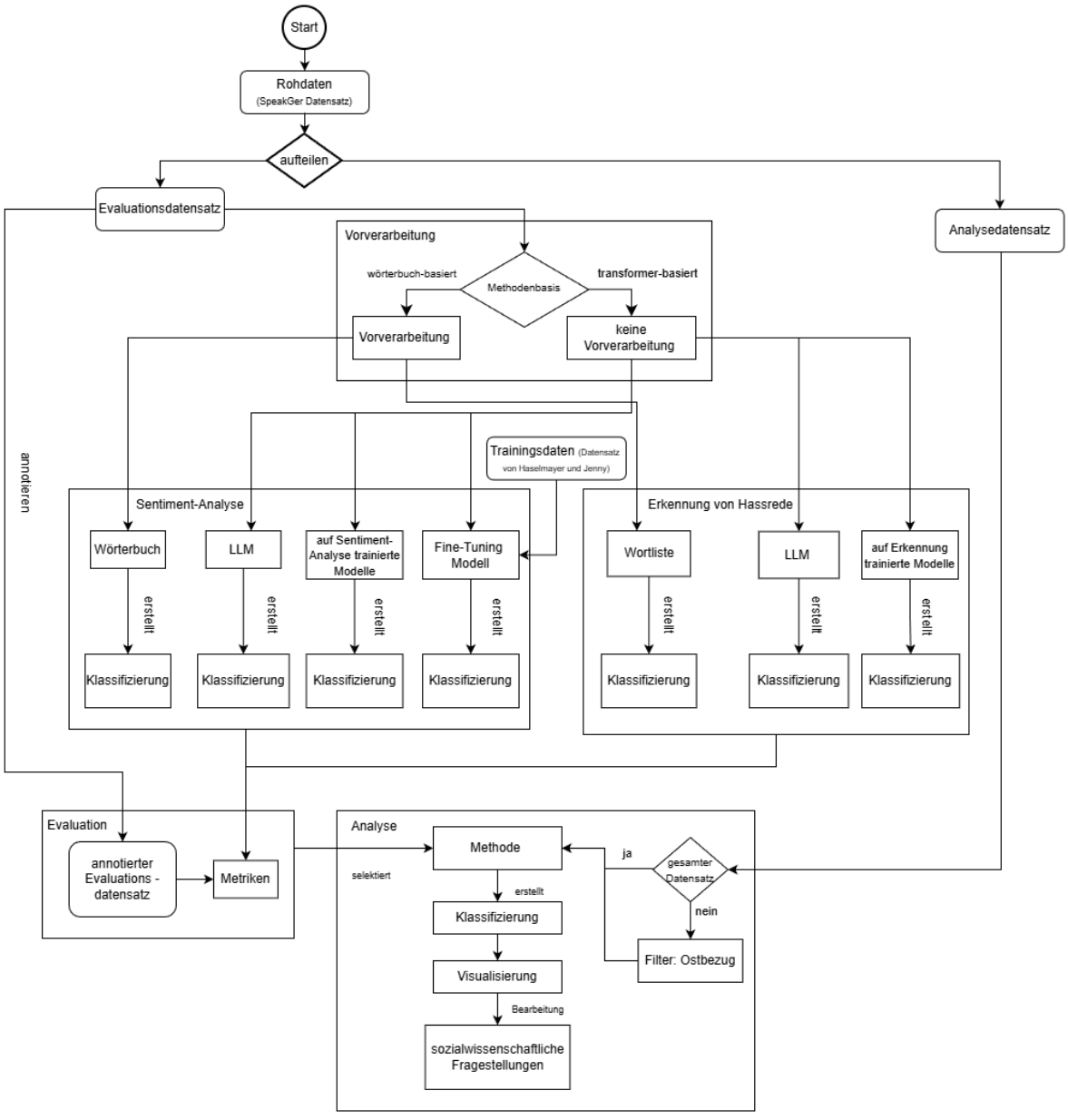


Abbildung 5.1: Flussdiagramm zur Darstellung der durchgeführten Untersuchungsmethode (erstellt mit draw.io)

5.1 Eingrenzen der Daten

Zunächst soll die Auswahl der Daten erläutert sowie die weiterführende Eingrenzung des Untersuchungszeitraums dargelegt werden. Ergebnis dieser Einschränkung sind die Daten, die zur Bearbeitung der Forschungsfrage eingesetzt werden.

Die Eingrenzung erfolgt in zwei Phasen. In der ersten wird der Zeitraum festgelegt, aus dem die Evaluationsdaten entnommen werden. Die zweite Phase grenzt die Daten auf den Zeitraum, der für die Untersuchung der sozialwissenschaftlichen Fragestellungen betrachtet wird, ein.

Die Dateien des *SpeakGer*-Datensatzes (Lange & Jentsch, 2023a) werden einzeln eingelesen und in einen Dataframe umgewandelt. Für die weitere Betrachtung werden die Spalten *Party* und *Constituency* entfernt, da diese für die Untersuchung der Forschungsfrage nicht genutzt werden. Anschließend erfolgt die Eingrenzung der Daten auf einen Intervall, beginnend mit dem 03.10.1990 bis zum 01.01.2022. Das Startdatum begründet sich daraus, dass für die ehemaligen ostdeutschen Bundesländer die Daten erst ab der Wiedervereinigung vorliegen (siehe Tabelle 4.1). Zusätzlich erfolgt das Entfernen von Einträgen, die aus einem einzelnen Buchstaben bzw. einer einzelnen Zahl, einer Zusammensetzung aus einzelnen Buchstaben oder Zahlen und runden Klammern aufgebaut sind.

Dem Datensatz hinzugefügt werden die Spalten *Jahr*, das aus dem Datum extrahiert wird und *land*. Die Spalte *land* gibt an, aus welchem Bundesland der Eintrag kommt. Diese wird benötigt, da die Datensätze im Anschluss vereinigt werden, um den Evaluationsdatensatz zu erstellen. Aufgrund der Größe der gesamten Daten werden die Datensätze der Gremien zusätzlich einzeln gespeichert. Zudem ist dadurch die getrennte Betrachtung der Gremien, im Kontext der sozialwissenschaftlichen Fragestellungen, sichergestellt. Aus der vereinigten Datenmenge werden die Einträge des Evaluationsdatensatzes ausgewählt. Anschließend erfolgt die engere Eingrenzung der Daten.

Im Rahmen dieser Arbeit soll für die Untersuchung der sozialwissenschaftlichen Fragestellungen der Zeitraum ab dem 01.01.2002 bis zum 01.01.2022 betrachtet werden. Das Ende des betrachteten Zeitintervalls wird so bestimmt, dass die vorletzte Bundestagswahl am 26.09.2021 mit eingeschlossen ist (Bundeswahlleiterin, 2025a). Für die Untersuchung werden Daten des Bundestages sowie aus vier Bundesländern berücksichtigt: Hessen, Schleswig-Holstein, Sachsen und Thüringen. Die Auswahl erfolgt so, dass jeweils zwei Bundesländer den ehemaligen Westen und zwei den ehemaligen Osten Deutschlands repräsentieren. Damit ist der Datensatz auf eine Größe von 4.048.205 Einträgen begrenzt.

Aus diesem eingegrenzten Datensatz wird eine weitere Teilmenge entnommen. Auf Basis dieser Daten ist die Untersuchung der sozialwissenschaftlichen Fragestellung (d) *Lässt sich ein negativer Bias gegenüber den Regionen des ehemaligen Ostens Deutschlands feststellen?* durchgeführt. Dabei werden Einträge erfasst, die am 03.10. eines jeden Jahres oder am nächstmöglichen Folgetermin erhoben wurden. Auf diese Weise werden gezielt Einträge ausgewählt, die einen direkten zeitlichen Bezug zum Tag der Deutschen Einheit aufweisen.

Beide eingegrenzten Datensätze bilden die Daten, die zur Analyse verwendet werden sollen, und sind folgend als Analysedaten bezeichnet.

5.2 Erstellen der Evaluationsdaten

Der nächste Prozessschritt besteht darin, dass für die Evaluation der Methoden ein Evaluationsdatensatz erstellt und annotiert werden soll.

Die Daten liegen im Datensatz in unverarbeiteter Form vor und sind nicht annotiert. Die Intention von (Lange & Jentsch, 2023a) besteht dabei darin, eine individuelle Vorverarbeitung zu ermöglichen. Durch das Fehlen von Labeln ist die Evaluation der angewendeten Methoden nicht durchführbar. Der Lösungsansatz besteht darin, einen Evaluationsdatensatz zu erstellen und diesen zu annotieren. Ein Vorgehen, dass beispielsweise in der Arbeit von (T. Schmidt et al., 2022) ebenfalls angewendet wird.

Der Evaluationsdatensatz wird in zwei Schritten erstellt. Im ersten Schritt wird ein Datensatz erzeugt, der 0.1% des ursprünglichen Datensatzes entspricht. Die Wahl dieser Größe erfolgt in der praktischen Arbeit unter Berücksichtigung der lokalen Rechenleistung. Notwendig ist dieser Schritt aufgrund der Größe des ursprünglichen Datensatzes. Die Einträge werden basierend auf dem Ansatz des geschichteten (stratified) Stichprobeverfahrens ausgewählt, um sicherzustellen, dass alle Bundesländer repräsentiert sind. Dabei erfolgt eine Einteilung in Gruppen (strata), die voneinander getrennt sind. Aus den einzelnen Gruppen erfolgt das Entnehmen von Zufallsstichproben. Die Anzahl der Stichproben wird über eine proportionale Zuordnung berechnet (Ahmed, 2024). Dieser Schritt ist angelehnt an den von (Ahmed, 2024; geeksforgeeks, 2025) beschriebenen Prozess, Stichproben zu erstellen und wird ebenfalls von (T. Schmidt et al., 2022) eingesetzt.

Der zweite Schritt in der Erstellung des Evaluationsdatensatzes besteht darin, mit Hilfe einer bereitgestellten Sentiment-Wortliste (Siegel, n. d.) die Anzahl an Sentimentwörtern für jeden Eintrag zu zählen. Ein Wert von 0 wird in der praktischen Arbeit als neutraler Beitrag interpretiert, während Werte größer als 0 ein Sentiment im Eintrag vermuten lassen. Auf Basis dieser Information wird der endgültige Evaluationsdatensatz erstellt.

Es werden zu 25 % neutrale und zu 75% Beiträge, die ein Sentiment vermuten lassen, ausgewählt. Die 25 % neutralen Beiträge werden randomisiert aus den Einträgen mit einem Wert 0 gezogen. Die restlichen Einträge für den Evaluationsdatensatz sind aus den Einträgen, deren Wert größer als 0 ist, entnommen. Die vorherig errechneten Werte werden dabei als Gewichtung eingesetzt. Ziel ist es dabei, Beiträge mit einem höheren Wert, mit einer größeren Wahrscheinlichkeit in den Evaluationsdatensatz zu übernehmen.

Idee und Ansatzpunkt für die Erstellung des beschriebenen Vorgehens, insbesondere in Schritt zwei, bilden die Ausführungen von (Cieliebak et al., 2017) zur Auswahl der Tweets in ihren Datensatz. Weiterhin nutzt der zweite Schritt erneut Ansätze des geschichteten (stratified) Stichprobenverfahrens.

Im Folgenden wird das Verfahren für die Annotation vorgestellt, das für die Sentiment-Analyse eingesetzt wird. Neben der eigenen manuellen Annotation werden keine weiteren Personen in das Verfahren eingebunden. Stattdessen wird die Annotation durch TextBlob und ein für Sentiment-Analyse trainiertes Modell durchgeführt. Das in der vorliegenden Arbeit durchgeführte Verfahren erfolgt in mehreren Schritten. Zunächst werden die Daten manuell annotiert. Dabei wird jedem Eintrag eine Kategorie zugewiesen. Als mögliche Kategorien stehen 'positiv', 'negativ' und 'neutral' sowie eine vierte Kategorie 'nicht_codierbar' zur Verfügung.

Die Einbeziehung einer Kategorie 'nicht_codierbar' erfolgt in Anlehnung an das Vorgehen in verschiedenen Forschungsarbeiten (Antypas et al., 2022; Cieliebak et al., 2017; Haselmayer & Jenny, 2017; Rudkowsky et al., 2018).

Folgende Beispiele sind den Evaluationsdaten (Auszug aus (Lange & Jentsch, 2023b)) entnommen und werden zur Verdeutlichung der Klasse 'nicht_codierbar' aufgeführt.

> **Beispiel 'nicht_codierbar' 1:** 'Drucksache 15/586'
>
> **Beispiel 'nicht_codierbar' 2:** ' Nächste Frage, Frage 17, Herr Abg. Milde. Hessischer Landtag · 17.Wahlperiode · 6. Sitzung · 13. Mai 2008 307'

In diese Kategorie entfallen weiterhin Beiträge, die ausschließlich aus Namen und Aufzählungen von Namen bestehen.

Im nächsten Schritt werden die Evaluationsdaten unter Verwendung der TextBlob Bibliothek annotiert. Mit der Bibliothek kann die Textverarbeitung für NLP Aufgaben durchgeführt werden. Diese stellt die Eigenschaft *sentiment* zur Verfügung. Als Annotation wird dabei der *polarity* Rückgabewert verwendet. Dieser Wert liegt in einem Wertebereich von -1 bis 1 (Loria & Mitwirkende, n. d.). Da als Label eine der drei Klassen 'positiv', 'neutral' oder 'negativ' zugewiesen werden soll, wird der zurückgegebene Wert in eine der Klassen umgewandelt. Die neutrale Klasse wird durch den Wert 0 repräsentiert, die positive Klasse durch alle Werte größer als 0 und die negative Klasse durch alle Werte kleiner als 0.

Die dritte Annotation wird mit einem auf Sentiment-Analyse trainierten Modell durchgeführt. Dabei wird darauf geachtet, dass dieses Modell nicht im Methodenteil verwendet wird. Ausgewählt ist das mehrsprachige Sentiment-Analyse Modell von (tabularisai et al., 2025): *'tabularisai/multilingual-sentiment-analysis'*. Dieses nutzt synthetisch generierte Daten, um das mehrsprachige Modell *'distilbert/distilbert-base-multilingual-cased'* (Sanh et al., 2019) weiterzutrainieren. Es klassifiziert die Eingabetexte in fünf Klassen. Neben 'neutral', 'positive' und 'negative' gibt es zudem die Steigerungen 'very positive' und 'very negative' (tabularisai et al., 2025). Im Rahmen dieser Arbeit werden die Klassen 'negative' und 'very negative' sowie 'positive' und 'very positive' jeweils zu einer Klasse vereinigt.

Das Vorgehen, Label zu simulieren, ist an das Verfahren von (Sazzed & Jayarathna, 2021) angelehnt. Die Verfassenden generieren Pseudolabels für den Datensatz im Training. Weiterhin hat das Vorgehen von (Moldovan, 2025) Einfluss auf das beschriebene Verfahren. (Moldovan, 2025) nutzt automatisierte Tools in Verbindung mit einem Mehrheitsentscheid, um die Label seines Trainingsdatensatzes zu erstellen.

Zwischen den manuell erstellten Annotationen und den beiden zusätzlichen Verfahren wird der Kohens Kappa Wert berechnet. Dieser ist ein verbreitete Metrik, um zwischen Annotatoren die Übereinstimmung zu untersuchen. Ziel ist es dabei die Qualität einzuschätzen (Demus et al., 2022). Der Kohens Kappa Wert hat einen Wertebereich von -1 bis 1. Liegt der Wert im positiven Bereich, ist die Übereinstimmung der Annotatoren größer als die zufällige Übereinstimmung. Ein Koeffizient im negativen Bereich bedeutet die Einigkeit der Annotatoren ist kleiner als ein zufälliger Konsens zwischen ihnen (Cohen, 1960). Zwischen den manuellen Labeln und den Annotationen mit TextBlob liegt der Wert bei 0.0502. Im Vergleich der manuellen und transformer-basierten Annotationen beträgt der Wert 0.116.

Aufgrund des niedrigen Kohens Kappa Wertes wird nach Lösungen gesucht. Bei Diskrepanzen zwischen den Ergebnissen von Sentiment-Analyse-Tools bietet der Mehrheitsentscheid eine Lösung. Das Zusammenfassen der Label in Verbindung mit einem Mehrheitsentscheid erzeugt eine Lösung, die auf Konsens beruht. Dieser steigert Zuverlässigkeit und Robustheit für die Sentiment-Analyse (Moldovan, 2025). Auch in der Arbeit von (Rauh, 2018a, 2018b) wird diese Vorgehensweise angewendet. In der vorliegenden Arbeit wird daher zwischen den drei Annotationen ein Mehrheitsentscheid durchgeführt. Die mit diesem Verfahren erzeugten Labels werden anschließend manuell untersucht. Insbesondere Einträge mit Diskrepanzen zwischen dem manuellen Label und dem Mehrheitsentscheid werden betrachtet. Die überarbeiteten Label bilden die Grundlage für die Evaluation.

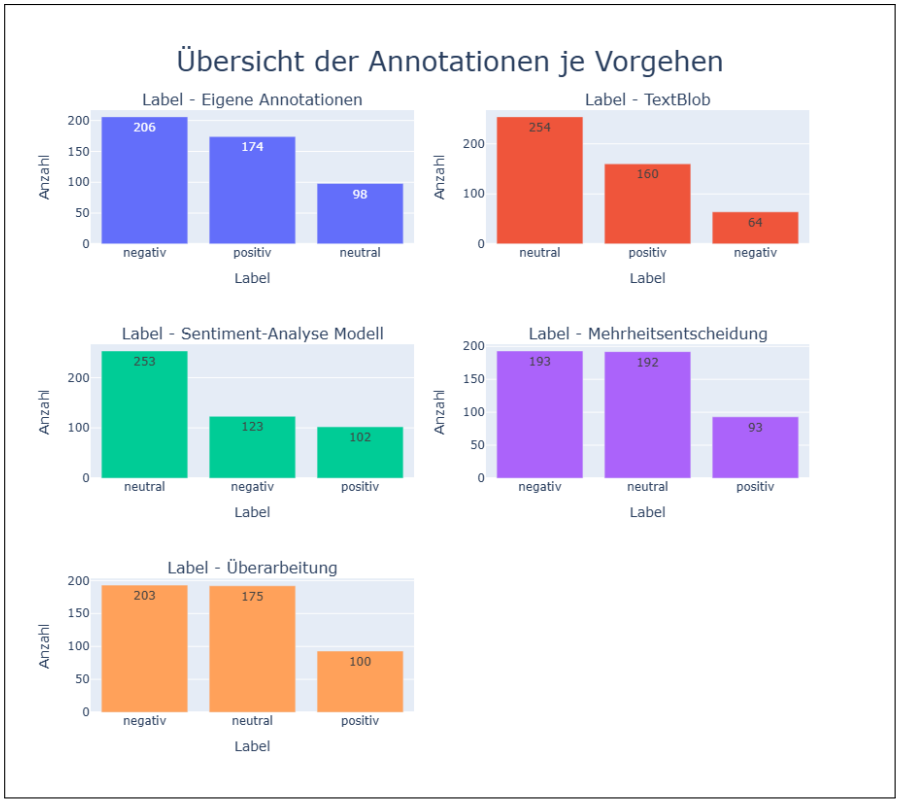


Abbildung 5.2: Darstellung der Label in den verschiedenen Vorgehensweisen der Annotation

Abbildung 5.2 zeigt die absolute Häufigkeit jeder Klasse nach den einzelnen Annotationsverfahren. Der Anstieg des Anteils neutraler Labels in den überarbeiteten Annotationen gegenüber der manuellen Klassifizierung erklärt sich durch die Beiträge, die in der Überarbeitung angepasst werden. Die größte Änderung bilden die Redebeiträge mit Beifallsbekundungen. In der ursprünglichen manuellen Annotation werden diese immer als positiv gelabelt. Bei der Überarbeitung der Labels wird sich dafür entschieden, dem Mehrheitsvotum zu folgen. Dieses ordnet Beifallsbekundungen allgemein in die neutrale Klasse ein. Der Grund für die Änderung der Entscheidung liegt in der Einschätzung, Beifall als Bestärkung des zuvor Ausgedrückten zu interpretieren. Aufgrund fehlender Kenntnis über den Kontext, in dem die Bestärkung erfolgt, wird diese als 'neutral' annotiert. Ausnahmen bilden die Fälle, in denen entweder mindestens drei Parteien oder das gesamte Parlament den Beifall äußern. Zurufe werden in der manuellen Annotation immer negativ gelabelt.

Die folgenden Beispiele sind aus den Evaluationsdaten (Auszug aus (Lange & Jentsch, 2023b)) entnommen:

> **Beispiel Zuruf 1:** Zuruf von der AfD: So ein Bullshit!
>
> **Beispiel Zuruf 2:** Zurufe von der Sozialdemokratische Partei Deutschlands (SPD)

Beide Beispiele sind im manuellen Verfahren als 'negativ' annotiert. Im ersten Beispiel wird durch den Text die Negativität deutlich. Das zweite Beispiel erhält das Label 'negativ', da es sich im Allgemeinen um einen Zuruf handelt. Diese Vorgehensweise liegt darin begründet, dass Zwischenrufe im Rahmen dieser Arbeit als Störungen oder Störversuche interpretiert werden und dies als negativ aufgefasst wird.

Für die Erkennung von Hassrede werden die Evaluationsdaten erneut manuell annotiert. Der Fokus liegt darin, für die Einträge einzuschätzen, ob Hassrede oder beleidigende Sprache vorliegt. Ausschließlich die bereits als 'negativ' gelabelten Redebeiträge werden untersucht, ob diese ebenfalls als Hassrede zu annotieren sind.

5.3 Auswahl und Implementierung der Methoden

In diesem Abschnitt werden die ausgewählten Methoden zur Untersuchung der Forschungsfrage erläutert sowie deren Implementierung beschrieben. Zunächst werden die verwendeten Vorverarbeitungsmethoden dargelegt und die Auswahl anhand von Forschungsliteratur begründet. Anschließend wird auf die Methoden der Sentiment-Analyse sowie der Erkennung von Hassrede eingegangen. Durch die Abschnitte werden die Intentionen bei der Auswahl verdeutlicht. Dabei werden sowohl die Besonderheiten der einzelnen Ansätze als auch deren praktische Implementierung dargestellt. Der Schwerpunkt bei der Bearbeitung der Forschungsfrage liegt dabei auf den Methoden der Sentiment-Analyse.

5.3.1 Vorverarbeitungsmethoden

Die nachfolgenden Ausführungen beschreiben die einzelnen Schritte der Vorverarbeitung. Dabei wird auf die Auswahl und Umsetzung der Methoden eingegangen. Die Begründung der Schritte soll den Entscheidungsprozess während der praktischen Arbeit nachvollziehbar darlegen.

Die Vorverarbeitung wird ausschließlich für die Durchführung der wörterbuch-basierten Methoden eingesetzt. Bei der Anwendung von transformer-basierten Methoden wird auf eine vorherige Vorverarbeitung verzichtet. Diese Entscheidung basiert darauf, dass (Widmann & Wich, 2022) keine manuelle Vorverarbeitung bei der Verwendung des ELECTRA Modelles eingesetzt haben. Als Begründung führen sie an, dass der Tokenizer diese Aufgabe übernimmt (Widmann & Wich, 2022). Da dieses Modell die Basis des Fine-Tuning-Modells ist, wird dem Vorgehen gefolgt. Daher wird zusätzlich bei allen transformer-basierten Modellen auf die vorherige manuelle Vorverarbeitung verzichtet.

Die Auswahl der geeigneten Vorverarbeitungsschritte basiert auf den Erkenntnissen der Forschungsarbeit von (Fehle et al., 2021). Die Verfassenden haben verschiedene Techniken wie Stemming, Stoppwörter entfernen, Lemmatisierung und weitere Methoden bei 19 Wörterbüchern getestet und auf Performanz verglichen. Dabei konnte festgestellt werden, dass *SentiMerge* (Emerson & Declerck, 2014) mit Lemmatisierung, Kleinschreibung, dem Entfernen von Stoppwörtern sowie dem Einbeziehen von Emoticons und der Verwendung kontinuierlicher Sentiment-Scores die beste Performanz zeigt (Fehle et al., 2021). Emoticons werden im Rahmen dieser Arbeit nicht betrachtet, da es sich bei der Datengrundlage um Parlamentsdaten von politischen Gremien handelt und diese nach stichprobenartiger manueller Durchsicht keine Emoticons beinhalten.

Neben *SentiMerge* (Emerson & Declerck, 2014) soll zudem der Einsatz eines domänenspezifischen Wörterbuches untersucht werden. Ausgewählt wird das von (Rauh, 2018a, 2018b) erstellte Wörterbuch. Dieses wird im Folgenden als domänenspezifisches Wörterbuch oder *Rauh* bezeichnet. Die von (Fehle et al., 2021) durchgeführte Untersuchung ergibt dabei, dass die Vorverarbeitungsmethoden Stemming, PoS-Tagging sowie Emoticons die beste Performanz erreichen. Emoticons werden analog zu *SentiMerge* (Emerson & Declerck, 2014) nicht betrachtet. Zudem wird Stemming durch Lemmatisierung ersetzt. Begründet wird dies dadurch, dass Lemmatisierung für komplexere Sprachstrukturen notwendig ist (Fehle et al., 2021). Das Entfernen von Stoppwörtern konnte in der Untersuchung von (Fehle et al., 2021) für das Wörterbuch von (Rauh, 2018a, 2018b) keine Verbesserung erreichen. Im Rahmen dieser Arbeit wird dennoch für die Anwendung dieses Wörterbuches (*Rauh*) die Stoppwörter entfernt. Eine Entscheidung, die aufgrund der Annahme, dass dadurch die Rechenleistung verringert werden kann, erfolgt (Fehle et al., 2021). Zudem erfolgt die Entfernung von Satzzeichen.

Vor der NLP-spezifischen Verarbeitung, werden häufig verwendete Redewendungen, beispielsweise zur Begrüßung der Abgeordneten, entfernt. Dies dient der Verhinderung von Ambiguitäten (Rheault, 2016). Hierfür werden verschiedene reguläre Ausdrücke definiert, die mit dem Text abgeglichen werden. Dieses Verfahren erfolgt ebenfalls in der Forschungsarbeit von (Rheault et al., 2016).

Eine Besonderheit bei der Verwendung des domänenspezifischen Wörterbuches liegt darin, dass zu Beginn die gefundenen Negationen durch ein Wort, das aus der dem Negationswort und dem eigentlichen Wort besteht, ersetzt werden sollen. Verwendet wird dabei das zur Verfügung gestellte Negations-Wörterbuch. Dieses umfasst neben den Sentiment-Scores ebenfalls die Ersetzungen (Rauh, 2018a, 2018b).

Für die Durchführung der NLP-spezifischen Vorverarbeitung wird zum einen spaCy (Honnibal et al., 2020; Montani et al., 2023; spaCy Doku, n. d.) mit dem Modell für die deutsche Sprache verwendet. Zum anderen wird Stanza (Qi et al., 2020) mit den deutschsprachigen Modellen eingesetzt. SpaCy stellt ebenfalls eine Liste für die Entfernung der Stoppwörter zur Verfügung.

In der spaCy Vorverarbeitungspipeline werden die Komponenten *tagger, parser, lemmatizer, trainable_lemmatizer, senter* und *sentencizer* geladen. Alle anderen verfügbaren Komponenten werden deaktiviert. Dadurch soll an Effizienz gewonnen werden (spaCy Usage Documentation, n. d.). Bei der von Stanza verwendeten Vorverarbeitungspipeline werden die Komponenten *tokenize, pos, lemma* und *mwt* ausgewählt. *mwt* steht dabei für *multi-word tokens* und stellt eine Voraussetzung für die *lemma* Komponente dar (Stanford NLP Group, 2020).

Der Einsatz von spaCy ebenfalls für PoS-Tags hat unzureichende Ergebnisse erzielt. Daher wird zum Erzeugen der PoS-Tags die Bibliothek Stanza (Qi et al., 2020) ausgewählt. Sie wird ebenfalls von (Fehle et al., 2021) getestet. Allerdings benötigt Stanza deutlich mehr Rechenzeit, weshalb sich gegen einen Einsatz für die gesamte Vorverarbeitung entschieden wird. Bislang wurde kein Multiprocessing in Stanza implementiert. Um die Rechenzeit zu verkürzen, wird in der Dokumentation empfohlen, mehrere Dokumente mit Hilfe des Einfügens einer leeren Zeile zu verbinden (Stanford NLP Group, n. d.).

Die Stanza-Pipeline wird ausschließlich dann eingesetzt, wenn das PoS-Tagging als Vorverarbeitungsschritt erforderlich ist. Dies wird mit Hilfe des Funktionsparameters *with_pos* der implementierten Vorverarbeitungsfunktion erreicht. In allen anderen Fällen erfolgt die Vorverarbeitung mit der spaCy-Pipeline. Auf die Evaluationsdaten werden beide Vorverarbeitungspipelines angewendet.

Die finalen Ergebnisse der Pipelines sind als neue Spalten an den ursprünglichen Datensatz angefügt. Dabei bleibt die Spalte der originalen Einträge erhalten, da diese für die transformerbasierten Modelle benötigt wird.

5.3.2 Methoden der Sentiment-Analyse

Der folgende Abschnitt untergliedert sich in die untersuchten Methoden der Sentiment-Analyse. Für jede Methode wird sowohl ihre Umsetzung in der praktischen Arbeit beschrieben als auch die Begründung für ihre Einbeziehung dargestellt.

Die Sentiment-Analyse wird zunächst auf dem eigens erstellten Evaluationsdatensatz durchgeführt. Ziel ist es, jeden Beitrag einer der Klassen positiv, negativ oder neutral zuzuordnen. Anschließend erfolgt die Evaluation der Methoden, um für die vorliegende Datengrundlage das beste Modell zu identifizieren. Dieses Modell wird danach für die weitere Untersuchung auf den Analysedaten angewendet.

Untersucht werden das wörterbuchbasierte Verfahren sowie transformer-basierte Modelle und ein LLM sowie die Kombination der Klassifizierungserbenisse. Die transformer-basierten Modelle sind unterteilt in öffentlich verfügbare Modelle, die auf Sentiment-Analyse trainiert sind, und ein eigenes erstelltes Fine-Tuning Modell.

Wörterbuch-basierte Methode

Als erste Methode ist der wörterbuch-basierte Ansatz zu untersuchen. Er kann als unüberwachtes Verfahren angesehen werden, da die Notwendigkeit von Trainingsdaten entfällt (Wankhade et al., 2022). In diesem Verfahren wird ein Wortlistenabgleich als Hauptmethode genutzt (Siegel & Alexa, 2020). Die Einbeziehung dieses Ansatzes wird durch die Ergebnisse von (Fehle et al., 2021) gestützt. Den Verfassenden zufolge ist ein wörterbuch-basierter Ansatz eine geeignete Methode, wenn Trainingsdaten entweder nicht verfügbar oder nur in begrenzter Qualität vorhanden sind. Daher kommt dieser Ansatz bei ressourcenarmen Domänen oder Sprachen zum Einsatz (Fehle et al., 2021). Da die vorliegende Arbeit Sentiment-Analyse in der politischen Domäne unter Verwendung deutschsprachiger Parlamentsdaten durchführt, begründen die vorherig beschriebenen Ausführungen den wörterbuch-basierten Ansatz.

Für die Untersuchung sind die beiden Wörterbücher *SentiMerge* (Emerson & Declerck, 2014) sowie das von (Rauh, 2018a, 2018b) erstellte Wörterbuch ausgewählt. Grundlage für diese Entscheidung bildet die Arbeit von (Fehle et al., 2021). Deren Betrachtungen ergeben, dass SenitMerge (Emerson & Declerck, 2014) die besten Ergebnisse erzielt. Aufgrund der spezifischen Domäne der in dieser Arbeit verwendeten Daten sollte ebenfalls ein Wörterbuch mit domänenspezifischen Wörtern untersucht werden. Dadurch ist die Abhängigkeit dieses Ansatzes von der Domäne beachtet (Bashiri & Naderi, 2024). Dafür wird das von (Rauh, 2018a, 2018b) erstellte Wörterbuch ausgewählt. Zudem ist es unter den besten drei Wörterbüchern in der Untersuchung von (Fehle et al., 2021) aufgelistet.

Im Folgenden sollen die Besonderheiten sowie die Zusammensetzung der ausgewählten Wörterbücher erläutert werden. Anschließend ist die Umsetzung des Ansatzes im Rahmen dieser Arbeit beschrieben.

SentiMerge (Emerson & Declerck, 2014) besteht aus vier Wörterbüchern. Aus dem von (Clematide & Klenner, 2010) erstellten Wörterbuch, sowie aus dem SentimentWortschatz von (Remus et al., 2010), dem GermanSentiSpin von (Takamura et al., 2005) und dem GermanPolarityClues von (Waltinger et al., 2010).

Beim Erstellen des Wörterbuches müssen die Sentiment-Scores zunächst normalisiert werden. Der Normalisierungsfaktor wird dabei mit Hilfe des quadratischen Mittelwertes berechnet. Dieser wird für jede Kombination aus zwei Ursprungswörterbüchern erstellt. Um den Normalisierungsfaktor für jedes Wörterbuch zu erhalten, wird der Mittelwert aus den zuvor paarweise berechneten quadratischen Mittelwerten genommen (Emerson & Declerck, 2014). entscheiden sich bei der Vereinigung der Wörterbücher für einen Ansatz nach Bayes. Die Evaluation erfolgt anhand des MLSA (Clematide et al., 2012) Datensatzes (Emerson & Declerck, 2014).

Das domänenspezifische Wörterbuch (*Rauh*) (Rauh, 2018a, 2018b) basiert auf den Wörterbüchern SentimentWortschatz (Remus et al., 2010) und dem GermanPolarityClues (Waltinger et al., 2010) Beide Wörterbücher werden im ersten Schritt des Erstellungsprozesses vereinigt. Für die Berücksichtigung von Negationen ist ein regulärer Ausdruck definiert, der die möglichen Varianten der direkten Verneinung abdeckt. Dabei wird ausschließlich die direkte Negation einbezogen. Eine Verneinung eines Wortes resultiert in das Invertieren des zugeordneten Wertes. Mögliche Werte innerhalb des Wörterbuches sind 1 und -1. Dabei repräsentiert 1 ein positives und -1 ein negatives Wort. Begriffe, die im politischen Kontext mehrdeutig sein können oder kein Sentiment aufweisen, werden verbessert oder ausgenommen. Dieser Schritt erfolgt manuell. Das Wörterbuch umfasst 37.080 Wörter, davon sind 17.330 als positiv annotiert und 19.750 negativ (Rauh, 2018a, 2018b).

Für die Implementierung des Wörterbuchansatzes erfolgt die Umsetzung auf Basis der vorverarbeiteten Daten. Für jedes Wort innerhalb der erstellten Tokenliste wird geprüft, ob dieses im Wörterbuch vorkommt. Trifft dies zu, wird aus dem Wörterbuch der zugewiesene Wert herausgesucht und aufaddiert. Dieses Vorgehen folgt dem in den theoretischen Grundlagen 2 beschrieben wörtbuch-basierten Ansatz von (Kirilenko et al., 2022). Die mit den Wörterbüchern ermittelten Werte werden in die Kategorien 'positiv', 'neutral' und 'negativ' umgewandelt. Ein Wert kleiner als 0 wird der Kategorie 'negativ' zugewiesen, ein Wert größer als 0 entsprechend der Kategorie 'positiv'. Ist der Wert gleich 0 erhält er die Kategorie 'neutral'.

Bei der Verwendung von *Rauh* (Rauh, 2018a, 2018b) sind weitere Punkte zu beachten. Die Liste der Tokens aus der Vorverarbeitungsfunktion enthält zusätzlich für jedes Lemma den zugehörigen PoS-Tag. Demnach muss neben dem Abgleich auf das Vorkommen des Lemmas im Wörterbuch ebenfalls die Übereinstimmung der PoS-Tags überprüft werden (Fehle et al., 2021). Eine weitere Besonderheit besteht darin, dass zusätzlich ein Wörterbuch für den Abgleich der Negationen zur Verfügung gestellt wird (Rauh, 2018a, 2018b).

Transformer-basierte Methode

Neben der Durchführung der wörterbuch-basierten Sentiment-Analyse werden ebenfalls transformer-basierte Modelle angewendet. Diese gelten zum aktuellen Zeitpunkt als Stand der Forschung in NLP-Bereichen (Kawintiranon & Singh, 2022; T. Schmidt et al., 2022; Wankhade et al., 2022; Widmann & Wich, 2022). Für die Domäne typische sprachliche Muster, können mit Hilfe dieser Methode erlernt werden. Dadurch erreichen sie bessere Resultate im Vergleich zu Ansätzen die Regeln oder Wörterbüchern einsetzen (Lossio-Ventura et al., 2024). Die Realisierung dieses Ansatzes erfolgt in zwei Varianten: der Einsatz öffentlich verfügbarer, für Sentiment-Analyse trainierter Modelle und die Entwicklung eines eigenen Fine-Tuning-Modells, das vorrangig mit Parlamentsdaten weitertrainiert wird.

Für die Untersuchung werden drei frei verfügbare Modelle berücksichtigt:

- *cardiffnlp/twitter-xlm-roberta-base-sentiment* (Barbieri et al., 2022)
- *cardiffnlp/xlm-twitter-politics-sentiment* (Cardiff NLP, n. d.)
- *oliverguhr/german-sentiment-bert* (Guhr et al., 2020)

Trainingsgrundlage dieser Art von Modellen bilden mehrere domänenübergreifende Datensätze. Dadurch wird eine gute Leistung und Robustheit erreicht (Naglik & Lango, 2025). Auf dieser Grundlage werden die drei Modelle in die praktische Arbeit einbezogen.

Bei dem ersten Modell wird ein cross-lingual language model (XLM)-T Modell eingesetzt. Dieses stellt ein Sprachmodell dar, das spezifisch für Twitter erstellt ist und auf einem mehrsprachigen Modell basiert. Die Trainingsdaten bildet ein Twitterdatensatz. Dieser wird verwendet, um das Modell für die Sentiment-Analyse zu trainieren. Der Datensatz umfasst acht Sprachen. Als deutscher Anteil wird der SB-10K (Cieliebak et al., 2017) benutzt (Barbieri et al., 2022). Die Einbeziehung dieses Modells begründet sich durch den vorhandenen deutschen Trainingsanteil.

Das zweite Modell stellt eine Erweiterung des zuvor beschriebenen Modells dar. Die Neuerung besteht darin, dass die Sentiment-Analyse auf die politische Domäne betrachtet wird. Dazu wird das bestehende Modell mit neuen Twitterbeiträgen von Parlamentsmitgliedern trainiert. Anzumerken ist, dass die Beiträge aus Griechenland, Großbritannien und Spanien stammen (Antypas et al., 2022; Cardiff NLP, n. d.). Obwohl keine deutschsprachigen politischen Daten einbezogen sind, wird das Modell dennoch berücksichtigt, da es auf einem mehrsprachigen Sprachmodell basiert, das deutsche Daten enthält, und eine politisch spezialisierte Ausrichtung besitzt.

Als drittes Modell wird ein deutschsprachig auf Sentiment-Analyse trainiertes Modell ausgewählt. Die Datengrundlage bilden acht Datensätze, die verschiedene Domänen abbilden. Einer dieser Datensätze weist einen politischen Bezug auf. Der *PotTS* (Sidarenka, 2016) Datensatz

umfasst Twitterbeiträge zu spezifischen Themen aus dem Jahr 2013. Zwei Themenschwerpunkte haben einen expliziten politischen Bezug: die Bundestagswahl sowie allgemeine politische Debatten. Eingesetzt wird für das Training eine vortrainierte BERT-Version. Die anschließende Verarbeitung erfolgt mit einem Feedforward Neuronalen Netz (Guhr et al., 2020).

Implementiert werden die Modelle über die HuggingFace *transformers* Bibliothek (Wolf et al., 2020). Die Inferenz erfolgt über die 'pipeline()-Funktion (Hugging Face, n. d. c). Dieser wird die Aufgabe sowie der Name des Modells übergeben (Pascual, 2022). Aufgrund der Länge einiger Einträge muss bei der Verwendung der Pipeline der *truncation* Parameter auf *True* gesetzt werden (Hugging Face, n. d. b). Die Klassifizierung der Daten erfolgt, indem die Daten der *Speech*-Spalte in die zuvor erstellte Pipeline übergeben werden. Aus der Rückgabe werden die klassifizierten Label ausgelesen.

Weiterhin wird ein transformer-basiertes Fine-Tuning Modell entwickelt. Die Begründung ergibt sich daraus, dass die zuvor ausgewählten transformer-basierten Modelle überwiegend auf Twitterdaten basieren oder nur einen geringen politischen Anteil in ihren Trainingsdaten aufweisen. Weitere Forschungsarbeiten, wie beispielsweise von (Abercrombie & Batista-Navarro, 2020b; Kawintiranon & Singh, 2022; T. Schmidt et al., 2022) und (Widmann & Wich, 2022) verwenden diesen Ansatz. Dadurch zeigt sich zusätzlich die Eignung dieses Verfahrens. Als Ausgangsmodell wird das deutschsprachig trainierte ELECTRA-Modell von *German NLP Group* (May & Reißel, 2020) ausgewählt. Angelehnt ist dieser Ansatz an das Vorgehen von (Widmann & Wich, 2022), die ebenfalls dieses Modell in der politischen Domäne trainiert haben. Die Entscheidung für ein ELECTRA-Modell basiert zum einen auf den Ergebnissen, die (Widmann & Wich, 2022) mit diesem erzielt haben. Zum anderen zeigt das verwendete deutsche ELECTRA-Modell eine bessere Leistung in dem Aufgabenbereich der deutschsprachigen Textklassifikation (Widmann & Wich, 2022).

Für das Training des Modells wird der Datensatz 4.2 von (Haselmayer & Jenny, 2020b) verwendet. Im Trainingsprozess werden die mit dem Wert '99' annotierten Sätze aus dem Datensatz entfernt, da diese Daten als nicht codierbar interpretiert werden (Haselmayer & Jenny, 2017, 2020b). Die Beschreibung des Datensatzes ist im Kapitel 4.2 dargestellt. Dieser ist zu entnehmen, dass ausschließlich negative Annotationen verwendet werden.

Für das Modelltraining erfolgt eine Reduktion der Labels auf binäre Klassen. Der Wert 0 wird dabei beibehalten und bildet im Rahmen dieser Arbeit die Klasse 'neutral' ab. Alle anderen Werte sind dem Wert '1' zugeordnet und repräsentieren die Klasse 'negativ'. Dadurch entfällt die Klasse 'positiv' und das selbst trainierte Fine-Tuning Modell kann ausschließlich in die Klassen 'neutral' oder 'negativ' klassifizieren. Der Trainingsprozess wird mithilfe der von Hugging Face bereitgestellten Trainer-Application Programming Interface (API) (Hugging Face, n. d. a, n. d. d) implementiert.

Der Trainingsprozess wird in vier separaten Durchläufen durchgeführt. Im ersten Durchgang wird der Trainingsdatensatz mit den binären Klassen als Eingabe übergeben. Die Anwendung des daraus trainierten Modells auf die Evaluationsdaten zeigt, dass alle Einträge in die Klasse 'negativ' klassifiziert werden. Der Grund liegt darin, dass im originalen Datensatz fünf Klassen existieren, während nach der Binarisierung für den Anwendungsfall zwei Klassen bleiben. Die Verteilung 5.3 der beiden Klassen deutet auf das *class imbalanced problem* hin. Darunter ist zu verstehen, dass die Aufteilung der Daten nicht gleichmäßig ist (Anjum & Katarya, 2023).

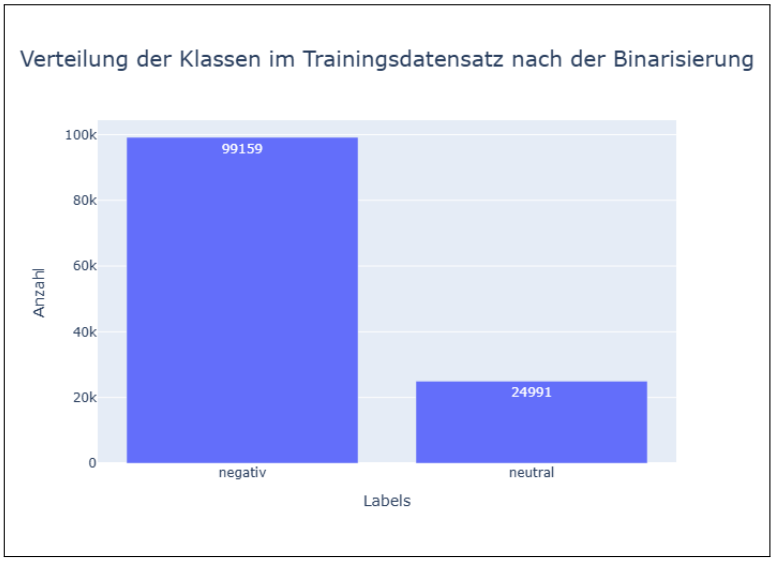


Abbildung 5.3: Visualisierung des Ungleichgewichts der beiden Klassen (neutral, negativ) in den Trainingsdaten (Haselmayer & Jenny, 2020b) nach der Binarisierung der ursprünglichen Klassen

Nach (Spelmen & Porkodi, 2018) ist eine Lösung des Problems die Methode des Oversamplings. Oversampling bedeutet, dass die Datengröße der unterrepräsentierten Klasse angehoben wird. Ziel ist dabei, dass die Klassen danach gleichmäßig verteilt sind (Rathi, 2024; Spelmen & Porkodi, 2018).

Anschließend wird die Lernrate variiert, um die Performanz zu verbessern. Die Lernrate ist als Optimierungs-Hyperparameter definiert. Dieser kontrolliert, wie weit ein Aktualisierungsschritt eines Parameters in die Gradientenrichtung ist. Sie wirkt sich erheblich auf die Modellleistung aus (Goodfellow et al., 2016). Die Variation der Lernrate wird von (Devlin et al., 2019) in den durchgeführten Experimenten ebenfalls untersucht. Die Lernrate ist zudem Gegenstand weiterer Forschungsarbeiten (Bu et al., 2024; Houlsby et al., 2019; Jin et al., 2023).

Im Rahmen der Arbeit werden drei Lernraten getestet: 5e-5, 3e-5 und 2e-5. Die Auswahl orientiert sich an den von (Devlin et al., 2019) durchgeführten Experimenten. Die Wahl von 5e-5 bildet den Ausgangspunkt, da dieser Wert sowohl der Standardeinstellung der verwendeten Trainer API (Hugging Face, n. d. d) entspricht als auch der Initialisierung in (Widmann & Wich, 2022).

LLM

Abschließend ist für die Klassifizierung der Daten ein LLM ausgewählt. Nach den Erkenntnissen aus (Zhang et al., 2023) besteht eine Empfindlichkeit gegenüber dem Prompt bei LLMs für Aufgaben, die anspruchsvolle Formate haben. LLMs zeigen ein gutes Textverständnis sowie Stärken bei Anwendungsfällen mit wenig annotierten Daten (Zhang et al., 2023). Zum einen stützen diese Erkenntnisse und der Forschungsstand 3 die Wahl eines LLMs für die Bearbeitung der Forschungsfrage, zum anderen begründen sie den Fokus auf die Gestaltung des Prompts.

Verwendet wird das *Llama 3.1 SauerkrautLM 70B Instruct*. Dabei handelt es sich um ein fine-getuntes Modell von *Meta-Llama 3.1 70B Instruct* (Meta Llama, n. d.). Im Prozess des Fine-Tunings wird die Spektrum-Methode angewendet. Bei der Spektrum Methode werden Modellschichten für das Training ausgewählt. Die anderen Schichten bleiben fixiert (Hartford et al., 2024). Es wurden 15 % der Modellschichten ausgewählt, die ins Training eingebunden werden (Hartford et al., 2024; VAGO solutions, n. d.).

Ausschlaggebend für die Entscheidung zugunsten dieses Modells war, dass auf der Modellauswahlseite die deutsche Sprache explizit in der Kategorie 'Vorteile' aufgelistet ist. Endpunkt der Wissensbasis liegt bei diesem Modell im Dezember 2023 (GWDG Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen, n. d. a).

Während der Arbeit mit dem LLM hat sich die Wichtigkeit eines klar und präzise formulierten Prompts für die zielführende Verarbeitung der Antwort gezeigt. Ohne eine präzise Formulierung führt dies zunächst zu Inkonsistenzen in der Antwort. Beispielsweise wird bei der Rückgabe am Anfang und am Ende zusätzlicher Text zurückgegeben.

Für die Erstellung des Prompts wird die von (Zhang et al., 2023) beschriebene Prompting Strategie als Grundlage genommen. Diese beschreibt drei Komponenten: den Namen der durchzuführenden Aufgabe, eine Definition sowie das gewünschte Ausgabeformat. In der Definition sind neben den zu beachtenden Regeln ebenfalls die möglichen Optionen anzugeben (Zhang et al., 2023).

Folgender Prompt wurde dem LLM übergeben:

> **Prompt:** 'Bestimme für jeden nummerierten Eintrag in der übergebenen Liste, ob der Text dieses Eintrages das Sentiment positiv, negativ oder neutral hat.', 'Antworte nur mit einem Wort (Optionen: negativ, neutral, positiv).', 'Gib das Ergebnis als JSON-Objekt zurück. In der Form: '1': 'positiv', '2': 'negativ', '3': 'neutral', ...'

Dieser folgt der beschriebenen Definition. Dabei wird verlangt, dass genau ein Wort für die Kategorie jedes Beitrags zurückgegeben wird. Zudem soll die Rückgabe in Form eines JSON-Objekts erfolgen. Dies wurde mit dem Parameter *response_format* zusätzlich festgelegt.

Die Parameter $seed = 42$ und $temperature = 0.0$ werden gesetzt. Dadurch soll die Variabilität in den Antworten des LLMs eingeschränkt werden. Die Variation der Antworten geht darauf zurück, dass die Determiniertheit nicht durch alle LLMs gegeben ist (Blackwell et al., 2024). Daraus folgt, dass ein identischer Prompt in variierenden Antworten resultieren kann (Ouyang et al., 2025).

Der Zugriff auf das Modell erfolgt über die von Chat AI (Doosthosseini et al., 2025) zur Verfügung gestellte API mit Hilfe eines entsprechenden API-Keys. Chat AI (Doosthosseini et al., 2025) wird von der (GWDG Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen, n. d. b) bereitgestellt.

Kombination der Klassifizierungsergebnisse

Zudem wird in dieser Arbeit untersucht, inwiefern sich die Kombination der Klassifizierungsergebnisse mehrerer Modelle für die Datengrundlage eignet. Das Zusammenführen der Resultate aus Methoden führt zu einer Verbesserung in der Vorhersage (Tan et al., 2023). Die Studie

von (Tan et al., 2023) veranschaulicht zudem den Einsatz des Zusammenführens im Bereich der Sentiment-Analyse. Daraus begründet sich die Untersuchung im Rahmen der Arbeit.

Die finale Klassifikation erfolgt mithilfe eines Mehrheitsvotums über die Klassifizierungsergebnisse der kombinierten Modelle. Die Nutzung eines solchen Mehrheitsentscheids orientiert sich am Vorgehen von (Suandi et al., 2024). Auch die Arbeiten von (Tan et al., 2022, 2023) dienen als methodische Referenz. Die nachfolgende Auflistung erläutert die Zusammensetzung der verschiedenen Kombinationen:

- Kombination 1: *rauh_multilang_ger_fine-tuned_V2*: Mehrheitsvotum aus dem domänenspezifischen Wörterbuch, dem mehrsprachig auf Sentiment-Analyse trainierten Modell, dem deutschsprachig auf Sentiment-Analyse trainierten Modell und dem Fine-Tuning Modell mit Lernrate 5e-5
- Kombination 2: *rauh_multilang_ger_fine-tuned_V2_fine-tuned_V3*: Mehrheitsvotum aus dem domänenspezifischen Wörterbuch, dem mehrsprachig auf Sentiment-Analyse trainierten Modell, dem deutschsprachig auf Sentiment-Analyse trainierten Modell, dem Fine-Tuning Modell mit Lernrate 5e-5 und Lernrate 3e-5
- Kombination 3: Mehrheitsvotum aller getesteten Modelle
- Kombination 4: *fine-tuned_V2_fine-tuned_V3 _fine-tuned_V4*: Mehrheitsvotum aus allen Versionen des Fine-Tuning Modells mit ausgeglichenen Trainingsdaten und den Lernraten 5e-5, 3e-5, 2e-5

5.3.3 Methoden zur Erkennung von Hassrede

Zur Untersuchung der Forschungsfrage wird zusätzlich zur Sentiment-Analyse die automatische Erkennung von Hassrede einbezogen. (A. Schmidt & Wiegand, 2017) verdeutlichen in ihrer Arbeit die enge Verbindung beider Bereiche sowie den Zusammenhang von Hassrede und negativer Stimmung. Dadurch ist der wissenschaftliche Ansatzpunkt für die unterstützende Betrachtung abgebildet.

Um für die vorliegende Datengrundlage geeignete Modelle auszuwählen, werden verschiedene Ansätze getestet. Eingesetzt wird eine Wortliste. Diese auf Schlagwörtern basierenden Methoden gelten als gut nachvollziehbar und schnell (MacAvaney et al., 2019). Für die Erkennung von Hassrede wird eine speziell dafür zusammengestellte Wortliste verwendet, die aus der Zusammenführung mehrerer durch das Forschungsprojekt BoTox („BoTox – Bot- und Kontexterkennung im Umfeld von Hasskommentaren“, 2025) bereitgestellter Listen entsteht. Vor dem Einsatz erfolgt die Bereinigung der Wortliste, um den Abgleich zu erleichtern und zu beschleunigen. Es sollten alle Beispiele mit Sonderzeichen wie beispielsweise '?' oder '\' entfernt werden. Weiterhin werden Beispiele ausgenommen, die unvollständige Klammerpaare beinhalten. Ein Beitrag wird als offensiv markiert, sobald ein Element der Wortliste in den Daten vorkommt.

Wie im Abschnitt *Stand der Forschung* 3 dargestellt, erreichen sowohl transformer-basierte Modelle (Weissenbacher & Kruschwitz, 2024) als auch LLMs (Albladi et al., 2025) in der Erkennung von Hassrede gute Ergebnisse.

Auf dieser Grundlage werden neben dem wortlistenbasierten Verfahren auch transformerbasierte Klassifikationsmodelle sowie ein LLM in die Untersuchung einbezogen. Ausgangspunkt für die Auswahl sind auf der *Hugging Face* Webseite verfügbare Modelle, die auf die Erkennung von Hassrede trainiert sind. Ein wesentliches Kriterium dabei ist, dass die Modelle entweder auf deutschsprachigen oder auf mehrsprachigen Datensätzen trainiert sind. Eine Übersicht über die gesamten getesteten Modelle kann der nachfolgenden Auflistung entnommen werden. Modell (1) stellt ein Modell dar, das das German BERT v1 Modell unter Verwendung der GermanEval18Coarse Daten als Trainingsdaten weitertrainiert (deepset, n. d.). Die Modelle (2)-(4) sind drei Varianten des *xlm-roberta-large* (Facebook AI community, n. d.) Modells. Dieses Modell nutzt als Trainingsdatensatz einen mit CommonCrawl (Wenzek et al., 2020) erstellten Datensatz, bestehend aus Daten für 100 Sprachen (Conneau et al., 2020). Die Trainingsdaten für das Fine-Tuning sind bei keinem der drei Versionen angegeben (Christodoulou, n. d. a, n. d. b, n. d. c). Bei dem Modell (5) handelt es sich um ein deutschsprachig trainiertes Modell. Grundlage ist dabei das *bert-base-german-cased* (BERT community, n. d.). Dieses wird mit den Daten aus HASOC (Mandl et al., 2023) weitertrainiert (Ortiz, n. d.). Im letzten Modell (6) werden zwei deutschsprachige Datensätze mit den Themenschwerpunkten der Flüchtlingskrise sowie Geflüchteten im Trainingsprozess eingesetzt (Aluru et al., 2020). Die Inferenz dieser Modelle erfolgt analog zum Vorgehen für die Methoden der Sentiment-Analyse. Für die Untersuchung eines LLMs wird das bereits für die Sentiment-Analyse eingesetzte LLM angewendet.

Auflistung der ausgewählten Modelle:

(1) *'deepset/bert-base-german-cased-hatespeech-GermEval18Coarse'* (deepset, n. d.)

(2) *'christinacdl/XLM_RoBERTa-Offensive-Language-Detection-8-langs-new'* (Christodoulou, n. d. c)

(3) *'christinacdl/XLM_RoBERTa-Multilingual-OpusMT-Offensive-Language-Detection'* (Christodoulou, n. d. b)

(4) *'christinacdl/XLM_RoBERTa-Multilingual-Hate-Speech-Detection-New'* (Christodoulou, n. d. a)

(5) *'jorgeortizv/BERT-hateSpeechRecognition-German'* (Ortiz, n. d.)

(6) *'Hate-speech-CNERG/dehatebert-mono-german'* (Hate-ALERT, n. d.)

(7) *'Llama 3.1 SauerkrautLM 70B Instruct'* (VAGO solutions, n. d.)

(8) *'Wortliste'* aus dem Forschungsprojekt BoTox („BoTox – Bot- und Kontexterkennung im Umfeld von Hasskommentaren“, 2025)

Dem LLM wird folgender Prompt übergeben:

> **Prompt:** 'Bestimme für jeden nummerierten Eintrag in der übergebenen Liste, ob der Text dieses Eintrages als Hassrede, Beleidigung oder Diskriminierung eingestuft werden kann.', 'Antworte nur mit einem Wort (Optionen: offensive, not_offensive).', 'Gib das Ergebnis als JSON-Objekt zurück. In der Form: '1': 'offensive', '2': 'not_offensive', ...'.

Zu erkennen ist, dass der Prompt für die Sentiment-Analyse an die Erkennung von Hassrede angepasst ist. Demnach folgt dieser Prompt ebenfalls der von (Zhang et al., 2023) dargelegten Strategie.

Auf der Basis dieser Klassifizierungen wird die Evaluation der getesteten Modelle auf den gesamten Evaluationsdaten durchgeführt. Dadurch soll das performanteste Modell ausgewählt werden. Dieses wird anschließend auf den Analysedaten ausgeführt.

5.4 Evaluation der Methoden

In diesem Abschnitt werden die ausgewählten Methoden evaluiert. Ziel ist es, das Modell mit dem besten Resultat auf der vorliegenden Datengrundlage zu identifizieren. Die Evaluation der verschiedenen Methoden erfolgt anhand ausgewählter Metriken. Berücksichtigt werden Precision, Accuracy und F1-Wert. Diese werden überwiegend in der Sentiment-Analyse betrachtet (Wankhade et al., 2022) und sind ebenfalls in der Erkennung von Hassrede (Rini et al., 2020) etabliert. Die Einbeziehung aller drei F1-Varianten basiert auf den Empfehlungen von (Hinojosa Lee et al., 2024), die den Einsatz differenzierter F1-Metriken insbesondere bei Klassifikationsaufgaben mit mehreren Klassen empfehlen.

Die Ergebnisse für die Methoden der Sentiment-Analyse sind der Tabelle 5.1 zu entnehmen. Aufgebaut sind die nachfolgenden Evaluationstabellen 5.1, 5.2, indem in der ersten Spalte die untersuchten Methoden und darauffolgend die Ergebnisse der Evaluationsmetriken aufgelistet sind. Die Bedeutung der Methodenbezeichnungen ist in der folgenden Auflistung erklärt:

- *sentiMerge*: sentiMerge-Wörterbuch (Emerson & Declerck, 2014)
- *rauh_with_pos*: domänenspezifisches Wörterbuch *Rauh* (Rauh, 2018a, 2018b) mit PoS-Tag-Abgleich
- *task_multi*: auf Sentiment-Analyse trainiertes mehrsprachiges Modell
- *task_multi_politics*: auf Sentiment-Analyse trainiertes mehrsprachiges Modell erneut trainiert mit politischen Daten
- *task_german*: deutschsprachiges auf Sentiment-Analyse trainiertes Modell
- *llm*: LLM-Modell
- *fine-tuned_V1*: Fine-Tuning Modell mit ungleich verteilten Trainingsdaten und Lernrate 5e-5
- *fine-tuned_V2*: Fine-Tuning Modell mit ausgeglichenen Trainingsdaten und Lernrate 5e-5
- *fine-tuned_V3*: Fine-Tuning Modell mit ausgeglichenen Trainingsdaten und Lernrate 3e-5
- *fine-tuned_V4*: Fine-Tuning Modell mit ausgeglichenen Trainingsdaten und Lernrate 2e-5

Die besten Ergebnisse werden in den Evaluationstabellen hervorgehoben.

Methode	**Accuracy**	**micro F1**	**macro F1**	**gewichteter F1**	**Precision**
sentiMerge	0.414	0.414	0.372	0.377	0.439
rauh_with_pos	0.471	0.471	0.436	0.439	**0.623**
task_multi	0.485	0.485	0.449	0.477	0.488
task_multi_politics	0.506	0.506	0.462	0.465	0.5402
task_german	0.4498	0.4498	0.385	0.424	0.452
llm	0.446	0.446	0.4499	0.459	0.556
fine-tuned_V1	0.425	0.425	0.199	0.253	0.1803
fine-tuned_V2	0.567	0.567	0.423	0.503	0.462
fine-tuned_V3	**0.644**	**0.644**	**0.482**	**0.576**	0.534
fine-tuned_V4	0.623	0.623	0.468	0.559	0.533

Tabelle 5.1: Evaluationsergebnisse für die Klassifikation in drei Klassen

Die Analyse der Ergebnisse der wörterbuchbasierten Methoden zeigt, dass das domänenspezifische Wörterbuch in Kombination mit dem Abgleich von PoS-Tags die beste Leistung erzielt.

Weiterhin werden die drei auf Sentiment-Analyse trainierten Modelle verglichen. Das ausschließlich deutschsprachig trainierte Modell erzielt dabei die schwächsten Ergebnisse. Der Vergleich der anderen beiden ergibt kein eindeutiges Ergebnis. Während das nochmals mit politischen Daten trainierte Modell im micro-F1 Wert leicht besser abschneidet, erzielt das mehrsprachige Modell einen höheren gewichteten F1-Wert.

Im Vergleich der Wörterbücher mit den auf Sentiment-Analyse trainierten Modellen zeigen sowohl das mehrsprachige als auch das Fine-Tuning Modell die besseren Ergebnisse. Das Modell für die deutsche Sprache liegt unter denen des domänenspezifischen Wörterbuchs.

Für die Einschätzung des verwendeten LLMs ist zwischen dem micro und gewichteten F1-Wert zu unterscheiden. Im Vergleich der micro Werte zeigt das LLM keine guten Ergebnisse. Im gewichteten F1-Wert sind deutlich bessere Ergebnisse erkennbar.

Im Gesamtvergleich aller Modelle anhand der Varianten des F1-Wertes erzielt das Fine-Tuning Modell mit ausgeglichener Verteilung der Klassen und einer Lernrate von 3e-5 die besten Ergebnisse. Die Betrachtung der Precision ergibt ein abweichendes Bild. In dieser Metrik erreicht das domänenspezifische Wörterbuch den besten Wert.

Die Konfusionsmatrizen 5.4 machen deutlich, dass alle Versionen des Fine-Tuning Modells keine positive Klasse vorhersagen können. Dieses Ergebnis ist auf den Trainingsprozess zurückzuführen. Das Modell wurde ausschließlich mit Beispielen der Klassen 'negativ' und 'neutral' trainiert. Der Grund sind fehlende Beispiele für eine positive Klasse in den Trainingsdaten 4.2.

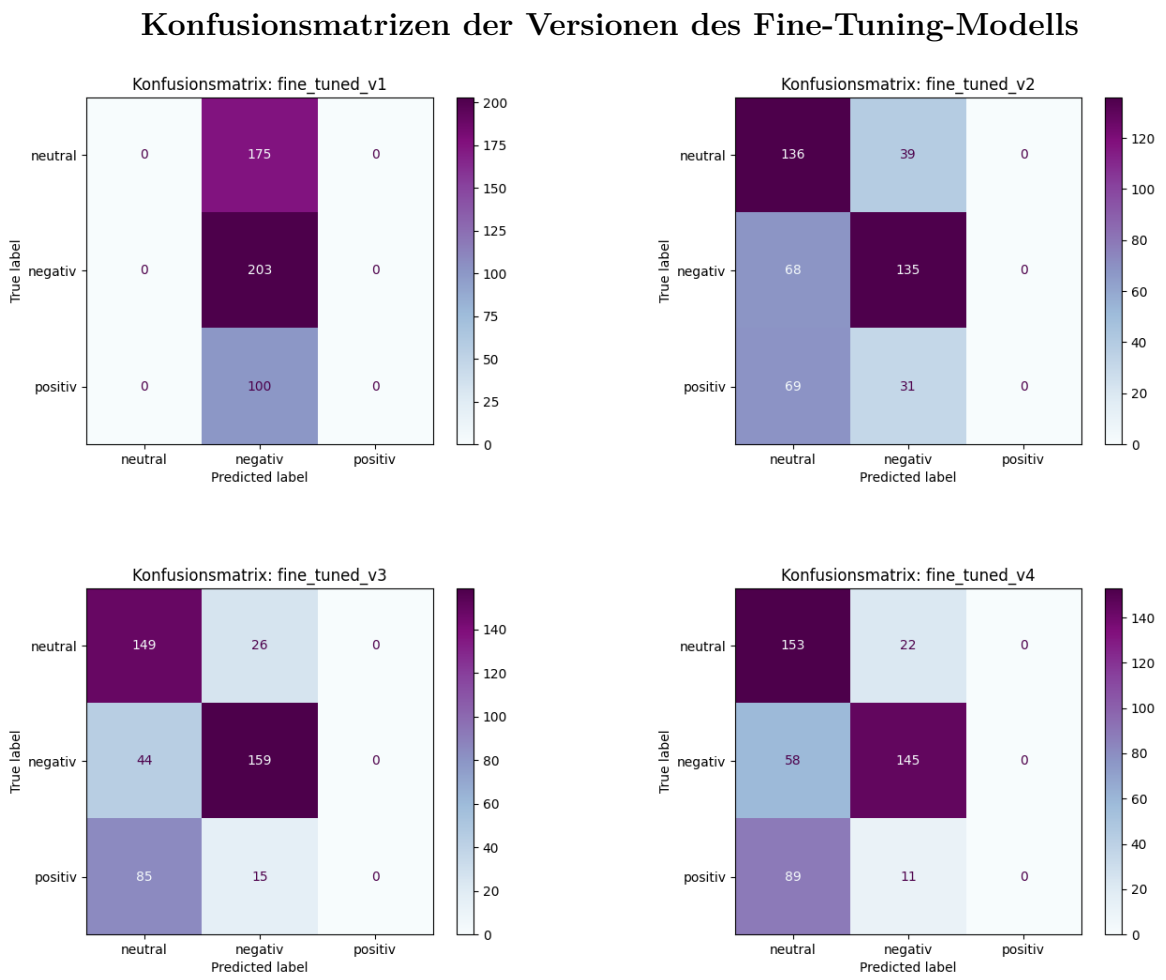


Abbildung 5.4: Übersicht über die Konfusionsmatrizen der Versionen des Fine-Tuning-Modells für die Klassifikation in drei Klassen

Das Fine-Tuning Modell mit ausgeglichenen Klassen hat bereits bei der Evaluation mit drei Klassen in den F1-Metriken die besten Ergebnisse. Dadurch zeigt sich dessen gute Performanz, die Klassen 'negativ' und 'neutral' zu erkennen. Auf Basis dieser Erkenntnis sind die Repräsentanten der positiven Klasse in den Evaluationsdaten eingehender zu betrachten. Im Fokus stand dabei die Qualität der positiven Beispiele. Es konnte festgestellt werden, dass 39 % der positiven Klasse aus Einträgen besteht, die aufgrund von Beifall, Heiterkeit, Lachen oder Bitten dort eingeordnet werden. Anschließend werden die verbliebenen Einträge manuell gesichtet. Dabei kann ebenfalls festgestellt werden, dass aussagekräftige und eindeutige positive Beispiele selten vorkommen. Aus diesen Erkenntnissen begründet sich die weiterführende Betrachtung der Sentiment-Analyse als binäre Klassifikation.

Daraus folgend wird die Evaluation erneut durchgeführt. In diesem Fall erfolgt die Vereinigung der beiden Klassen 'positiv' und 'neutral' zur 'neutralen' Klasse. Die Ergebnisse sind der Tabelle 5.2 zu entnehmen.

Methode	Accuracy	micro F1	macro F1	gewichteter F1	Precision
sentiMerge	0.636	0.636	0.626	0.635	0.635
rauh_with_pos	0.651	0.651	0.547	0.579	0.738
task_multi	0.626	0.626	0.611	0.622	0.621
task_multi_politics	0.573	0.573	0.5699	0.564	0.626
task_german	0.579	0.579	0.564	0.576	0.575
llm	0.709	0.709	0.662	0.681	0.744
fine-tuned_V1	0.425	0.425	0.298	0.253	0.1804
fine-tuned_V2	0.711	0.711	0.705	0.711	0.712
fine-tuned_V3	**0.822**	**0.822**	**0.818**	**0.822**	**0.822**
fine-tuned_V4	0.8096	0.8096	0.801	0.808	0.81

Tabelle 5.2: Evaluationsergebnisse für die binäre Klassifikation

Der Vergleich der Evaluationsergebnisse lässt Unterschiede in den Ergebnissen der Modelle erkennen. Es zeigt sich, dass die Ergebnisse der Evaluationsmetriken für alle Modelle besser geworden sind. Eine besondere Verbesserung kann für das LLM festgestellt werden. Zudem wird aus den Ergebnissen deutlich, dass die Accuracy der beiden Wörterbücher über den Werten für die auf Sentiment-Analyse trainierten Modelle liegt. Ein Vergleich mit dem gewichteten F1-Wert zeigt, dass für diesen Wert nur das *SentiMerge* (Emerson & Declerck, 2014) Wörterbuch einen höheren Wert als die auf Sentiment-Analyse trainierten Modellen erreicht. Weiterhin wird zwischen den auf Sentiment-Analyse trainierten Modellen ersichtlich, dass das deutschsprachig trainierte Modell eine deutliche Verbesserung aufweist. Für die Metriken Accuracy und gewichteter F1-Wert übertrifft es das mehrsprachig-politisch trainierte Modell.

Eine sehr deutliche Verbesserung in den Ergebnissen zeigt sich bei den Versionen des selbst erstellten Fine-Tuning Modells mit ausgeglichenen Trainingsdaten. Erkennbar ist erneut, dass das Modell mit einer Lernrate von 3e-5 das beste Ergebnis erzielt. Es erreicht einen gewichteten F1-Wert, eine Accuracy und eine Precision von 0.822.

Die Evaluation der kombinierten Klassifizierungsergebnisse wird in Tabelle 5.3 dargestellt. Dabei ist in der ersten Zeile das beste Ergebnis der einzelnen Modelle angegeben. Die Bezeichnung der Methoden ist entsprechend der Auflistung der Kombinationen zu entnehmen.

Methode	Accuracy	micro F1	macro F1	gewichteter F1	Precision
fine-tuned_V3	**0.822**	**0.822**	**0.818**	**0.822**	**0.822**
Kombination 1	0.688	0.688	0.679	0.687	0.686
Kombination 2	0.764	0.764	0.742	0.754	0.776
Kombination 3	**0.822**	**0.822**	0.816	0.821	**0.822**
Kombination 4	0.82	0.82	0.815	0.819	0.819

Tabelle 5.3: Evaluationsergebnisse für die Kombination ausgewählter Methoden (binäre Klassifikation)

Der Vergleich der Evaluationsmetriken zwischen den einzelnen Modellen und der Kombinationen zeigt, dass die dritte Kombination in Bezug auf Accuracy, micro F1-Wert und Precision identische Werte erzielt. Die Varianten macro und gewichtet der F1-Metrik sind dagegen für das einzelne Modell minimal besser. Aufgrund der Größe der Analysedaten wird es als nicht praktikabel eingeschätzt, Kombination 3 (Mehrheitsvotum über alle Klassifizierungsergebnisse) anzuwenden. Daher soll dieses Ergebnis ausschließlich im Theoretischen zeigen, dass eine Kombination gute Ergebnisse in diesem Anwendungsfall erreichen kann.

Aus den gesamten Evaluationsergebnissen zeigt sich, dass das Fine-Tuning-Modell mit einer Lernrate von 3e-5 in den Metriken für die binäre Klassifikation die beste Performanz aufweist. Daher wird dieses für die Klassifizierung der Analysedaten verwendet.

Eine weitere Evaluation wird durchgeführt, um ein geeignetes Modell für die Erkennung von Hassrede zu identifizieren. Die Ergebnisse werden in Tabelle 5.4 dargestellt. Die Modelle sind entsprechend ihrer Nummerierung aus der vorherig aufgeführten Auflistung zu entnehmen.

Methode	**micro F1**	**macro F1**	**gewichteter F1**	**Precision**
(1)	0.87	0.507	0.851	0.125
(2)	0.836	0.588	0.848	0.231
(3)	0.866	0.587	0.863	0.262
(4)	**0.91**	0.517	0.871	**1.0**
(5)	0.902	0.494	0.863	0.25
(6)	0.888	0.4703	0.852	0.0
(7)	0.878	**0.696**	**0.887**	0.394
(8)	0.654	0.527	0.726	0.175

Tabelle 5.4: Evaluationsergebnisse der ausgewählten Methoden für die Erkennung von Hassrede

Es zeigt sich, dass das Modell (4) *christinacdl/XLM_RoBERTa-Multilingual-Hate-Speech-Detection-New* für den micro F1-Wert das beste Resultat erzielt. Für die anderen Varianten der F1-Metrik erreicht das verwendete LLM die besten Ergebnisse. Das Entscheidungskriterium ist in diesem Fall der gewichtete F1-Wert. Diese Variante der F1-Metrik geht auf Änderungen in der Verteilung der Klassen ein (Hinojosa Lee et al., 2024). Da die Einträge, die als Hassrede annotiert sind, im Evaluationsdatensatz geringer sind, wird diese Metrik betrachtet. Eine Anwendung des LLMs direkt auf den Analysedaten kann aufgrund der Größe der Daten und der daraus entstehenden Antwortdauer nicht durchgeführt werden. Daher werden die Daten vorab gefiltert, bevor sie dem LLM zur Verarbeitung übergeben werden. Der Fokus bei der Entscheidung für das Modell zur Datenfilterung lag auf der Precision. Gemäß der Definition 2.5 dieser Metrik gibt sie an, wie zuverlässig als Hassrede klassifizierte Beiträge tatsächlich Hassrede darstellen. Die Betrachtung dieser Metrik zeigt für das Modell (4) mit einem Wert von 1.0 die beste Performanz. Basierend auf diesen Ergebnissen wird eine Kombination der beiden Methoden verwendet. Eine manuelle Überprüfung der klassifizierten Beiträge bestätigt, dass diese Vorgehensweise funktioniert.

5.5 Durchführung der Datenanalyse

Das Modell, das auf Basis des Evaluationsdatensatzes die beste Performanz erzielt, wird verwendet, um die Klassifizierungsergebnisse für die Sentiment-Analyse und die Erkennung von Hassrede über den gesamten Untersuchungszeitraum (2002–2021) zu erstellen. Diese Ergebnisse dienen als Grundlage für die anschließende Datenanalyse, in der die aufgestellten sozialwissenschaftlichen Fragestellungen untersucht werden.

Für die weitere Analyse werden die Daten auf Jahres- und auf Quartalsebene betrachtet. Die klassifizierten Daten werden zunächst in numerische Werte umgewandelt. Für jedes Datum wird die Anzahl der Einträge pro Klasse ermittelt und anschließend durch die Gesamtanzahl der Einträge an diesem Datum geteilt. Die finalen Werte ergeben sich aus der Differenz der relativen Häufigkeiten der neutralen und der negativen Klasse. Diese Berechnung folgt der Gleichung 2.2.1 aus (Rauh, 2018b) zur Berechnung von Sentiment-Scores. Ein niedriger Wert zeigt dabei eine steigende Negativität an. Die aggregierten Werte je Jahr und Quartal werden über den Mittelwert gebildet.

Die Betrachtung der jährlichen Trendkurve liefert einen allgemeinen Überblick über den Verlauf der Werte. Die Analyse auf Quartalsebene ermöglicht hingegen detailliertere Einblicke in den Verlauf der Trendkurve und erlaubt eine präzisere Identifikation von Anstiegs- und Abstiegsphasen. Durch die Untersuchung beider Ebenen kann zudem ein Vergleich mit den Ergebnissen von (Rheault et al., 2016) sowie (Lehtosalo & Nerbonne, 2020) hinsichtlich der Polarität in Parlamenten vorgenommen werden. Für die Analyse des Teildatensatzes zur Untersuchung der Einstellung gegenüber dem ehemaligen Osten Deutschlands erfolgt die Visualisierung ausschließlich auf der jährlichen Ebene. Aufgrund dessen, dass dieser nur Einträge für den Oktober jedes Jahres enthält.

Der Wahltag entnommen aus (Bundeswahlleiterin, 2025a, 2025b) wird in den Visualisierungen der Ergebnisse durch eine vertikale Linie gekennzeichnet. Zusätzliche Markierungen stellen die untersuchten Ereignisse im Zusammenhang mit der Fragestellung (b): *Lassen sich Wendepunkte in der Grundstimmung identifizieren und fallen diese mit markanten Ereignissen zusammen?* dar. Ausgewählt werden die folgenden Ereignisse: Flutkatastrophe 2002, Finanzkrise 2008, Beginn der Flüchtlingsdebatte und der Pandemie sowie der Einzug der AfD in die jeweiligen Gremien. Für die Flutkatastrophe wird der 12. August 2002 als Markierung ausgewählt. An diesem Tag wird im Zinnwald-Georgenfeld ein Niederschlag von 312 Liter auf den Quadratmeter gemessen (Weber, 2022). Die Markierung der Finanzkrise ist der 5. Oktober 2008. An diesem Tag findet die Pressekonferenz von Angela Merkel und Peer Steinbrück zum Schutz der Sparguthaben statt (Dohmen, 2024; tagesschau.de, 2008). Als Datum für den Beginn der Flüchtlingsdebatte ist der 21. August 2015 hervorgehoben. Dieser repräsentiert den Tag der Anweisung des Bundesamtes für Migration und Flüchtlinge, syrische Flüchtlinge nicht zurückzuweisen (Herbert & Schönhagen, 2020). Der Beginn der Pandemie wird durch den 27. Januar 2020 markiert. An diesem Tag wurde der erste Deutsche infizierte bekannt gegeben (Bundesministerium für Gesundheit, 2023). Der Einzug der AfD wird durch die Wahlen markiert, nach denen die Partei im jeweiligen Gremium vertreten ist (Decker, 2022).

Einen besseren Einblick in die Daten soll mit Hilfe von Zeitreihenzerlegung erreicht werden. Dafür wird die STL - Methode (Cleveland et al., 1990) verwendet. Ausgewählt wird diese Methode, da sie mit einer veränderlichen saisonalen Komponente umgehen kann. Zudem lässt

sich über Parameter der Funktion die Stärke der Glättung für die trend-zyklische Komponente einstellen (Hyndman & Athanasopoulos, 2014). Nach (Dudek, 2023) zählt sie weiterhin in diesem Bereich zu den verbreitetsten Verfahren.

Darüber hinaus soll die Möglichkeit bestehen, Ausreißer visuell darzustellen. Die Ermittlung der Ausreißer erfolgt basierend auf den Residuen der zuvor berechneten Zeitreihenzerlegung. Angewendet wird dabei die STL-Methode und anschließend der Quartilsabstand. Dieses Vorgehen folgt dem von (Giachanou & Crestani, 2016) eingesetzten Verfahren zur Erkennung von Ausreißern und einer Variante die von (Eslava, 2023) implementiert ist. Repräsentiert werden die Ausreißer in den Grafiken durch ein 'x'-Symbol.

6 Ergebnisse

Das folgende Kapitel ist in zwei Schwerpunkte unterteilt. Der erste Teil bildet die Auseinandersetzung mit den Ergebnissen der Klassifizierung aus der Sentiment-Analyse und der Erkennung von Hassrede. Die Ergebnisse werden im Kontext der aufgestellten sozialwissenschaftlichen Fragestellungen interpretiert. Dadurch wird der Bezug zur Forschungsfrage hergestellt. Im zweiten Teil dieses Kapitels erfolgt eine Reflexion der durchgeführten Untersuchungsmethode sowie die Einordnung der Erkenntnisse aus der praktischen Arbeit.

6.1 Ergebnisse der Datenanalyse

Nachfolgende Ausführungen stellen die Analyse der berechneten Werte aus den Methoden der Sentiment-Analyse und der Erkennung von Hassrede dar. Die Beschreibung der wichtigsten Erkenntnisse aus den Abbildungen dient anschließend der Bearbeitung der sozialwissenschaftlichen Fragestellungen. Die Darstellung der Resultate ist in vier Abschnitte gegliedert, die jeweils eine der aufgestellten Fragestellungen thematisieren.

In der nachfolgenden Auflistung sind die zu Beginn aufgestellten sozialwissenschaftlichen Fragestellungen erneut aufgeführt. Diese sollen mit Hilfe der Klassifizierungsergebnisse untersucht werden.

(a) *Sind die Debatten in den deutschen Gremien aggressiver und negativer geworden?*

(b) *Lassen sich Wendepunkte in der Grundstimmung identifizieren und fallen diese mit markanten Ereignissen zusammen?*

(c) *Lässt sich eine Veränderung des Stimmungsbildes vor und nach Wahlen feststellen?*

(d) *Lässt sich ein negativer Bias gegenüber den Regionen des ehemaligen Ostens Deutschlands feststellen?*

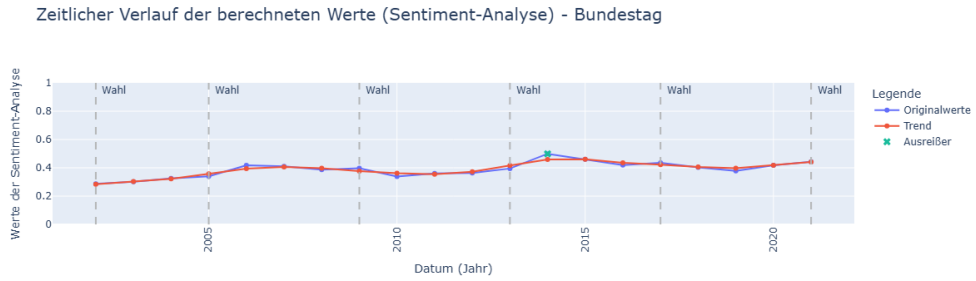
Unterstützend zu den berechneten Sentiment-Scores werden auf der Jahres- und Quartalsebene die Werte für die Erkennung von Hassrede betrachtet. Durch die Analyse des allgemeinen Verlaufs der Trendkurven ist ein erster Überblick über die Entwicklung der Stimmung in den Gremien gegeben. Das wird eingesetzt, um Bezug auf die sozialwissenschaftlichen Fragestellungen (a) und (d) zu nehmen. Eine differenziertere und detailliertere Untersuchung erfolgt basierend auf den berechneten Werten der Quartalsebene. Anhand der Trendkurvenverläufe auf dieser Ebene sowie der markierten Ausreißer wird die zweite sozialwissenschaftliche Fragestellung (b) untersucht. Die Markierungen der Wahltermine ermöglichen die Untersuchung des Stimmungsbildes im Zeitraum um die Wahlen (Fragestellung (c)).

Bei der Interpretation der Grafiken ist zu beachten, dass ein positiver Anstieg eine Entwicklung zu größeren Werten bedeutet. Größere Werte beschreiben in diesem Zusammenhang, dass der

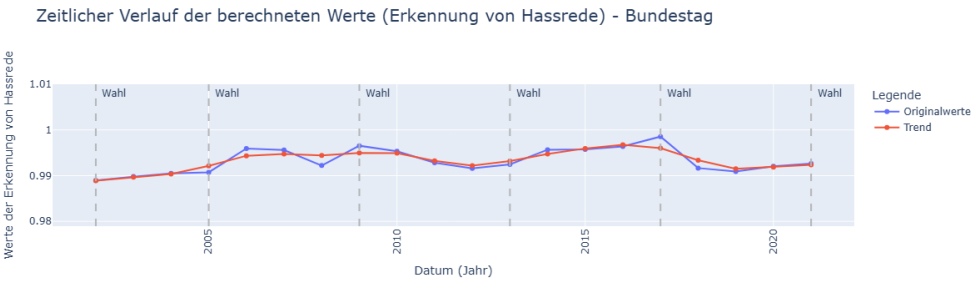
Anteil an neutral klassifizierten Beiträgen steigt und die als negativ klassifizierten Beiträge sinken. Eine positive Entwicklung lässt dementsprechend eine neutralere Stimmung vermuten. Analog sind die Werte in den Abbildungen zur Erkennung von Hassrede zu interpretieren.

Anzumerken ist dabei, dass die berechneten Werte nahe dem Wert 1 sind. Für die bessere Darstellung der An- und Abstiegsphasen im Trendkurvenverlauf ist die Skalierung der y-Achse angepasst. Ein Wert nahe 1 bedeutet in diesem Kontext, dass der überwiegende Teil der Redebeiträge als keine Hassrede klassifiziert wird. Kleinere Werte sind demnach als ein Anstieg in der Verwendung von beleidigender Sprache zu interpretieren. Die folgenden Abbildungen der Ergebnisse sind zweigeteilt. Im oberen Teil werden die Werte der Sentiment-Analyse und im unteren Teil die der Erkennung von Hassrede dargestellt.

Die erste sozialwissenschaftliche Fragestellung (a): *Sind die Debatten in den deutschen Gremien aggressiver und negativer geworden* thematisiert die Entwicklung von Negativität und Aggressivität innerhalb deutscher Gremien. Für die Untersuchung dieser Fragestellung werden die jährlichen Trendkurvenverläufe der ausgewählten Gremien analysiert. Zur Unterstützung der dabei gewonnenen Erkenntnisse werden die Trendkurvenverläufe je Quartal betrachtet.



(a) Darstellung des zeitlichen Verlaufs der Sentiment-Scores im Deutschen Bundestag auf der jährlichen Betrachtungsebene

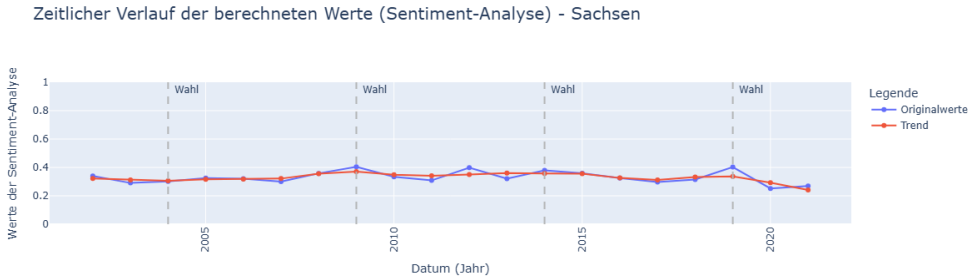


(b) Darstellung des zeitlichen Verlaufs der Werte für die Erkennung von Hassrede im Deutschen Bundestag auf der jährlichen Betrachtungsebene

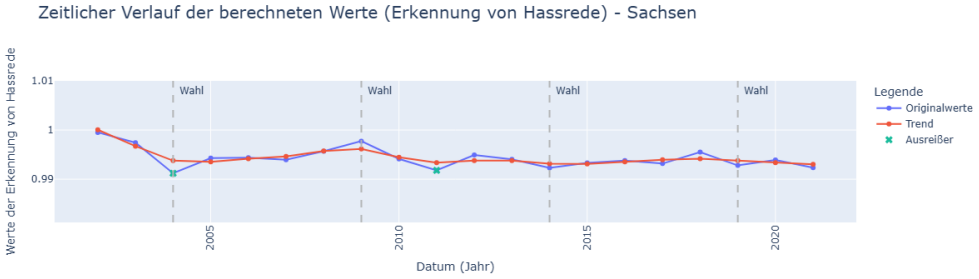
Abbildung 6.1: Visualisierung der Ergebnisse der Sentiment-Analyse und Erkennung von Hassrede für den Deutschen Bundestag

In der Abbildung 6.1 sind die Trendkurven für die jährlich berechneten Sentiment-Scores sowie die Werte der Erkennung von Hassrede für den Deutschen Bundestag dargestellt. Der Beginn beider Kurven markiert die jeweils niedrigsten Werte des Beobachtungsintervalls von 2002 bis 2021. Es ist in beiden Grafiken zunächst eine positive Entwicklung erkennbar. Der Verlauf in den Sentiment-Scores zeigt eine leicht erkennbare Steigung. Die Trendkurve

der Untersuchung auf Hassrede zeigt dagegen ein uneinheitliches Bild. Hierbei ist bis 2016 ein allgemeiner Rückgang der Negativität erkennbar, gefolgt von einem Anstieg zwischen 2016 und 2018. Ab 2018 deutet sich wieder ein leichter Trend zu neutraleren Stimmung an. Die beschriebenen Erkenntnisse lassen sich im Allgemeinen ebenfalls durch die je Quartal berechneten Werte in der Abbildung 6.6 erkennen.



(a) Darstellung des zeitlichen Verlaufs der Sentiment-Scores im Sächsischen Landtag auf der jährlichen Betrachtungsebene

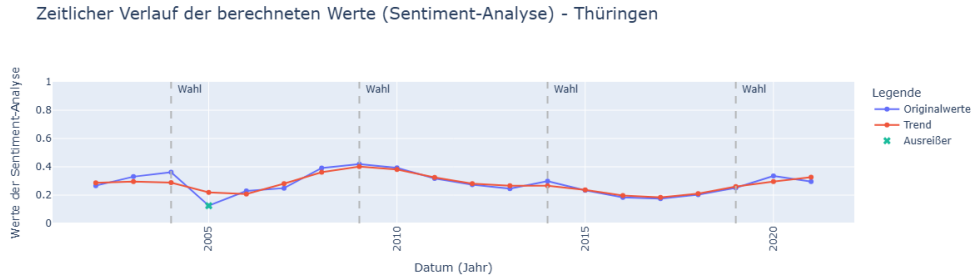


(b) Darstellung des zeitlichen Verlaufs der Werte für die Erkennung von Hassrede im Sächsischen Landtag auf der jährlichen Betrachtungsebene

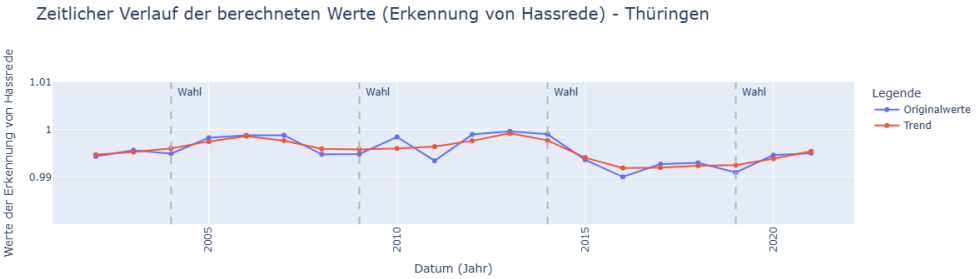
Abbildung 6.2: Visualisierung der Ergebnisse der Sentiment-Analyse und Erkennung von Hassrede für den Sächsischen Landtag

Die sächsische Trendkurve 6.2a der jährlichen Sentiment-Scores bleibt über das gesamte Betrachtungsintervall (2002-2022) auf einem annähernd gleichen Niveau. Bis ins Jahr 2008 ist ein leichter Rückgang an Negativität erkennbar. Ab dem Jahr 2018 deutet sich dagegen eine Entwicklung zu steigender Negativität an. Diese Entwicklung bestätigt sich, indem zusätzlich die Abbildung 6.2b analysiert wird. In dieser sind die Ergebnisse für Sachsen aus der Untersuchung auf Hassrede dargestellt. Für weitere Erkenntnisse werden die visualisierten Werte je Quartal aus der Abbildung 6.7 genutzt. In dieser ist die festgestellte Entwicklung zu steigender Negativität am Ende des Untersuchungszeitraums (2002-2021) deutlicher zu erkennen.

Die Abbildung 6.3a zeigt den jährlichen Verlauf der Trendkurve der Sentiment-Scores für das Bundesland Thüringen. Ab 2006 ist zunächst ein Trend zu neutralerer Stimmung sichtbar. Im Jahr 2009 deutet sich allgemein eine Steigung zu negativerer Stimmung an, die bis 2017 anhält. Die Trendkurve zeigt im Folgejahr einen Rückgang der Negativität. Diese Entwicklung bleibt anschließend bestehen. Die Trendkurve 6.3b der Untersuchung auf beleidigende Sprache zeigt ab 2013 einen deutlichen Anstieg an Negativität. Durch den im Jahr 2018 leicht positiven



(a) Darstellung des zeitlichen Verlaufs der Sentiment-Scores im Thüringischen Landtag auf der jährlichen Betrachtungsebene

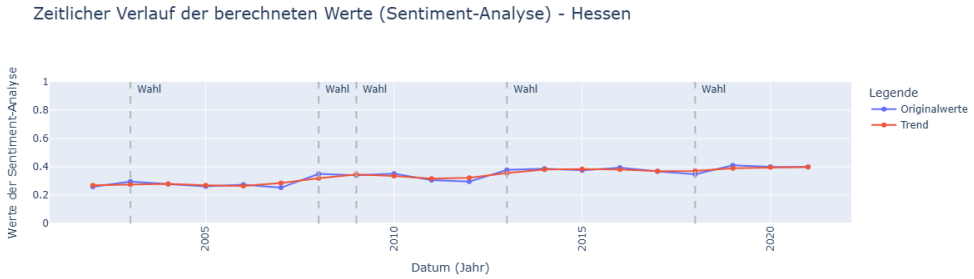


(b) Darstellung des zeitlichen Verlaufs der Scores für die Erkennung von Hassrede im Thüringischen Landtag auf der jährlichen Betrachtungsebene

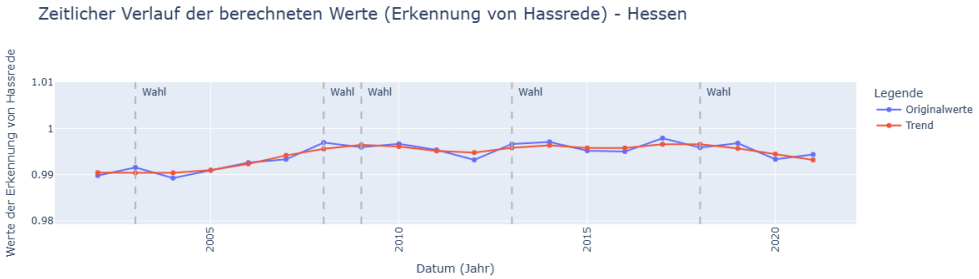
Abbildung 6.3: Visualisierung der Ergebnisse der Sentiment-Analyse und Erkennung von Hassrede für den Thüringischen Landtag

Verlauf der Trendkurve deutet sich eine Abschwächung dessen an. Die Betrachtung des Trendkurvenverlaufs je Quartal 6.8 bestätigt diese Erkenntnisse.

Die Ergebnisse der Sentiment-Analyse sowie der Erkennung von Hassrede im Hessischen Landtag sind in Abbildung 6.4 dargestellt. Beide jährlichen Trendkurvenverläufe zeigen einen allgemeinen Rückgang an Negativität. Die Endphasen in beiden Abbildungen werden dagegen durch einen leichten Anstieg an Negativität charakterisiert. In beiden Grafiken lässt sich der Beginn dieser Entwicklung auf 2019 eingrenzen. Für den Hessischen Landtag zeigt die Abbildung 6.9 der je Quartal berechneten Werte Unterschiede im Verlauf der Trendkurven auf. Die Auffälligkeiten sind in der Endphase des betrachteten Zeitraums (2002-2021) sichtbar. Auf der jährlichen Betrachtungsebene ist in beiden Darstellungen eine einsetzende Entwicklung zu mehr Negativität sichtbar. Aus den je Quartal betrachteten Trendkurvenverläufen ist dagegen eine Entwicklung zu sinkender Negativität abzulesen.



(a) Darstellung des zeitlichen Verlaufs der Sentiment-Scores im hessischen Landtag auf der jährlichen Betrachtungsebene

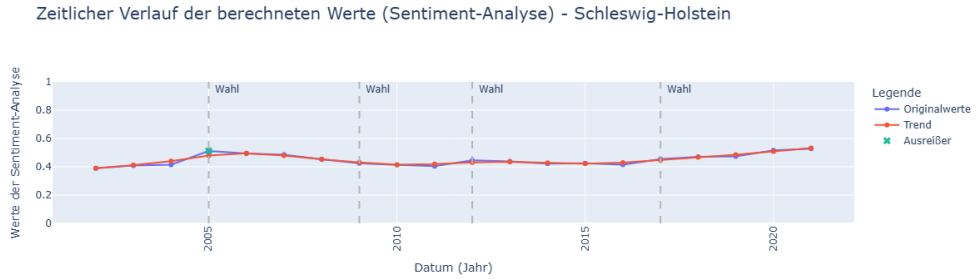


(b) Darstellung des zeitlichen Verlaufs der Scores für die Erkennung von Hassrede im Hessischen Landtag auf der jährlichen Betrachtungsebene

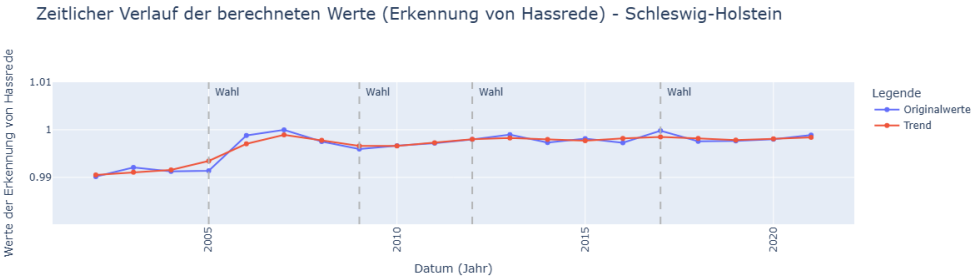
Abbildung 6.4: Visualisierung der Ergebnisse der Sentiment-Analyse und Erkennung von Hassrede für den Hessischen Landtag

Die jährlich aggregierten Ergebnisse für das Bundesland Schleswig-Holstein sind in der Abbildung 6.5 dargestellt. Beide Trendkurven zeigen über den Betrachtungszeitraum hinweg einen Rückgang der Negativität. In der Darstellung 6.5b der berechneten Werte für die Erkennung von Hassrede lässt sich ein deutlicher Anstieg der neutralen Beiträge zwischen 2004 und 2007 erkennen. Diese Erkenntnisse bestätigen sich in der Betrachtung der Quartalsebene. 6.10.

Basierend auf den Erkenntnissen der Trendkurvenverläufe wird folgend thematisiert, inwieweit die Stimmungen in den untersuchten Gremien negativer und aggressiver geworden sind (Fragestellung (a)). Die Untersuchung des Sächsischen Landtages zeigt, dass alle analysierten Trendkurvenverläufe eine Entwicklung zu steigender Negativität vermuten lassen. Die Untersuchung für den Landtag in Schleswig-Holstein deutet in allen zugrundeliegenden Trendkurvenverläufen deuten eine Abnahme an Negativität an. Die Landtage in Thüringen und Hessen sowie der Bundestag lassen keine eindeutigen Erkenntnisse zu. Für die thüringischen Trendkurvenverläufe zeigen alle eine Zunahme der neutralen Stimmung gegen Ende des untersuchten Zeitraums (2002-2021) an. Der zuvor festgestellte Verlauf stellt dagegen eine Entwicklung zu steigender Negativität dar. Es lässt sich vermuten, dass ein leichter Rückgang an Negativität und Aggressivität eingesetzt hat. Am deutlichsten wird diese Annahme in der Abbildung 6.8b der auf Quartalsebene berechneten Werte aus der Erkennung von Hassrede. Diese Ergebnisse zeigen sich ebenfalls für den Bundestag. Unter Berücksichtigung der Untersuchung auf beleidigende Sprache in den Debatten zeigt sich ab der Wahl 2017 eine deutliche Entwicklung zu steigender Negativität. In der Endphase kann eine Abschwächung dieses



(a) Darstellung des zeitlichen Verlaufs der Sentiment-Scores im Landtag von Schleswig-Holstein auf der jährlichen Betrachtungsebene



(b) Darstellung des zeitlichen Verlaufs der Scores für die Erkennung von Hassrede im Landtag von Schleswig-Holstein auf der jährlichen Betrachtungsebene

Abbildung 6.5: Visualisierung der Ergebnisse der Sentiment-Analyse und Erkennung von Hassrede für den Landtag in Schleswig-Holstein

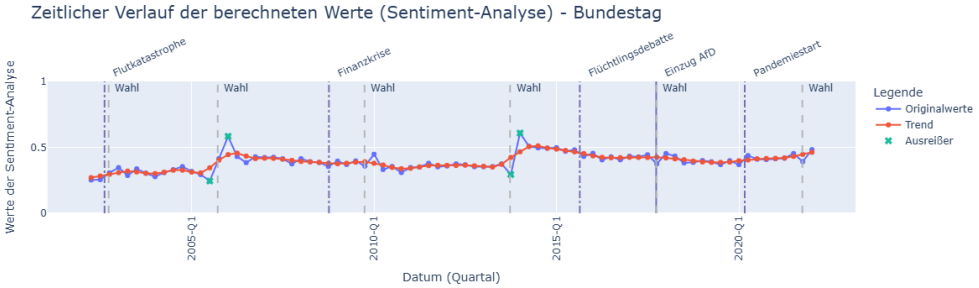
Trends festgestellt werden. Im Hessischen Landtag stellen die Trendkurven unterschiedliche Verläufe dar. Im Allgemeinen zeichnet sich ein Rückgang negativer Stimmung ab. Dieser Annahme stehen die Verläufe in den jährlichen Trendkurven am Ende des untersuchten Zeitraums entgegen. In diesen wird eine leichte negative Entwicklung sichtbar. Daher lässt sich gegen Ende eine leichte Tendenz zu steigender negativer Stimmung vermuten.

Im Rahmen der zweiten sozialwissenschaftlichen Fragestellung: *Lassen sich Wendepunkte in der Grundstimmung identifizieren und fallen diese mit markanten Ereignissen zusammen?*, sollen Besonderheiten in den berechneten Werten auf der Quartalsebene betrachtet werden. Nach der Beschreibung der gefundenen Auffälligkeiten in den deutschen Gremien ist zu untersuchen, ob diese mit ausgewählten Ereignissen zusammenfallen. Folgende Ereignisse sind ausgewählt: die Flutkatastrophe von 2002, die Finanzkrise 2008, der Beginn der Flüchtlingsdebatte und der Corona-Pandemie sowie der Einzug der AfD in die Gremien.

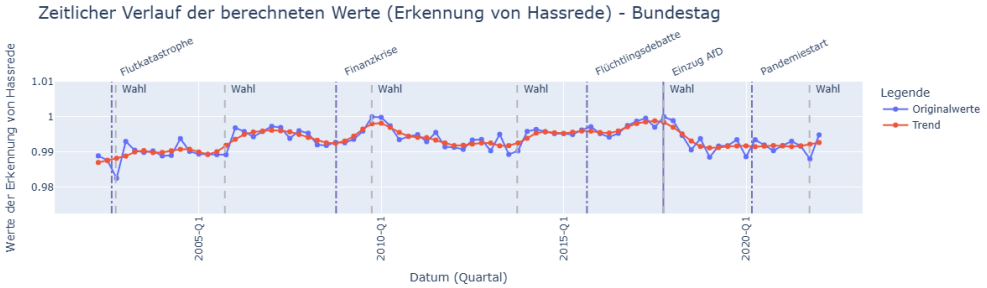
Da die dritte Fragestellung den Fokus auf den Zeitraum um die Wahltermine legt, werden Besonderheiten, die in diesem Zusammenhang stehen, erst im darauffolgenden Abschnitt thematisiert.

Die Betrachtungsebene der Sentiment-Analyse auf Quartalsebene zeigt für die Trendkurve (Abbildung 6.6a) des deutschen Bundestages zwei auffällige Anstiegsphasen. Die Zeiträume des Anstiegs lassen sich zum einen auf den Zeitraum von Beginn (2002) bis zum ersten Quartal 2006 und zum anderen auf das erste Quartal 2013 bis zum zweiten in 2014 eingrenzen. Zusätzlich ist die Darstellung 6.6b mit den Ergebnissen der Erkennung von Hassrede zu

analysieren. Rückgänge der Negativität treten insbesondere in den Perioden Q1/2005–Q4/2006 sowie Q3/2008–Q4/2009 auf. Markante Phasen steigender Negativität bilden die Zeiträume zwischen den Wahlen 2009 und 2013 sowie Q2/2017-Q4/2018.



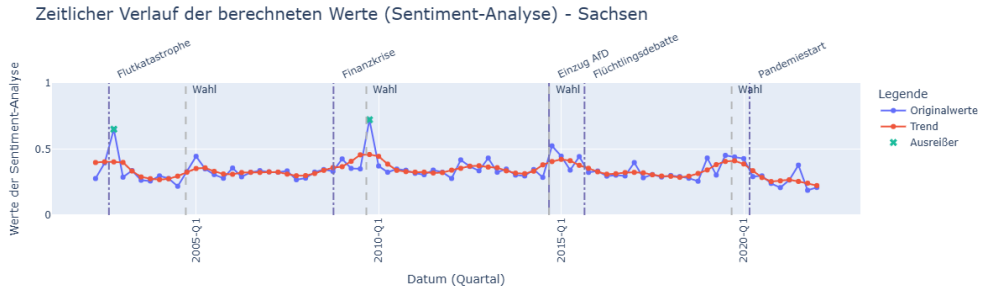
(a) Darstellung des zeitlichen Verlaufs der Sentiment-Scores für den Deutschen Bundestag mit markierten Ereignissen



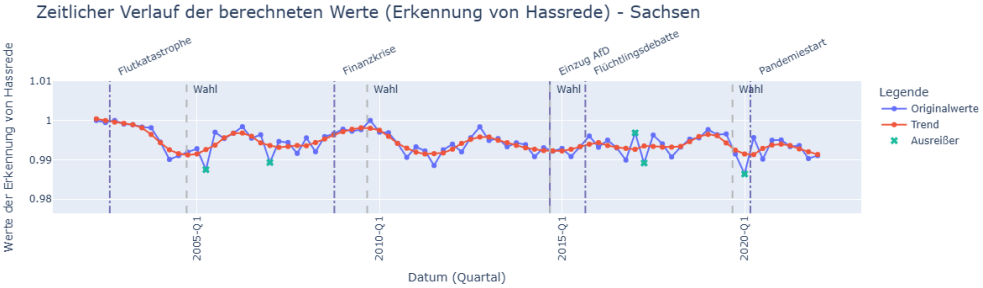
(b) Darstellung des zeitlichen Verlaufs der Werte für die Erkennung von Hassrede für den Deutschen Bundestag mit markierten Ereignissen

Abbildung 6.6: Visualisierung der Ergebnisse der Sentiment-Analyse und Erkennung von Hassrede für den Deutschen Bundestag inklusive der Markierung ausgewählter Ereignisse

In Abbildung 6.7a sind die Sentiment-Scores auf der Quartalsebene für das Bundesland Sachsen dargestellt. Auffällig ist die starke Zunahme an Negativität beginnend in Q4/2004 bis Q4/2003. Eine weitere Besonderheit in dieser Abbildung stellt der Ausreißer in Q3/2002 dar. Die deutliche Entwicklung zu steigender Negativität gegen Ende des Untersuchungsintervalls (2002-2021) ist ab Q3/2019 zu erkennen. Weitere Besonderheiten konnten ebenfalls aus den Ergebnissen der Erkennung von Hassrede festgestellt werden. In Abbildung 6.7b ist die steigende Negativität zu Beginn des Zeitintervalls (2002-2021) zu nennen, die in Q3/2004 endet. Eine weitere auffällige steigende Negativität ist zwischen Q2/2009 - Q1/2011 erkennbar. Markierte Ausreißer in dieser Abbildung bilden Q1/2005, Q4/2006, Q4/2016, Q1/2017 und Q4/2019.



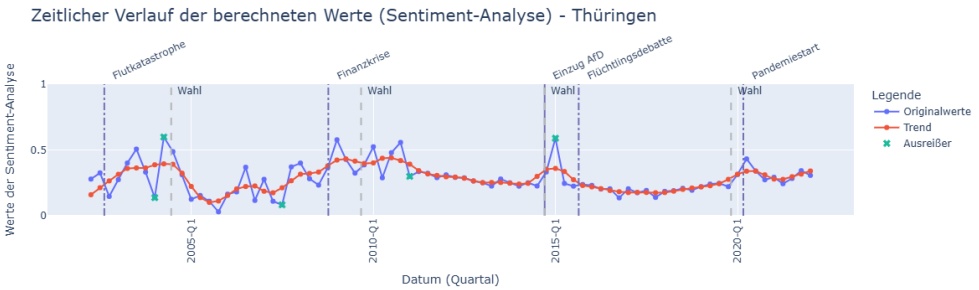
(a) Darstellung des zeitlichen Verlaufs der Sentiment-Scores für den Sächsischen Landtag mit markierten Ereignissen



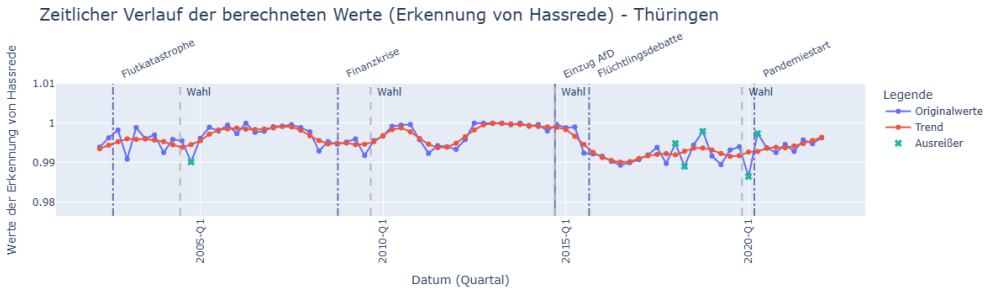
(b) Darstellung des zeitlichen Verlaufs der Werte für die Erkennung von Hassrede für den Sächsischen Landtag mit markierten Ereignissen

Abbildung 6.7: Visualisierung der Ergebnisse der Sentiment-Analyse und Erkennung von Hassrede für den Sächsischen Landtag inklusive der Markierung ausgewählter Ereignisse

Für die Identifizierung von Besonderheiten im Thüringischen Landtag ist die Abbildung 6.8 zu betrachten. Im Verlauf der Trendkurve der Sentiment-Scores fällt die starke Schwankung zwischen Q2/2004 und Q1/2009 auf. Der Beginn dieser Schwankung stellt die stärkste Entwicklung zu steigender Negativität dar. Zeitlich kann diese auf das Intervall von Q2/2004 - Q2/2005 eingegrenzt werden. Ein markanter Ausreißer ist in Q4/2014 markiert. Auffälligkeiten aus der Erkennung von Hassrede sind der Abbildung 6.8b zu entnehmen. Dabei werden zwei markante Schwankungen sowie eine ausgeprägte Periode zunehmender Negativität deutlich. Die Schwankungen lassen sich auf einen Zeitraum von Q1/2007 bis Q4/2012 eingrenzen. Die niedrigsten Werte innerhalb dieses Intervalls liegen in Q1/2009 und Q2/2011, während der Höchstwert in Q2/2010 erreicht ist. Ab Q3/2014 beginnt die deutlichste Entwicklung zu steigender Negativität der Trendkurve.



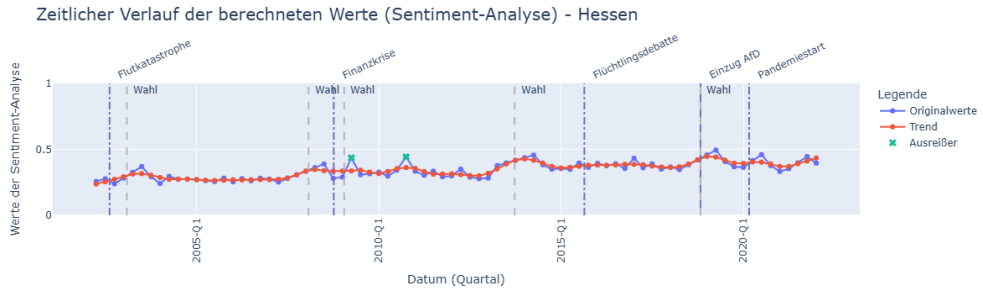
(a) Darstellung des zeitlichen Verlaufs der Sentiment-Scores für den thüringischen Landtag mit markierten Ereignissen



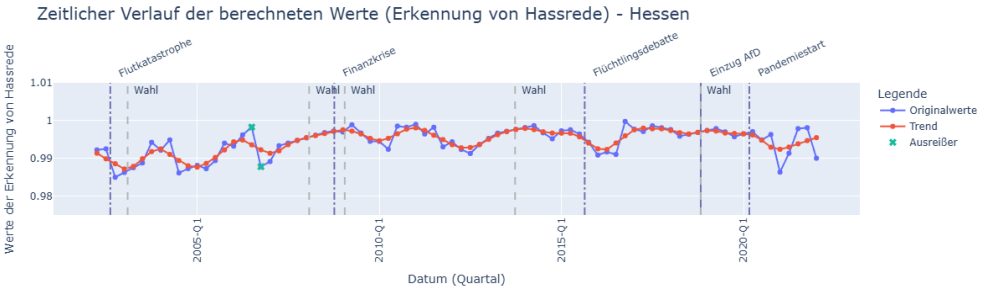
(b) Darstellung des zeitlichen Verlaufs der Werte für die Erkennung von Hassrede für den Thüringischen Landtag mit markierten Ereignissen

Abbildung 6.8: Visualisierung der Ergebnisse der Sentiment-Analyse und Erkennung von Hassrede für den Thüringischen Landtag inklusive der Markierung ausgewählter Ereignisse

Die Untersuchung der Ergebnisse des Hessischen Landtages erfolgt anhand der Abbildung 6.9. Die Sentiment-Scores zeigen einen Rückgang an Negativität, der das darauffolgende Niveau der Kurve anhebt. Diese Periode des Rückgangs lässt sich auf den Zeitraum Q3/2012 bis Q4/2013 eingrenzen. Eine Besonderheit in den markierten Ausreißern bildet das dritte Quartal 2010. Auffällig im Trendkurvenverlauf der Erkennung von Hassrede sind die Schwankungen zu Beginn des Beobachtungsintervalls. In der Abbildung 6.9b ist die Kurve durch mehrere An- und Abstiegsphasen bis zum vierten Quartal 2006 charakterisiert. Die Höchstwerte in diesem Zeitraum liegen in Q4/2004 und Q1/2006, während die niedrigsten Werte im vierten Quartal 2002 und 2004 liegen. Zwei weitere Schwankungen in die Richtung steigender Negativität lassen sich auf die Zeiträume Q4/2010 - Q4/2013 sowie Q1/2015 Q1/2017 eingrenzen. Die niedrigsten Werte liegen zum einen in Q1/2012 und zum anderen in Q1/2016. Ab Q1/2020 ist eine erneute Entwicklung zu steigender Negativität erkennbar. Die Abbildung 6.9b weist Ausreißer in Q2/2006 sowie im folgenden Quartal auf.



(a) Darstellung des zeitlichen Verlaufs der Sentiment-Scores für den Hessischen Landtag mit markierten Ereignissen

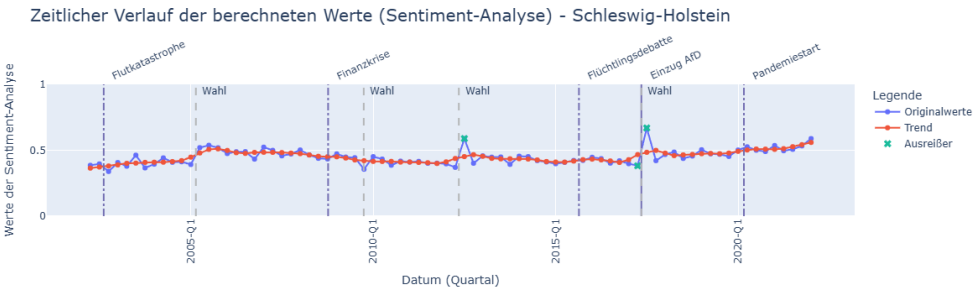


(b) Darstellung des zeitlichen Verlaufs der Werte für die Erkennung von Hassrede für den Hessischen Landtag mit markierten Ereignissen

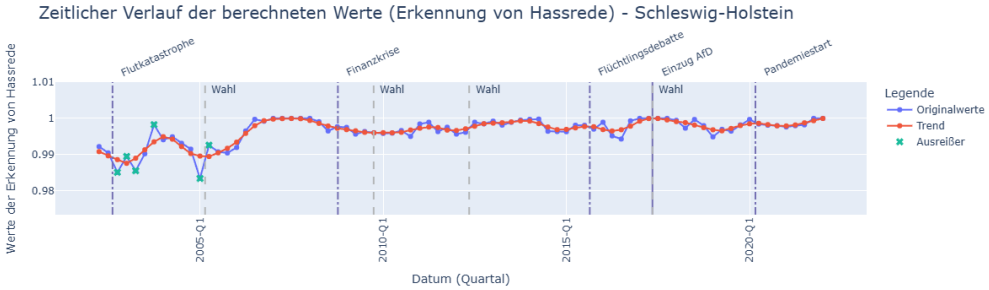
Abbildung 6.9: Visualisierung der Ergebnisse der Sentiment-Analyse und Erkennung von Hassrede für den Hessischen Landtag inklusive der Markierung ausgewählter Ereignisse

Abschließend ist die Abbildung 6.10 für den Landtag in Schleswig-Holstein auf Besonderheiten zu untersuchen. Zu Beginn des Untersuchungszeitraums lassen sich in der Trendkurve der Erkennung von Hassrede markante Perioden abnehmender Negativität erkennen. Zeitlich sind diese zum einen auf Q4/2002 bis Q4/2003 und zum anderen auf Q1/2005 bis Q3/2007 einzugrenzen.

Die ausgewählten Ereignisse: Flutkatastrophe 2002, Finanzkrise 2008, Beginn der Flüchtlingsdebatte und der Corona-Pandemie sowie der Einzug der AfD in die Gremien sind in den Abbildungen (6.6 - 6.10) der Quartalsebene markiert. Anhand dieser wird analysiert, ob die im vorherigen beschriebenen Änderungen in der Stimmung mit den markierten Ereignissen zusammenfallen.



(a) Darstellung des zeitlichen Verlaufs der Sentiment-Scores für den Landtag in Schleswig-Holstein mit markierten Ereignissen



(b) Darstellung des zeitlichen Verlaufs der Werte für die Erkennung von Hassrede für den Landtag in Schleswig-Holstein mit markierten Ereignissen

Abbildung 6.10: Visualisierung der Ergebnisse der Sentiment-Analyse und Erkennung von Hassrede für den Landtag in Schleswig-Holstein inklusive der Markierung ausgewählter Ereignisse

Im Sächsischen Landtag zeigen die Ergebnisse der Sentiment-Analyse 6.7a, einen positiven Ausreißer kurz nach der Markierung der Flutkatastrophe 2002. Die Ergebnisse aus der Erkennung von Hassrede im Landtag von Schleswig-Holstein weisen um dieses Ereignis Ausreißer auf.

Weiterhin zeigt die Abbildung 6.8b zur Erkennung von Hassrede in Thüringen, dass die Markierung der Finanzkrise in die Schwankung zunehmender Negativität fällt, die von Q1/2007 bis Q2/2010 erkennbar ist. In der Darstellung der Sentiment-Analyse 6.8a ist die Markierung dagegen innerhalb einer Phase abnehmender Negativität in der Trendkurve. Die Betrachtung der Finanzkrise im Bundestag zeigt in der Abbildung 6.6b für die Erkennung von Hassrede, dass die Markierung im Rahmen der Schwankungen zwischen Q4/2006 und Q4/2009 liegt.

In den Abbildungen 6.8b und 6.6b, für die Untersuchung auf das Vorkommen von Hassrede ist zu erkennen, dass eine zunehmende Negativität nach den Wahlen 2014 in Thüringen und 2017 im Bundestag einsetzt. Den Grafiken ist weiterhin zu entnehmen, dass diese beiden Wahlen den Einzug der AfD darstellen.

Im Zusammenhang mit dem Beginn der Flüchtlingsdebatte zeigt Abbildung 6.8b zur Erkennung von Hassrede in Thüringen, dass die Markierung in die bereits nach der Wahl 2014 einsetzende Entwicklung hin zu steigender Negativität fällt. Im Hessischen Landtag 6.9b fällt dieses Ereignis ebenfalls in eine bereits eingesetzte Entwicklung zu steigender Negativität.

Das letzte ausgewählte Ereignis ist der Beginn der Pandemie. Dabei zeigt sich, dass die Markierung des Ereignisses in der Abbildung der Sentiment-Scores für Sachsen 6.7b innerhalb einer Phase steigender Negativität liegt.

Basierend auf den beschriebenen Besonderheiten in den Kurvenverläufen sowie markierten Ausreißern ist die Fragestellung: *Lassen sich Wendepunkte in der Grundstimmung identifizieren und fallen diese mit markanten Ereignissen zusammen?* zu untersuchen. Die ausgewählten Methoden ermöglichen die Identifikation von Wendepunkten sowie Änderungen in Stimmung und Tonalität innerhalb der Gremien. Eine zeitliche Eingrenzung ist ebenfalls möglich. Eine spezifische Zuordnung der genauen Ereignisse, die Änderungen in der Tonalität verursachen können, ist mit den eingesetzten Methoden nicht möglich.

Im Rahmen der Untersuchung ausgewählter Ereignisse: Flutkatastrophe 2002, Finanzkrise 2008, Beginn der Flüchtlingsdebatte und der Corona-Pandemie und Einzug der AfD in die Gremien zeigt sich, dass es mit den Methoden realisierbar ist, zu untersuchen, ob ausgewählte Ereignisse und identifizierte Änderungen zusammenfallen. Dabei können ausschließlich Hypothesen über mögliche Zusammenhänge formuliert werden. Eine Überprüfung dieser Hypothesen ist Gegenstand weiterführender Untersuchungen, die nicht im Rahmen dieser Arbeit durchgeführt werden.

Die Analyse zeigt, dass die ausgewählten Ereignisse ausschließlich in vereinzelten Abbildungen mit markanten Änderungen der Stimmung oder markierten Ausreißern zusammenfallen. Eine allgemeine Aussage kann daher für die Ereignisse nicht getroffen werden. Auffällig ist jedoch die deutliche Zunahme an Negativität in Thüringen und dem Bundestag nach dem Einzug der AfD, erkennbar in den Ergebnissen aus der Erkennung von Hassrede. Die verwendeten Methoden können somit als Indikator und Ausgangspunkt für die Beantwortung der sozialwissenschaftlichen Fragestellung (b) dienen.

In der dritten sozialwissenschaftlichen Fragestellung (c): *Lässt sich eine Veränderung der Sprache vor und nach Wahlen feststellen?* wird der Zeitraum um die jeweiligen Wahltermine untersucht.

Aus der Darstellung der Sentiment-Analyse für den Bundestag 6.6a ist zu entnehmen, dass um die Wahltermine 2005 und 2013 eine sprunghafte Abnahme an Negativität in den originalen Werten zu verzeichnen ist. Der Trendkurvenverlauf in dieser Darstellung zeigt zudem um diese Wahltermine denselben Trend. Zusätzlich werden für den Bundestag die Ergebnisse der Erkennung von Hassrede 6.6b analysiert. Die Analyse des Trendkurvenverlaufs lässt eine Abnahme an Negativität um die Wahlen in den Jahren 2005 und 2009 erkennen. Dieser Trend zeigt sich ebenfalls um die Wahl 2013 in abgeschwächter Form. Nach der Wahl 2017 ist dagegen eine starke Zunahme an Negativität sichtbar.

Die Abbildung 6.7a der berechneten Sentiment-Scores in Sachsen zeigt, dass um alle Wahltermine zunächst Phasen mit abnehmender Negativität auftreten, denen nach den Terminen Phasen zunehmender Negativität folgen. Erfolgt die Einbeziehung der Ergebnisse aus der Erkennung von Hassrede, liegt der Zeitraum um die Wahlen 2004 und 2019 in Schwankungen zu steigender Negativität der Trendkurve (siehe 6.7b). Nach der Wahl 2014, die noch im Hochpunkt der Phase abnehmender Negativität liegt, folgt eine deutliche Zunahme an Negativität.

Die Analyse im Thüringischen Landtag zeigt aus den Ergebnissen der Sentiment-Analyse 6.8a, dass die Wahlen 2014 und 2019 innerhalb von Entwicklungen zu sinkender Negativität

liegen. Auf beide Entwicklungen folgt nach den Wahlterminen eine Zunahme an Negativität. Auffällig ist der Zeitraum um die Wahlen 2004 und 2014. In den originalen Werten aus der Sentiment-Analyse ist in beiden Fällen ein Ausreißer, der einen hohen Anteil neutraler Klassifizierungen darstellt, zu erkennen. Nach der Wahl 2004 zeigt sich erneut eine starke Entwicklung zu steigender Negativität. In den Werten für die Erkennung von Hassrede (siehe Abbildung 6.8b) zeigt sich, dass auf die Wahlen 2004, 2009 und 2019 Phasen sinkender Negativität folgen. Ausnahme ist dabei die Wahl 2014. Nach dieser Wahl ist eine deutliche Zunahme an Negativität zu erkennen.

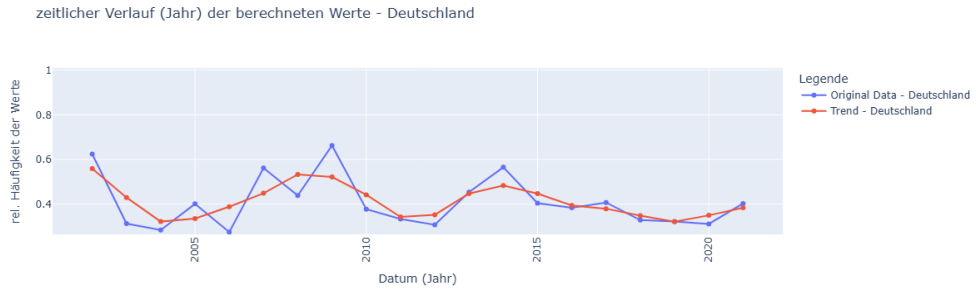
Ein erkennbares Muster in den originalen aggregierten Sentiment-Scores für das Bundesland Hessen 6.9a sind Anstiege, deren Höhepunkte nach den Wahlen auftreten. Auf diese Anstiege folgt der deutliche Abfall auf das Niveau der vorherigen Werte. Dabei entspricht ein Anstieg einer Entwicklung zu sinkender Negativität, ein Abfall einer Zunahme an Negativität. Nach der Wahl 2008 ist zudem ein Ausreißer markiert, der einen besonders hohen Anteil neutral klassifizierter Einträge aufweist. Die Analyse der Trendkurvenverläufe zeigt, dass im Zeitraum vor jeder Wahl eine Phase abnehmender Negativität erkennbar ist. In der Abbildung 6.9b der Untersuchungsergebnisse zur Erkennung von Hassrede ist um die Wahlen 2008, 2009 und 2013 ein ähnliches Muster in abgeschwächter Form zu erkennen. Die Wahl 2003 liegt im Tiefpunkt einer Phase steigender Negativität. Anzumerken ist, dass nach der Wahl im Jahr 2009 eine leichte Zunahme an Negativität erkennbar ist.

Die Abbildung 6.10a visualisiert die berechneten Sentiment-Scores für Schleswig-Holstein. Nach den Wahlterminen 2012 und 2017 wurden in den originalen Werten Ausreißer markiert. Diese stellen einen hohen Anteil an neutral klassifizierten Beiträgen dar. Aus dem Trendkurvenverlauf um die Wahltermine ist zu entnehmen, dass die Wahlen 2005, 2012 und 2017 in Phasen der sinkenden negativen Stimmung liegen. Die Abbildung zur Erkennung von Hassrede 6.10b zeigt eine weitere Auffälligkeit in den originalen Werten um den Wahltermin 2005. Vor diesem ist ein Ausreißer markiert, der einen gestiegenen Anteil an negativ klassifizierten Beiträgen darstellt. Der darauffolgende Datenpunkt ist ebenfalls markiert und repräsentiert den wieder gestiegenen Anteil an neutralen Beiträgen. Der Verlauf der Trendkurve zeigt, dass die Wahl 2005 am Ende einer Phase zunehmender Negativität liegt. Im Anschluss daran folgt eine Phase abnehmender Negativität. Weiterhin zeigt sich eine leichte Entwicklung zu mehr Negativität nach der Wahl 2017.

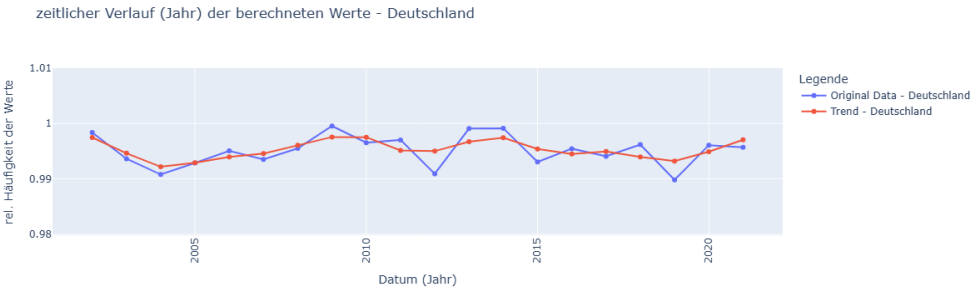
Auf Grundlage der dargestellten Erkenntnisse lässt sich feststellen, dass der Zeitraum um Wahltermine häufig durch Phasen von sinkender Negativität gekennzeichnet ist. Besonders markant sind die An- und Abstiegsphasen der Trendkurve in den Sentiment-Scores des Sächsischen Landtags 6.7a. Ein ähnliches Muster zeigt sich in den originalen Sentiment-Scores des Hessischen Landtags (6.9a). Dies deutet darauf hin, dass vor Wahlen tendenziell auf negative Sprache verzichtet wird. Auffällig sind zudem die deutlichen Entwicklungen zu steigender Negativität in den Sentiment-Scores 6.8a nach der Wahl 2004 und in den Werten der Erkennung von Hassrede 6.8b nach der Wahl 2014 in Thüringen. Ein ähnliches Verhalten ist, für den Bundestag aus der Erkennung von Hassrede 6.6b nach der Wahl 2017 zu erkennen. Es ist auf dieser Grundlage festzustellen, dass um Wahltermine sichtbare Veränderungen in der Stimmung auftreten.

Die vierte aufgestellte sozialwissenschaftliche Fragestellung (d): *Ist ein negativer Bias gegenüber den Regionen des ehemaligen Ostens Deutschlands zu erkennen?*, thematisiert spezifisch die Regionen des ehemaligen Ostens Deutschlands. Dafür wird ein eigenständiger Datenauszug

erstellt, der ausschließlich Einträge in zeitlicher Nähe zum Tag der Deutschen Einheit umfasst. Für die anschließende Visualisierung und Analyse werden die Daten entsprechend gruppiert. Zunächst werden die Ergebnisse ohne eine Differenzierung in die Bundesländer und den Bundestag dargestellt. Weiterhin sollen die Ergebnisse für die Repräsentanten der ostdeutschen sowie der westdeutschen Bundesländer getrennt untersucht werden. Abschließend sind die Ergebnisse für den Bundestag dargestellt. Die Analyse erfolgt hierbei ausschließlich auf der jährlichen Ebene.



(a) Darstellung des zeitlichen Verlaufs der Sentiment-Scores für den gesamten Datensatz der Redebeiträge um den Tag der Deutschen Einheit



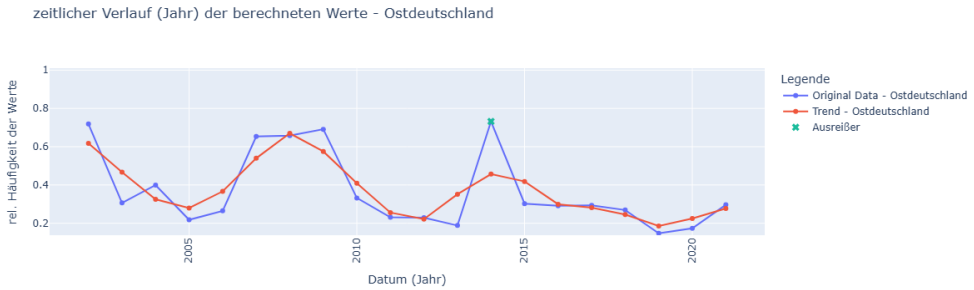
(b) Darstellung des zeitlichen Verlaufs der Werte für die Erkennung von Hassrede für den gesamten Datensatz der Redebeiträge um den Tag der Deutschen Einheit

Abbildung 6.11: Visualisierung der Ergebnisse der Sentiment-Analyse und Erkennung von Hassrede für den gesamten Datensatz der Redebeiträge um den Tag der Deutschen Einheit

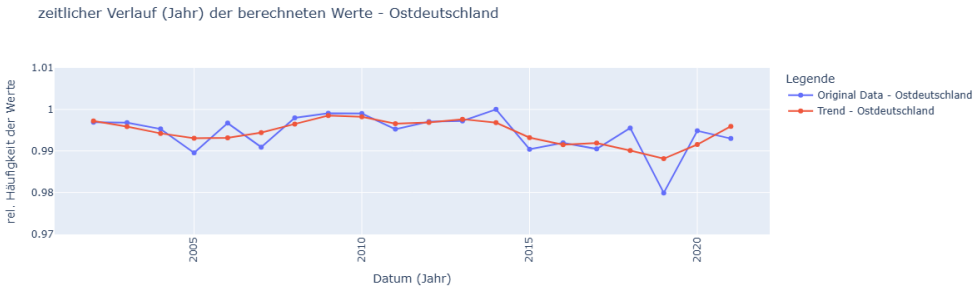
Die Trendkurve der Sentiment-Scores ohne eine Differenzierung (Abbildung 6.11a) zeigt keinen eindeutig allgemein feststellbaren Trend. Die Trendkurve ist durch starke Schwankungen gekennzeichnet. Eine einsetzende Entwicklung zu sinkender Negativität beginnt im Jahr 2019 und bleibt bis zum Ende des Zeitintervalls (2002-2021) bestehen. Im Verlauf der Trendkurve 6.11b für die Ergebnisse der Erkennung von Hassrede ist zu Beginn eine Entwicklung zu steigender Negativität erkennbar. Im Allgemeinen zeigt die Abbildung hingegen eine Abnahme an Negativität.

Eine Analyse der Sentiment-Scores 6.12a, die sich aus den Werten für Sachsen und Thüringen zusammensetzen, lässt ebenfalls keinen eindeutigen allgemeinen Trend erkennen. Der Kurvenverlauf weist erneut zu Beginn des Untersuchungszeitraums (2002-2021) starke Schwankungen auf. Es lässt sich zunächst eine Entwicklung zu steigender Negativität vermuten. Ab dem

Jahr 2020 ist ein entgegenwirkender Trend zu erkennen. Die Betrachtung der Ergebnisse zur Erkennung von Hassrede in Abbildung 6.12b verdeutlicht die ab dem Jahr 2020 einsetzende Abnahme an Negativität.



(a) Darstellung des zeitlichen Verlaufs der Sentiment-Scores für die Repräsentanten des Osten Deutschlands aus dem Datensatz der Redebeiträge um den Tag der Deutschen Einheit

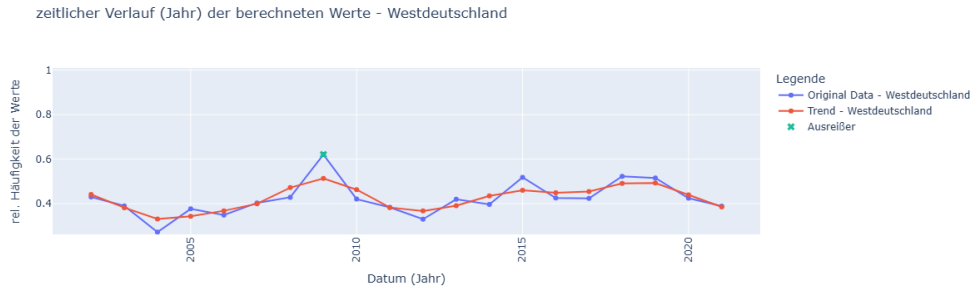


(b) Darstellung des zeitlichen Verlaufs der Werte für die Erkennung von Hassrede für die Repräsentanten des Osten Deutschlands aus dem Datensatz der Redebeiträge um den Tag der Deutschen Einheit

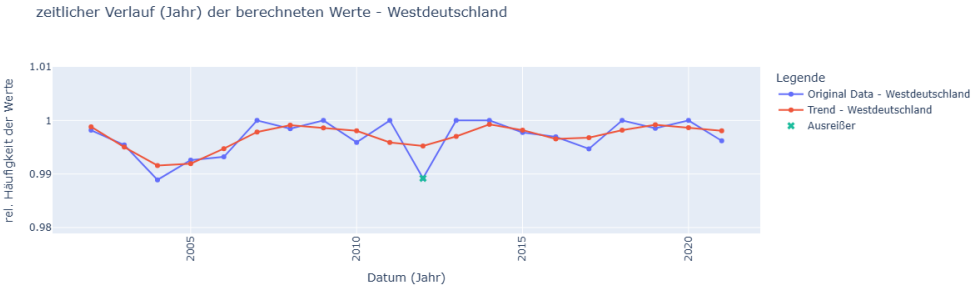
Abbildung 6.12: Visualisierung der Ergebnisse der Sentiment-Analyse und Erkennung von Hassrede für die Repräsentanten des Osten Deutschlands aus dem Datensatz der Redebeiträge um den Tag der Deutschen Einheit

Die Abbildung 6.13a stellt die vereinigten Sentiment-Scores von Hessen und Schleswig-Holstein dar. Bis ins Jahr 2009 lässt sich eine allgemeine Abnahme an Negativität vermuten. Der weitere Verlauf der Trendkurve zeigt nach einer Zunahme an Negativität im Zeitraum von 2009 bis 2012 erneut eine Entwicklung zu sinkender Negativität, die im Jahr 2019 endet. Bis zum Ende des Untersuchungszeitraums (2002-2021) zeigt sich eine Zunahme an Negativität. Die zusätzliche Betrachtung der Ergebnisse aus der Erkennung von Hassrede bestätigt diese Erkenntnisse. Der Trendkurvenverlauf in dieser Abbildung 6.13b deutet ebenfalls zunächst auf eine allgemeine Abnahme an Negativität hin, die ab dem Jahr 2019 unterbrochen wird.

Für den Bundestag ist in der Visualisierung (Abbildung 6.14a) der Sentiment-Analyse zu erkennen, dass die Trendkurve mehrere Phasen der Zu- und Abnahme der Negativität aufweist. Das generelle Niveau der Kurve bleibt bestehen. Ein allgemeiner Trend kann nicht eindeutig festgestellt werden. Die ab dem Jahr 2013 einsetzende Zunahme an Negativität wird im Jahr 2019 durch eine starke Entwicklung zu sinkender Negativität bis zum Ende des



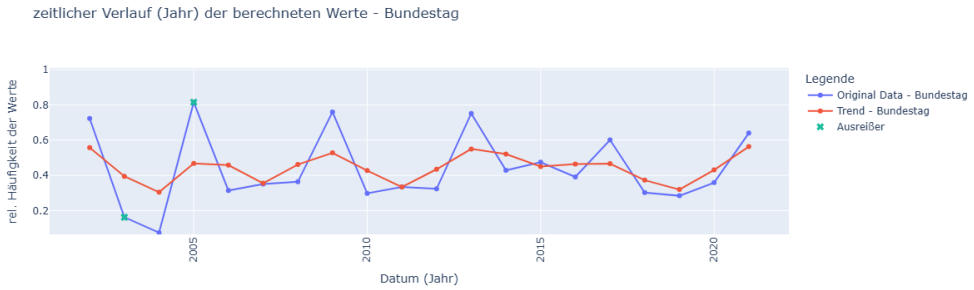
(a) Darstellung des zeitlichen Verlaufs der Sentiment-Scores für die Repräsentanten des Westen Deutschlands aus dem Datensatz der Redebeiträge um den Tag der Deutschen Einheit



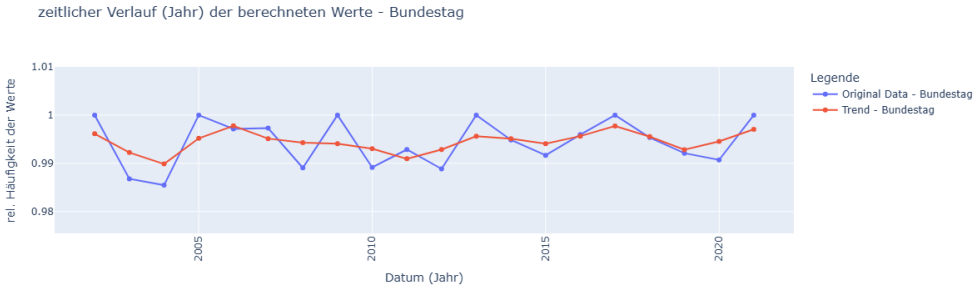
(b) Darstellung des zeitlichen Verlaufs der Werte für die Erkennung von Hassrede für die Repräsentanten des Westen Deutschlands aus dem Datensatz der Redebeiträge um den Tag der Deutschen Einheit

Abbildung 6.13: Visualisierung der Ergebnisse der Sentiment-Analyse und Erkennung von Hassrede für die Repräsentanten des Westen Deutschlands aus dem Datensatz der Redebeiträge um den Tag der Deutschen Einheit

Untersuchungszeitraums (2002-2021) beendet. Ein analoges Verhalten der Trendkurve ist ebenfalls aus den in Abbildung 6.14b dargestellten Ergebnissen der Erkennung von Hassrede abzulesen.



(a) Darstellung des zeitlichen Verlaufs der Sentiment-Scores für den Deutschen Bundestag aus dem Datensatz der Redebeiträge um den Tag der Deutschen Einheit



(b) Darstellung des zeitlichen Verlaufs der Werte für die Erkennung von Hassrede für den Deutschen Bundestag aus dem Datensatz der Redebeiträge um den Tag der Deutschen Einheit

Abbildung 6.14: Visualisierung der Ergebnisse der Sentiment-Analyse und Erkennung von Hassrede für den Deutschen Bundestag aus dem Datensatz der Redebeiträge um den Tag der Deutschen Einheit

Diese Untersuchungen in Bezug auf die Fragestellung: *Ist ein negativer Bias gegenüber den Regionen des ehemaligen Ostens Deutschlands zu erkennen?*, haben gezeigt, dass die ausgewählten Methoden der Sentiment-Analyse sowie der Erkennung von Hassrede nicht ausreichen, um diese direkt zu beantworten. Mit den ausgewählten Methoden ist es möglich, den Verlauf der berechneten Werte in Bezug auf die Stimmung in zeitlicher Nähe zum Tag der Deutschen Einheit zu analysieren. Es ist weiterhin möglich, den Verlauf der Stimmung über die Jahre zu diesem Datum einzuschätzen. Die untersuchten Methoden reichen dabei nicht aus, um das Verhältnis bzw. die Einstellung in Bezug auf den Osten Deutschlands abzubilden. Aussagen über eine eventuell vorliegende negative Voreingenommenheit können mit diesen Methoden nicht getroffen werden.

6.2 Einordnung der Untersuchungsmethode

Ziel der dargestellten praktischen Arbeit ist es, mit geeigneten Methoden aus dem Bereich der Sentiment-Analyse sowie ergänzend der Erkennung von Hassrede, das Stimmungsbild der Debatten in deutschen politischen Gremien zu untersuchen. Nachdem das Verfahren mit der besten Modellleistung aus beiden Bereichen zur Datenanalyse eingesetzt werden, zeigt die Untersuchung, inwieweit die ausgewählten NLP-Methoden zur Bearbeitung der sozialwissenschaftlichen Fragestellungen geeignet sind.

Die Evaluation der verwendeten Methoden hat gezeigt, dass transformer-basierte Methoden die besten Ergebnisse erreichen. Für die Sentiment-Analyse erzielte das auf Parlamentsdaten trainierte Fine-Tuning Modell, basierend auf ELECTRA, die besten Resultate. Die Untersuchung der Analysedaten im Hinblick auf Hassrede wird mit einer Kombination aus einem auf die Erkennung von Hassrede trainierten Modell und einem LLM durchgeführt. Damit bestätigt sich die Erkenntnis des Forschungsstandes 3 hinsichtlich des Potenzials transformer-basierter Methoden.

Insbesondere zeigt das Training eines eigenen Fine-Tuning-Modells, dass eine deutlich bessere Modellleistung erkennbar ist, wenn der Trainingsprozess überwiegend auf Parlamentsdaten basiert. Die hauptsächlich mit Twitterdaten trainierten transformer-basierten Sentiment-Analyse Modelle erreichen vergleichsweise schlechtere Ergebnisse. Daher ist zu vermuten, dass ein Modell, das mit Daten trainiert wird, die in Art und Struktur den in der vorliegenden Arbeit zu untersuchenden Parlamentsdaten ähnlich sind, für diesen Anwendungsfall bessere Ergebnisse erreichen kann.

Deutlich wird weiterhin, dass die auf Sentiment-Analyse trainierten Modelle in der binären Klassifikation hinsichtlich der Accuracy schlechter als beide Wörterbücher sind. In der gewichteten F1-Metrik erzielt nur das mehrsprachig trainierte Modell einen besseren Wert als das domänenspezifische Wörterbuch. Das deutschsprachige und das mehrsprachig-politisch trainierte Modell erreichen dagegen in dieser Metrik geringere Werte. Dies verdeutlicht, dass der direkte Einsatz trainierter Sentiment-Analyse-Modelle auf die vorliegende Datengrundlage nicht zwangsläufig zu besseren Ergebnissen führt.

(Rheault et al., 2016) unterstützen diese Ergebnisse indem darauf hingewiesen wird, dass Modelle deren Trainingsdaten überwiegend aus sozialen Medien stammen, auf Parlamentsdaten schlechtere Ergebnisse erzielen. Als Begründung wird angeführt, dass Reden in Parlamenten anspruchsvollere Sprache aufweisen, während in sozialen Medien Umgangssprache dominiert (Rheault et al., 2016).

Die praktische Untersuchung ermöglicht es, für die vorliegende Datengrundlage das jeweils beste Verfahren aus den ausgewählten Methoden beider Bereiche auszuwählen. Darüber hinaus zeigt der gesamte Prozess, dass NLP-Methoden geeignet sein können, sozialwissenschaftliche Fragestellungen zu bearbeiten. Gleichzeitig wird deutlich, dass zur Beantwortung der aufgestellten Fragestellungen – insbesondere für (b) und (d) – weiterführende Forschung und weiterentwickelte methodische Ansätze erforderlich sind. Vor allem die Fragestellung (d) hat die Grenzen der Untersuchungsmethode und der ausgewählten Methoden verdeutlicht. In Hinblick auf diese Fragestellung wäre die zusätzliche Einbeziehung der Aspektebene in die Sentiment-Analyse zu beachten.

Es hat sich gezeigt, dass die praktische Arbeit zielführend ist, um die Forschungsfrage zu beantworten. Sie eignet sich zur Analyse des Stimmungsbildes der Debatten in politischen Gremien im Hinblick auf Fragestellungen aus dem Bereich der Sozialwissenschaften. Gleichzeitig zeigt die Untersuchungsmethode, in Bezug auf die Bearbeitung der sozialwissenschaftlichen Fragestellungen, Randbedingungen und Grenzen der ausgewählten Methoden.

Entgegen der mehrheitlichen Verwendung von BERT Modellen wie beispielsweise in (Abercrombie & Batista-Navarro, 2020b; Kawintiranon & Singh, 2022) oder (T. Schmidt et al., 2022) wird im Training dieser Untersuchung ein ELECTRA-basiertes Modell aufgrund der Erkenntnisse aus (Widmann & Wich, 2022) verwendet. (Widmann & Wich, 2022) analysierten in ihrer Forschungsarbeit verschiedene Methoden der Sentiment-Analyse. Ihr trainiertes ELECTRA-Modell erreicht dabei die besten Resultate. Das Fine-Tuning Modell der vorliegenden Arbeit basiert auf dem gleichen Ausgangsmodell und zeigt ebenfalls die besten Ergebnisse. Insbesondere die Modifikation der Lernrate hat zu einer weiteren Verbesserung der Modellleistung geführt.

Die unterstützende Betrachtung zur Erkennung von Hassrede und beleidigender Sprache stellt ein Merkmal dieser Arbeit dar. Forschungsarbeiten, die die Stimmung in Gremien untersuchen, wenden mehrheitlich Methoden der Sentiment-Analyse an (Lehtosalo & Nerbonne, 2020; Pätz et al., 2025; Rheault et al., 2016). Für die Untersuchung der Fragestellungen (a): *Sind die Debatten in den deutschen Gremien aggressiver und negativer geworden?* und (b): *Lassen sich Wendepunkte in der Grundstimmung identifizieren und fallen diese mit markanten Ereignissen zusammen?* zeigt sich der Mehrwert für die Analyse. Insbesondere für den Punkt 'aggressiver' in Fragestellung (a) ist die Betrachtung dieser Methode zielführend. Eine starke Entwicklung zu steigender Negativität in den Visualisierungen der Erkennung von Hassrede lässt daher ein deutlich negatives Stimmungsbild vermuten.

Im Forschungsstand wird das von (Erhard et al., 2024) erstellte Modell zur Analyse von Parlamentsdaten aufgeführt. Dieses wird nicht im Rahmen der Erkennung von Hassrede eingesetzt. Es verwendet im Gegensatz zu den ausgewählten Methoden im Training Redebeiträge des Bundestages und wäre daher grundsätzlich geeigneter. Allerdings steht im Rahmen der Forschungsfrage sowie der formulierten sozialwissenschaftlichen Fragestellungen nicht die Analyse populistischer Rhetorik im Fokus.

Eine besondere methodische Erweiterung stellt die Einbindung eines LLMs dar. Die Darlegung im Forschungsstand 3 zeigt deren Einsatz in den Aufgabenbereichen von NLP. Zudem wird durch die Arbeit von (Li et al., 2024) der Einsatz in der politischen Domäne thematisiert. In Bezug auf Forschungsarbeiten mit deutschsprachigen Parlamentsdaten und sozialwissenschaftlichen Fragestellungen wird auch in aktuellen Arbeiten wie beispielsweise (Pätz et al., 2025) dieser Ansatz nicht mitbetrachtet. Die Ergebnisse der binären Klassifikation im Bereich der Sentiment-Analyse, verdeutlicht das Potential von LLMs. Es erreicht nach den Versionen 2-4 des Fine-Tuning-Modells die besten Ergebnisse für die Metriken Accuracy und gewichteter F1-Wert. Zudem zeigt sich dieses Potential vor allem bei der Erkennung von Hassrede. In diesem Bereich erreicht es für die gewichtete F1-Metrik den besten Wert. Eine Grenze dieses Ansatzes, besteht in der Größe des Datensatzes und der daraus resultierenden langen Antwortdauer.

In einigen im Forschungsstand erwähnten Arbeiten die Parlamentsdaten untersuchen, werden Methoden aus dem Bereich des maschinellen Lernens eingesetzt (Abercrombie & Batista-Navarro, 2020b; Pätz et al., 2025). Diese Methoden werden in der vorliegenden Arbeit nicht

berücksichtigt, da der Fokus auf Methoden wie transformer-basierten Modellen und LLMs liegt. Die Intention der Arbeit besteht darin, Methoden aus den Bereichen von Sentiment-Analyse und Erkennung von Hassrede, die aktuell Gegenstand der Forschung sind, in der politischen Domäne anzuwenden.

Die Untersuchung von deutschsprachigen Parlamentsdaten basiert in der aktuellen Forschung mehrheitlich auf Daten aus dem Deutschen Bundestag. Repräsentativ sind dabei die Arbeiten von (Pätz et al., 2025; Rauh, 2018b) und (Erhard et al., 2024) zu nennen. (Lange & Jentsch, 2023a) beziehen in ihre Experimente ebenfalls Daten der Bundesländer mit ein, allerdings erfolgt dabei die Differenzierung in Parteien. In der vorliegenden Arbeit liegt der Fokus neben dem Bundestag zusätzlich auf vier ausgewählten Landtagen. Die Ergebnisse der Methoden werden differenziert für die fünf betrachteten Gremien dargestellt und die sozialwissenschaftlichen Fragestellungen für jedes Gremium analysiert. Zudem ist die Auswahl der Landtage darauf ausgelegt die für Deutschland besondere Unterteilung von Osten und Westen nachzubilden.

Die Analyse der Polarität in politischen Gremien untersuchen (Rheault et al., 2016) für das englische und (Lehtosalo & Nerbonne, 2020) für das finnische Parlament. Beide Studien konnten einen positiven Trend feststellen. Im Rahmen dieser Arbeit ist eine eindeutige Entwicklung zu einem Rückgang an Negativität für den Landtag in Schleswig-Holstein zu erkennen. Im Gegensatz zu den Ergebnissen der beiden Studien zeigt der Sächsische Landtag hingegen eine steigende Negativität. Der Vergleich der Ergebnisse von (Rheault et al., 2016) und (Lehtosalo & Nerbonne, 2020) mit den Ergebnissen für den Deutschen Bundestag verdeutlicht, dass für das deutsche Parlament kein eindeutiger Trend festzustellen ist. Anzumerken ist, dass die vorliegende Arbeit keine positive Entwicklung feststellen kann, da die Klassifizierung in 'neutral' und 'negativ' durchgeführt wird.

(Lehtosalo & Nerbonne, 2020) untersuchen in ihrer Forschungsarbeit zudem ausgewählte Ereignisse. Sie entscheiden sich für die Corona-Pandemie sowie die Invasion des russischen Staates in die Ukraine (Lehtosalo & Nerbonne, 2020). Im Rahmen der sozialwissenschaftlichen Fragestellung nach markanten Ereignissen, die mit identifizierten Änderungen in der Tonalität zusammenfallen könnten (Fragestellung (b)), wird ebenfalls die Pandemie mit einbezogen. (Lehtosalo & Nerbonne, 2020) sind in ihrer Arbeit zu dem Ergebnis gekommen, dass die Möglichkeit eines Zusammenhangs zwischen dem Datum der ersten Beschränkung und einer negativen Periode besteht. Im Rahmen der vorliegenden Arbeit lässt sich diese Erkenntnis für den Sächsischen Landtag vermuten. In den anderen betrachteten Gremien gibt es keine eindeutigen Erkenntnisse. Anzumerken ist, dass im Gegensatz zu (Lehtosalo & Nerbonne, 2020) die Markierung am Tag der ersten Infektion gesetzt wird.

7 Schluss

Die vorliegende Arbeit untersucht Methoden aus dem Bereich von NLP zur Bearbeitung von sozialwissenschaftlichen Fragestellungen mit dem Schwerpunkt auf der Sentiment-Analyse. Im Zentrum steht die Analyse des Stimmungsbildes von Debatten in deutschen Gremien, wobei zwischen Bundes- und Landesebene differenziert wird. Zur Bearbeitung der Forschungsfrage besteht das Vorgehen darin Methoden aus dem Bereich von NLP auszuwählen und anschließend die geeignetste Methode für die Datengrundlage zu identifizieren. Der Fokus liegt dabei auf der Sentiment-Analyse mit einer unterstützenden Betrachtung der Erkennung von Hassrede. Das methodische Vorgehen unterteilt sich in die Evaluation der Methoden basierend auf der Datengrundlage und die Durchführung der Datenanalyse. Ziel der Evaluation ist es, für die Klassifizierung der Parlamentsdaten, die performanteste Methode aus beiden Bereichen zu identifizieren. Zur Realisierung wird zunächst ein Evaluationsdatensatz erstellt, der anschließend annotiert wird. Nach der Evaluation werden die performantesten Modelle auf die Daten angewendet. Die anschließende Untersuchung der sozialwissenschaftlichen Fragestellungen erfolgt basierend auf den Klassifizierungsergebnissen dieser Modelle. Durch die Untersuchung wird gezeigt, inwieweit die ausgewählten Methoden für die Analyse der aufgestellten sozialwissenschaftlichen Fragestellungen geeignet sind. Nachfolgende Ausführungen fassen die relevantesten Ergebnisse aus der Bearbeitung der Forschungsfrage zusammen und beschreiben darauffolgend Ansatzpunkte für weiterführende Forschungsarbeiten.

7.1 Zusammenfassung der Erkenntnisse

Die Ergebnisse teilen sich in zwei Punkte auf. Zum einen in die Auswahl der performantesten Methoden und zum anderen in die Analyse der Daten. Im Vergleich der ausgewählten Methoden zeigt sich, dass für die Sentiment-Analyse und die Erkennung von Hassrede transformerbasierte Modelle in diesem Anwendungsfall die besten Resultate erzielen. Weiterhin zeigt die Evaluation, dass das Training eines Modells auf der Grundlage von Parlamentsdaten für die Sentiment-Analyse der Datengrundlage zielführend ist. Zudem verdeutlicht die beste Methode zur Erkennung von Hassrede den methodischen Einsatz von LLMs für die Parlamentsdaten.

Die durchgeführte praktische Arbeit führt zur Beantwortung der Forschungsfrage: *'Welche Methoden aus dem Bereich von NLP sind geeignet, um ausgewählte sozialwissenschaftliche Fragestellungen zu untersuchen'*. Die ausgewählten Methoden der Sentiment-Analyse und der Erkennung von Hassrede erweisen sich als grundsätzlich geeignet, um sozialwissenschaftliche Fragestellungen zu adressieren. Insbesondere die Fragestellung (a): *'Sind die Debatten in den deutschen Gremien aggressiver und negativer geworden?'* und (c): *'Lässt sich eine Veränderung des Stimmungsbildes vor und nach Wahlen feststellen?'* lassen sich dadurch beantworten. Die Analyse zeigt für Fragestellung (a), dass sich im Sächsischen Landtag eine Entwicklung hin zu steigender Negativität abzeichnet, während für den Landtag in Schleswig-Holstein

eine Tendenz zu neutralerer Stimmung erkennbar ist. In den anderen Gremien konnte kein eindeutiges Ergebnis festgestellt werden. In Hinblick auf die Fragestellung (c) ergibt die Untersuchung, dass Änderungen im Stimmungsbild in den Zeiträumen um die Wahltermine erkennbar sind.

Für die Fragestellung (b): *'Lassen sich Wendepunkte in der Grundstimmung identifizieren und fallen diese mit markanten Ereignissen zusammen?'* sind die Methoden als hilfreiche Indikatoren geeignet und bilden die Grundlage weiterführender Analysen. Eine Veränderung in der Grundstimmung, die in den Ergebnissen aus der Erkennung von Hassrede erkennbar ist, ist der deutliche Anstieg an Negativität im Bundestag und in Thüringen nach dem Einzug der AfD. Die Untersuchung der vierten Fragestellung (d): *'Lässt sich ein negativer Bias gegenüber den Regionen des ehemaligen Ostens Deutschlands feststellen?'* zeigt, dass die ausgewählten Methoden nicht geeignet sind, um dies zu thematisieren.

Das Ziel, politische Debatten im Kontext sozialwissenschaftlicher Fragestellungen, mit ausgewählten Methoden aus dem Bereich von NLP zu analysieren, konnte erreicht werden. Die Untersuchung zeigt, dass die Einbeziehung der Erkennung von Hassrede einen Mehrwert für die Analyse des Stimmungsbildes in politischen Gremien darstellt. Dieser methodische Ansatz sollte verstärkter berücksichtigt werden. Weiterhin ist das Ziel erreicht herauszuarbeiten, inwieweit die Methoden sich für die jeweiligen Fragestellungen eignen und an welchen Stellen deren Grenzen liegen.

Der Einsatz transformer-basierter Methoden und LLMs entspricht dem Forschungsstand in der Sentiment-Analyse und Erkennung von Hassrede. Dennoch kommen in der Analyse deutschsprachiger Parlamentsdaten weiterhin Verfahren des überwachten maschinellen Lernens zum Einsatz, wie die aktuelle Arbeit von (Pätz et al., 2025) zeigt. Die Ergebnisse der vorliegenden Untersuchung sollen dazu anregen, transformer-basierte auf Parlamentsdaten weitertrainierte Modelle und LLMs stärker in der Analyse zu berücksichtigen. Bereits die Studie von (Widmann & Wich, 2022) verdeutlicht diesen Ansatz. In dieser erreicht das von ihnen weitertrainierte deutschsprachige ELECTRA Modell die beste Modellleistung. Die durchgeführte Evaluation der untersuchten Sentiment-Analyse Methoden bestätigt diese Erkenntnis. Allerdings wird in der vorliegenden Arbeit kein Vergleich zu einem überwachten maschinellen Lernverfahren durchgeführt, dies stellt zudem eine methodische Begrenzung der Untersuchungsmethode dar.

Während der Implementierung wurden weitere Grenzen deutlich. Dazu zählt insbesondere die Größe des Datensatzes. Die Anwendung des ausgewählten LLMs auf die gesamten Analysedaten war aufgrund der damit verbundenen langen Antwortzeiten der API nicht möglich. Zudem musste die Klassifizierung der Analysedaten für jedes Gremium getrennt durchgeführt werden. Die Daten des Deutschen Bundestages werden darüber hinaus in Teildatensätze aufgeteilt. Ein weiterer Punkt stellt die daraus resultierende Dauer während der Klassifizierung der Analysedaten sowie die Trainingszeit des Fine-Tuning Modells dar. Die Dauer des Trainings ist in (Abercrombie & Batista-Navarro, 2020b) ein Punkt, der gegen die weitere Untersuchung von transformer-basierten Methoden spricht. Die zeitliche Komponente sollte daher in die Entscheidung einbezogen werden.

Basierend auf der Analyse der Parlamentsdaten über den Zeitraum von 2002 bis 2022 sind die in der Einleitung erwähnten Äußerungen von Bärbel Bas und Julia Klöckner erneut zu betrachten. Beide Politikerinnen mahnten in ihren Interviews den Umgangston im Deutschen Bundestag an (Kathe, 2025; tagesschau.de, 2024). Diese empirisch begründeten Vermutungen

lassen sich mit Hilfe der Ergebnisse der praktischen Arbeit datenbasiert einordnen. Eine eindeutige Zunahme von Aggressivität und Negativität konnte für das Plenum nicht festgestellt werden. Die Ergebnisse der Sentiment-Analyse deuten vielmehr auf eine leichte Abnahme bis 2021 hin.

Die deutliche Zunahme, die in den Werten zur Erkennung von Hassrede nach der Bundestagswahl 2017 auftritt, stützt hingegen die Einschätzungen von Bas und Klöckner. Da beide Interviews zeitlich nach dem Untersuchungszeitraum liegen, könnte eine Erweiterung bis in die aktuelle Wahlperiode in weiterführenden Forschungen zu einer präziseren Bewertung führen. Zusätzliche Ansatzpunkte die Forschung an Parlamentsdaten auszubauen sind im Folgenden Abschnitt beschrieben.

7.2 Ausblick

Aus der durchgeführten praktischen Arbeit konnten verschiedene weiterführende Ansätze abgeleitet werden. Diese unterteilen sich in zwei Themenschwerpunkte. Zum einen in Möglichkeiten, zur Verbesserung des entwickelten Fine-Tuning Modells, und zum anderen in weiterführende Forschungsrichtungen auf Grundlage der Parlamentsdaten.

Einen Verbesserungsansatz bildet der Datensatz, der für das Training des deutschsprachigen ELECTRA Modells genutzt wird. Der Datensatz könnte mit positiven Beispielen angereichert werden. Dadurch wird es ermöglicht, mit dem Fine-Tuning Modell auch die positive Klasse repräsentieren zu können. Darauf aufbauend lässt sich untersuchen, ob Entwicklungen zu steigender Positivität erkennbar sind.

Die Analyse von Parlamentsdaten bietet eine Vielzahl an Möglichkeiten für weiterführende Forschungsprojekte. Ein Ansatz könnte darin bestehen, die Zwischenrufe und Unterbrechungen in den Kontext des dafür verantwortlichen Redebeitrags zu setzen. Ziel wäre es dabei, die Intention und Stimmung beispielsweise einer Beifallsbekundung oder eines Lachens besser einschätzen zu können.

Zudem bietet die Visualisierung einen weiteren Ansatzpunkt, die Analyse politischer Parlamentsdaten zu bereichern. In der vorliegenden Arbeit werden die Jahres- und Quartalsebene betrachtet und einfache Methoden der Zeitreihenzerlegung eingesetzt. Für die Analyse könnten hierbei neuere Visualisierungstechniken und Methodenansätze aus dem Bereich der Zeitreihenzerlegung und Erkennung von Anomalien eingesetzt werden.

Weiterhin könnte die Erkennung von Hassrede auf parlamentarischen Debatten mit methodischen Ansätzen ausgebaut werden. Beispielsweise könnte ein Fine-Tuning Modell analog zum beschriebenen Ansatz in der Sentiment-Analyse trainiert werden. Zugleich könnte für eine Analyse populisitscher Rhetorik das Modell von (Erhard et al., 2024) auf die Daten angewendet werden.

Ein Punkt von Interesse in weiterführenden Untersuchungen stellt die Differenzierung der Reden in die jeweiligen Parteien dar. Dadurch könnte die Analyse insbesondere der ersten Fragestellung (a): *'Sind die Debatten in den deutschen Gremien aggressiver und negativer geworden?'* auf einer feingranularen Ebene durchgeführt werden und würde das Stimmungsbild differenzierter darstellen.

In Bezug auf die sozialwissenschaftliche Fragestellung (d): *'Lässt sich ein negativer Bias gegenüber den Regionen des ehemaligen Ostens Deutschlands feststellen?'* sollten weiterführende Projekte den Fokus auf die Aspektebene in der Sentiment-Analyse legen. Ein vertiefender Ansatz zu Fragestellung (b): *'Lassen sich Wendepunkte in der Grundstimmung identifizieren und fallen diese mit markanten Ereignissen zusammen?'* besteht in der genaueren Untersuchung der Redebeiträge an den identifizierten auffälligen Punkten. Hierfür könnten die gespeicherten Daten der Ausreißer sowie das Filtern der Daten auf die identifizierten An- und Abstiegsphasen die Grundlage bilden. Auf diese Daten könnten beispielsweise Methoden aus dem Bereich des Topic Modelings angewendet werden.

Literatur

Abercrombie, G., & Batista-Navarro, R. (2020a). Sentiment and position-taking analysis of parliamentary debates: a systematic literature review. *Journal of Computational Social Science*, *3*(1), 245–270. https://doi.org/10.1007/s42001-019-00060-w

Abercrombie, G., & Batista-Navarro, R. (2020b). ParlVote: A Corpus for Sentiment Analysis of Political Debates. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (Hrsg.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (S. 5073–5078). European Language Resources Association. https://aclanthology.org/2020.lrec-1.624/

Abrami, G., Bagci, M., & Mehler, A. (2024). German Parliamentary Corpus (GerParCor) Reloaded. In N. Calzolari, M. Kan, V. Hoste, A. Lenci, S. Sakti & N. Xue (Hrsg.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (S. 7707–7716). ELRA; ICCL.

Agarwal, P., Hawkins, O., Amaxopoulou, M., Dempsey, N., Sastry, N., & Wood, E. (2021). Hate Speech in Political Discourse: A Case Study of UK MPs on Twitter. *Proceedings of the 32st ACM Conference on Hypertext and Social Media*, 5–16. https://doi.org/10.1145/3465336.3475113

Ahmed, S. K. (2024). How to choose a sampling technique and determine sample size for research: A simplified guide for researchers. *Oral Oncology Reports*, *12*, 100662. https://doi.org/10.1016/j.oor.2024.100662

Albladi, A., Islam, M., Das, A., Bigonah, M., Zhang, Z., Jamshidi, F., Rahgouy, M., Raychawdhary, N., Marghitu, D., & Seals, C. (2025). Hate Speech Detection Using Large Language Models: A Comprehensive Review. *IEEE Access*, *13*, 20871–20892. https://doi.org/10.1109/access.2025.3532397

Alkomah, F., & Ma, X. (2022). A Literature Review of Textual Hate Speech Detection Methods and Datasets. *Information*, *13*(6), 273. https://doi.org/10.3390/info13060273

Aluru, S. S., Mathew, B., Saha, P., & Mukherjee, A. (2020). *Deep Learning Models for Multilingual Hate Speech Detection* [arXiv preprint]. arXiv: 2004.06465 [cs.SI]. https://doi.org/10.48550/ARXIV.2004.06465

Anjum & Katarya, R. (2023). Hate speech, toxicity detection in online social media: a recent survey of state of the art and opportunities. *International Journal of Information Security*, *23*(1), 577–608. https://doi.org/10.1007/s10207-023-00755-2

Antypas, D., Preece, A., & Camacho-Collados, J. (2022). Negativity Spreads Faster: A Large-Scale Multilingual Twitter Analysis on the Role of Sentiment in Political Communication. *Online Social Networks and Media, 2023, Volume 33 Online Social Networks and Media*, *33*, 100242. https://doi.org/10.1016/j.osnem.2023.100242

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). *Layer Normalization* [arXiv preprint]. arXiv: 1607.06450 [stat.ML]. https://doi.org/10.48550/ARXIV.1607.06450

Barbieri, F., Anke, L. E., & Camacho-Collados, J. (2022). XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk & S. Piperidis (Hrsg.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (S. 258–266). European Language Resources Association.

Bashiri, H., & Naderi, H. (2024). Comprehensive review and comparative analysis of transformer models in sentiment analysis. *Knowledge and Information Systems*, *66*(12), 7305–7361. https://doi.org/10.1007/s10115-024-02214-3

BERT community. (n. d.). google-bert/bert-base-german-cased. Verfügbar 13. Oktober 2025 unter https://huggingface.co/google-bert/bert-base-german-cased

Blackwell, R. E., Barry, J., & Cohn, A. G. (2024). *Towards Reproducible LLM Evaluation: Quantifying Uncertainty in LLM Benchmark Scores* [arXiv preprint]. arXiv: 2410.03492 [cs.CL]. https://doi.org/10.48550/ARXIV.2410.03492

Blázquez-García, A., Conde, A., Mori, U., & Lozano, J. A. (2021). A Review on Outlier/Anomaly Detection in Time Series Data. *ACM Computing Surveys*, *54*(3), 1–33. https://doi.org/10.1145/3444690

BoTox – Bot- und Kontexterkennung im Umfeld von Hasskommentaren. (2025). Verfügbar 28. Oktober 2025 unter https://botox.h-da.de/

Brockwell, P. J., & Davis, R. A. (2002). *Introduction to Time Series and Forecasting* (2. Aufl.). Springer New York, NY. https://doi.org/https://doi.org/10.1007/b97391

Bu, C., Liu, Y., Huang, M., Shao, J., Ji, S., Luo, W., & Wu, X. (2024). Layer-Wise Learning Rate Optimization for Task-Dependent Fine-Tuning of Pre-Trained Models: An Evolutionary Approach. *ACM Transactions on Evolutionary Learning and Optimization*, *4*(4), 1–23. https://doi.org/10.1145/3689827

Bundesministerium für Gesundheit. (2023). Coronavirus-Pandemie: Was geschah wann? Verfügbar 31. Oktober 2025 unter https://www.bundesgesundheitsministerium.de/coronavirus/chronik-coronavirus.html

Bundeswahlleiterin. (2025a). Ergebnisse früherer Bundestagswahlen. Verfügbar 24. Juli 2025 unter https://www.bundeswahlleiterin.de/dam/jcr/397735e3-0585-46f6-a0b5-2c60c5b83de6/btw_ab49_gesamt.pdf

Bundeswahlleiterin. (2025b). Ergebnisse früherer Landtagswahlen. Verfügbar 24. Juli 2025 unter https://www.bundeswahlleiterin.de/dam/jcr/a333e523-0717-42ad-a772-d5ad7e7e97cc/ltw_erg_gesamt.pdf

Cardiff NLP. (n. d.). XLM-T-Sent-Politics. Verfügbar 1. August 2025 unter https://huggingface.co/cardiffnlp/xlm-twitter-politics-sentiment

Chinchor, N. (1992). MUC-4 Evaluation Metrics. *Fourth Message Understanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992*. https://aclanthology.org/M92-1002/

Christodoulou, C. (n. d. a). christinacdl/XLM_RoBERTa-Multilingual-Hate-Speech-Detection-New. Verfügbar 13. Oktober 2025 unter https://huggingface.co/christinacdl/XLM_RoBERTa-Multilingual-Hate-Speech-Detection-New

Christodoulou, C. (n. d. b). christinacdl/XLM_RoBERTa-Multilingual-OpusMT-Offensive-Language-Detection. Verfügbar 13. Oktober 2025 unter https://huggingface.co/christinacdl/XLM_RoBERTa-Multilingual-OpusMT-Offensive-Language-Detection

Christodoulou, C. (n. d. c). christinacdl/XLM_RoBERTa-Offensive-Language-Detection-8-langs-new. Verfügbar 13. Oktober 2025 unter https://huggingface.co/christinacdl/XLM_RoBERTa-Offensive-Language-Detection-8-langs-new

Cieliebak, M., Deriu, J. M., Egger, D., & Uzdilli, F. (2017). A twitter corpus and benchmark resources for german sentiment analysis. *5th International Workshop on Natural Language Processing for Social Media, Boston MA, USA, 11 December 2017*, 45–51.

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). *ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators* [arXiv preprint]. arXiv: 2003.10555 [cs.CL]. https://doi.org/10.48550/ARXIV.2003.10555

Clematide, S., Gindl, S., Klenner, M., Petrakis, S., Remus, R., Ruppenhofer, J., Waltinger, U., & Wiegand, M. (2012). MLSA — A Multi-layered Reference Corpus for German Sentiment Analysis. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. http://www.lrec-conf.org/proceedings/lrec2012/index.html

Clematide, S., & Klenner, M. (2010). Evaluation and extension of a polarity lexicon for German. In A. Montoyo, P. Martínez-Barco, A. Balahur & E. Boldrini (Hrsg.), *Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA)* (S. 7–13). https://doi.org/10.5167/UZH-45506

Cleveland, R. B., Cleveland, W. S., McRae, J. E., Terpenning, I., et al. (1990). STL: A seasonal-trend decomposition. *J. off. Stat*, *6*(1), 3–73.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/2020.acl-main.747

De Oliveira, A. B., de Souza Baptista, C., Firmino, A. A., & De Paiva, A. C. (2024). A Large Language Model Approach to Detect Hate Speech in Political Discourse Using Multiple Language Corpora. *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, 1461–1468. https://doi.org/10.1145/3605098.3635964

Decker, F. (2022). Etappen der Parteigeschichte der AfD. Verfügbar 31. Oktober 2025 unter https://www.bpb.de/themen/parteien/parteien-in-deutschland/afd/273130/etappen-der-parteigeschichte-der-afd/#node-content-title-3

deepset. (n. d.). deepset/bert-base-german-cased-hatespeech-GermEval18Coarse. Verfügbar 13. Oktober 2025 unter https://huggingface.co/deepset/bert-base-german-cased-hatespeech-GermEval18Coarse

Demus, C., Pitz, J., Schütz, M., Probol, N., Siegel, M., & Labudde, D. (2022). DeTox: A Comprehensive Dataset for German Offensive Language and Conversation Analysis. In K. Narang, A. Mostafazadeh Davani, L. Mathias, B. Vidgen & Z. Talat (Hrsg.), *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)* (S. 143–153). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.woah-1.14

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran & T. Solorio (Hrsg.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume*

1 (Long and Short Papers) (S. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/n19-1423

Dohmen, C. (2024). Die Finanzkrise von 2007/2008 und ihre Folgen. Verfügbar 31. Oktober 2025 unter https://www.bpb.de/themen/wirtschaft/finanzwirtschaft/524122/die-finanzkrise-von-2007-2008-und-ihre-folgen/

Doosthosseini, A., Decker, J., Nolte, H., & Kunkel, J. (2025). SAIA: A Seamless Slurm-Native Solution for HPC-Based Services. https://doi.org/10.21203/rs.3.rs-6648693/v1

Dudek, G. (2023). STD: A Seasonal-Trend-Dispersion Decomposition of Time Series. *IEEE Transactions on Knowledge and Data Engineering*, *35*(10), 10339–10350. https://doi.org/10.1109/tkde.2023.3268125

Egger, R., & Gokce, E. (2022). Natural Language Processing (NLP): An Introduction: Making Sense of Textual Data. In *Applied Data Science in Tourism* (S. 307–334). Springer International Publishing. https://doi.org/10.1007/978-3-030-88389-8_15

Emerson, G., & Declerck, T. (2014). SentiMerge: Combining sentiment lexicons in a Bayesian framework. *Proceedings of workshop on lexical and grammatical resources for language processing*, 30–38.

Erhard, L., Hanke, S., Remer, U., Falenska, A., & Heiberger, R. H. (2024). PopBERT. Detecting Populism and Its Host Ideologies in the German Bundestag. *Political Analysis*, *33*(1), 1–17. https://doi.org/10.1017/pan.2024.12

Eslava, A. (2023). Outlier Detection Techniques for Time Series. Verfügbar 15. September 2025 unter https://medium.com/@alex.eslava96/outlier-detection-techniques-for-time-series-9868db2875c2

Facebook AI community. (n. d.). FacebookAI/xlm-roberta-large. Verfügbar 13. Oktober 2025 unter https://huggingface.co/deepset/bert-base-german-cased-hatespeech-GermEval18Coarse

Fehle, J., Schmidt, T., & Wolff, C. (2021). Lexicon-based Sentiment Analysis in German: Systematic Evaluation of Resources and Preprocessing Techniques". In K. Evang, L. Kallmeyer, R. Osswald, J. Waszczuk & T. Zesch (Hrsg.), *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)* (S. 86–103). KONVENS 2021 Organizers. https://aclanthology.org/2021.konvens-1.8/

Fortuna, P., Dominguez, M., Wanner, L., & Talat, Z. (2022). Directions for NLP Practices Applied to Online Hate Speech Detection. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 11794–11805. https://doi.org/10.18653/v1/2022.emnlp-main.809

Gage, P. (1994). A new algorithm for data compression. *C Users Journal*, *12*(2), 23–38.

Gandhi, A., Ahir, P., Adhvaryu, K., Shah, P., Lohiya, R., Cambria, E., Poria, S., & Hussain, A. (2024). Hate speech detection: A comprehensive review of recent works. *Expert Systems*, *41*(8). https://doi.org/10.1111/exsy.13562

geeksforgeeks. (2025). What is Data Sampling - Types, Importance, Best Practices. Verfügbar 24. Juli 2025 unter https://www.geeksforgeeks.org/data-analysis/what-is-data-sampling-types-importance-best-practices/

Giachanou, A., & Crestani, F. (2016). Tracking Sentiment by Time Series Analysis. *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 1037–1040. https://doi.org/10.1145/2911451.2914702

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning* (1. Aufl.). MIT Press.

Guhr, O., Schumann, A.-K., Bahrmann, F., & Böhme, H. J. (2020). Training a Broad-Coverage German Sentiment Classification Model for Dialog Systems. In N. Calzolari, F. Béchet,

P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (Hrsg.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (S. 1627–1632). European Language Resources Association. https://aclanthology.org/2020.lrec-1.202/

GWDG Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen. (n. d. a). Available Models. Verfügbar 30. Juli 2025 unter https://docs.hpc.gwdg.de/acknowledgements/index.html

GWDG Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen. (n. d. b). HOME. Verfügbar 30. Juli 2025 unter https://docs.hpc.gwdg.de/index.html

Hartford, E., Atkins, L., Neto, F. F., & Golchinfar, D. (2024). *Spectrum: Targeted Training on Signal to Noise Ratio* [arXiv preprint]. arXiv: 2406.06623 [cs.LG]. https://doi.org/10.48550/ARXIV.2406.06623

Haselmayer, M., & Jenny, M. (2017). Sentiment analysis of political communication: combining a dictionary approach with crowdcoding. *Quality & Quantity*, *51*(6), 2623–2646. https://doi.org/10.1007/s11135-016-0412-4

Haselmayer, M., & Jenny, M. (2020a). 10274_mr_en_v1_0.pdf. https://doi.org/10.11587/7PFLIU/RURINK

Haselmayer, M., & Jenny, M. (2020b). Training Data for German Sentiment Analysis of Political Communication (SUF edition) [V1]. https://doi.org/10.11587/EOPCOB

Hate-ALERT. (n. d.). Hate-speech-CNERG/dehatebert-mono-german. Verfügbar 13. Oktober 2025 unter https://huggingface.co/Hate-speech-CNERG/dehatebert-mono-german

Hawkins, D. M. (1980). Introduction. In *Identification of Outliers* (S. 1–12). Springer Netherlands. https://doi.org/10.1007/978-94-015-3994-4_1

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition CVPR*, 770–778. https://doi.org/10.1109/cvpr.2016.90

He, P., Liu, X., Gao, J., & Chen, W. (2020). *DeBERTa: Decoding-enhanced BERT with Disentangled Attention* [arXiv preprint]. arXiv: 2006.03654 [cs.CL]. https://doi.org/10.48550/ARXIV.2006.03654

Henze, N. (2013). Grundbegriffe der deskriptiven Statistik. In *Stochastik für Einsteiger: Eine Einführung in die faszinierende Welt des Zufalls* (S. 20–36). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-03077-3_5

Herbert, U., & Schönhagen, J. (2020). Vor dem 5. September. Die 'Flüchtlingskrise' 2015 im historischen Kontext. Verfügbar 31. Oktober 2025 unter https://www.bpb.de/shop/zeitschriften/apuz/312832/vor-dem-5-september/

Hinojosa Lee, M. C., Braet, J., & Springael, J. (2024). Performance Metrics for Multilabel Emotion Classification: Comparing Micro, Macro, and Weighted F1-Scores. *Applied Sciences*, *14*(21). https://doi.org/10.3390/app14219863

Honnibal, M., Montani, I., & Van Landeghem, A., S. amd Boyd. (2020). spaCy: Industrial-strength Natural Language Processing in Python. https://doi.org/10.5281/zenodo.1212303

Hou, Y., & Huang, J. (2025). Natural language processing for social science research: A comprehensive review. *Chinese Journal of Sociology*, *11*(1), 121–157. https://doi.org/10.1177/2057150X241306780

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-Efficient Transfer Learning for NLP. In K. Chaudhuri & R. Salakhutdinov (Hrsg.), *Proceedings of the 36th International*

Conference on Machine Learning (S. 2790–2799, Bd. 97). PMLR. https://proceedings.mlr.press/v97/houlsby19a.html

Hugging Face. (n. d. a). Fine-tune a pretrained model. Verfügbar 1. August 2025 unter https://huggingface.co/docs/transformers/v4.18.0/en/training

Hugging Face. (n. d. b). Padding and truncation. Verfügbar 1. August 2025 unter https://huggingface.co/docs/transformers/pad_truncation

Hugging Face. (n. d. c). Pipelines. Verfügbar 21. Juli 2025 unter https://huggingface.co/docs/transformers/v4.17.0/en/main_classes/pipelines#transformers.pipeline

Hugging Face. (n. d. d). Trainer. Verfügbar 1. August 2025 unter https://huggingface.co/docs/transformers/main_classes/trainer

Hyndman, R. J., & Athanasopoulos, G. (2014). *Forecasting: Principles and Practice*. OTexts.

Jin, H., Wei, W., Wang, X., Zhang, W., & Wu, Y. (2023). Rethinking Learning Rate Tuning in the Era of Large Language Models. *2023 IEEE 5th International Conference on Cognitive Machine Intelligence (CogMI)*, 112–121. https://doi.org/10.1109/cogmi58952.2023.00025

Kannan, S., Gurusamy, V., Vijayarani, S., Ilamathi, J., Nithya, M., Kannan, S., & Gurusamy, V. (2014). Preprocessing techniques for text mining. *International Journal of Computer Science & Communication Networks*, *5*(1), 7–16.

Kathe, S. (2025). War noch nie so polarisiert wie heute: Klöckner schlägt wegen Stimmung im Bundestag Alarm. Verfügbar 25. November 2025 unter https://www.merkur.de/politik/war-noch-nie-so-polarisiert-wie-heute-kloeckner-schlaegt-wegen-stimmung-im-bundestag-alarm-zr-93937041.html

Kawintiranon, K., & Singh, L. (2022). PoliBERTweet: A Pre-trained Language Model for Analyzing Political Content on Twitter. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk & S. Piperidis (Hrsg.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (S. 7360–7367). European Language Resources Association.

Kirilenko, A. P., Wang, L., & Stepchenkova, S. O. (2022). Sentiment Analysis: Gaging Opinions of Large Groups. In *Applied Data Science in Tourism* (S. 363–374). Springer International Publishing. https://doi.org/10.1007/978-3-030-88389-8_17

Lange, K.-R., & Jentsch, C. (2023a). SpeakGer: A meta-data enriched speech corpus of German state and federal parliaments. In C. Klamm, G. Lapesa, V. Gold, T. Gessler & S. P. Ponzetto (Hrsg.), *Proceedings of the 3rd Workshop on Computational Linguistics for the Political and Social Sciences* (S. 19–28). Association for Computational Linguistics. https://aclanthology.org/2023.cpss-1.3/

Lange, K.-R., & Jentsch, C. (2023b). SpeakGer: A meta-data enriched speech corpus of German state and federal parliaments. Verfügbar 4. Juli 2025 unter https://berd-platform.de/records/g3225-rba63

Lange, K.-R., Rieger, J., & Jentsch, C. (2024). Lex2Sent: A bagging approach to unsupervised sentiment analysis. In P. H. Luz de Araujo, A. Baumann, D. Gromann, B. Krenn, B. Roth & M. Wiegand (Hrsg.), *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)* (S. 281–291). Association for Computational Linguistics. https://aclanthology.org/2024.konvens-main.28/

Le, Q., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. In E. P. Xing & T. Jebara (Hrsg.), *Proceedings of the 31st International Conference on Machine Learning* (S. 1188–1196, Bd. 32). PMLR. https://proceedings.mlr.press/v32/le14.html

Lehtosalo, S., & Nerbonne, J. (2020). Detecting emotional polarity in Finnish parliamentary proceedings. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 90–100.

Li, L., Li, J., Chen, C., Gui, F., Yang, H., Yu, C., Wang, Z., Cai, J., Zhou, J. A., Shen, B., Qian, A., Chen, W., Xue, Z., Sun, L., He, L., Chen, H., Ding, K., Du, Z., Mu, F., … Dong, Y. (2024). Political-LLM: Large Language Models in Political Science [arXiv preprint]. https://doi.org/10.48550/ARXIV.2412.06864

Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool.

Liu, B. (2015). *Sentiment analysis : mining opinions, sentiments, and emotions*. Cambridge Univ. Press.

Liu, B. (2017). Many Facets of Sentiment Analysis. In *A Practical Guide to Sentiment Analysis* (S. 11–39). Springer International Publishing. https://doi.org/10.1007/978-3-319-55394-8_2

Loria, S., & Mitwirkende. (n. d.). Tutorial: Quickstart. Verfügbar 17. Juli 2025 unter https://textblob.readthedocs.io/en/dev/quickstart.html

Lossio-Ventura, J. A., Weger, R., Lee, A. Y., Guinee, E. P., Chung, J., Atlas, L., Linos, E., & Pereira, F. (2024). A Comparison of ChatGPT and Fine-Tuned Open Pre-Trained Transformers (OPT) Against Widely Used Sentiment Analysis Tools: Sentiment Analysis of COVID-19 Survey Data. *JMIR Mental Health*, *11*, e50150. https://doi.org/10.2196/50150

MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions (M. Huang, Hrsg.). *PLOS ONE*, *14*(8), e0221152. https://doi.org/10.1371/journal.pone.0221152

Malik, J. S., Qiao, H., Pang, G., & van den Hengel, A. (2024). Deep learning for hate speech detection: a comparative study. *International Journal of Data Science and Analytics*, *20*(4), 3053–3068. https://doi.org/10.1007/s41060-024-00650-6

Mandl, T., Modha, S., Shahi, G. K., Jaiswal, A. K., Nandini, D., Daksh Patel, P. M., & Schäfer, J. (2023). *Accessed: Dec*, *16*.

Manning, C., & Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

May, P., & Reißel, P. (2020). German Electra Uncased. Verfügbar 1. August 2025 unter https://huggingface.co/german-nlp-group/electra-base-german-uncased

Meta Llama. (n. d.). meta-llama/Llama-3.1-70B-Instruct. Verfügbar 7. August 2025 unter https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct

Moldovan, D. (2025). A majority voting framework for reliable sentiment analysis of product reviews. *PeerJ Computer Science*, *11*, e2738. https://doi.org/10.7717/peerj-cs.2738

Montani, I., Honnibal, M., Boyd, A., Van Landeghem, S., & Peters, H. (2023). explosion/spaCy: v3.7.2: Fixes for APIs and requirements. https://doi.org/10.5281/ZENODO.1212303

Naglik, I., & Lango, M. (2025). Fine-Tuning Fine-Tuned Models: Towards a Practical Methodology for Sentiment Analysis with Small In-Domain Supervised Dataset. In *Neural Information Processing* (S. 1–16). Springer Nature Singapore. https://doi.org/10.1007/978-981-96-7005-5_1

Németh, R. (2022). A scoping review on the use of natural language processing in research on political polarization: trends and research prospects. *Journal of Computational Social Science*, *6*(1), 289–313. https://doi.org/10.1007/s42001-022-00196-2

Ortiz, J. (n. d.). jorgeortizv/BERT-hateSpeechRecognition-German. Verfügbar 13. Oktober 2025 unter https://huggingface.co/jorgeortizv/BERT-hateSpeechRecognition-German

Ouyang, S., Zhang, J. M., Harman, M., & Wang, M. (2025). An Empirical Study of the Non-Determinism of ChatGPT in Code Generation. *ACM Transactions on Software Engineering and Methodology*, *34*(2), 1–28. https://doi.org/10.1145/3697010

Parihar, A. S., Thapa, S., & Mishra, S. (2021). Hate Speech Detection Using Natural Language Processing: Applications and Challenges. *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, 1302–1308. https://doi.org/10.1109/icoei51242.2021.9452882

Pascual, F. (2022). Getting Started with Sentiment Analysis using Python. Verfügbar 19. Juli 2025 unter https://huggingface.co/blog/sentiment-analysis-python

Pätz, L., Beyer, M., Späth, J., Bohlen, L., Zschech, P., Kraus, M., & Rosenberger, J. (2025). *Analyzing German Parliamentary Speeches: A Machine Learning Approach for Topic and Sentiment Classification* [arXiv preprint]. arXiv: 2508.03181 [cs.CL]. https://doi.org/10.48550/ARXIV.2508.03181

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2020). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, *55*(2), 477–523. https://doi.org/10.1007/s10579-020-09502-8

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. https://doi.org/10.18653/v1/2020.acl-demos.14

Raiaan, M. A. K., Mukta, M. S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M. J., Ahmad, J., Ali, M. E., & Azam, S. (2024). A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. *IEEE Access*, *12*, 26839–26874. https://doi.org/10.1109/access.2024.3365742

Rathi, D. (2024). Handling Imbalanced Data: Key Techniques for Better Machine Learning. Verfügbar 15. September 2025 unter https://medium.com/@dakshrathi/handling-imbalanced-data-key-techniques-for-better-machine-learning-6e33b466f8b7

Rauh, C. (2018a). Replication Data for: Validating a sentiment dictionary for German political language. https://doi.org/10.7910/DVN/BKBXWD

Rauh, C. (2018b). Validating a sentiment dictionary for German political language—a workbench note. *Journal of Information Technology & Politics*, *15*(4), 319–343. https://doi.org/10.1080/19331681.2018.1485608

Remus, R., Quasthoff, U., & Heyer, G. (2010). SentiWS-A Publicly Available German-language Resource for Sentiment Analysis. *Proceedings of the 7th International Language Resources and Evaluation (LREC)*, 1168–1171.

Rheault, L. (2016). remove-decorum-words.sh. Verfügbar 15. Juli 2025 unter https://github.com/lrheault/emotion/blob/master/remove-decorum-words.sh

Rheault, L., Beelen, K., Cochrane, C., & Hirst, G. (2016). Measuring Emotion in Parliamentary Debates with Automated Textual Analysis (J. Najbauer, Hrsg.). *PLOS ONE*, *11*(12), 1–18. https://doi.org/10.1371/journal.pone.0168843

Rini, R., Utami, E., & Hartanto, A. D. (2020). Systematic Literature Review Of Hate Speech Detection With Text Mining. *2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS)*, 1–6. https://doi.org/10.1109/icoris50180.2020.9320755

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., & Sedlmair, M. (2018). More than Bags of Words: Sentiment Analysis with Word Embeddings. *Communication Methods and Measures*, *12*(2–3), 140–157. https://doi.org/10.1080/19312458.2018.1455817

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter* [arXiv preprint]. arXiv: 1910.01108 [cs.CL]. https://doi.org/10.48550/ARXIV.1910.01108

Sathyanarayanan, S., & Tantri, B. R. (2024). Confusion matrix-based performance evaluation metrics. *African Journal of Biomedical Research*, *27*(45), 4023–4031.

Sazzed, S., & Jayarathna, S. (2021). SSentiA: A Self-supervised Sentiment Analyzer for classification from unlabeled data. *Machine Learning with Applications*, *4*, 100026. https://doi.org/10.1016/j.mlwa.2021.100026

Schmidt, A., & Wiegand, M. (2017). A Survey on Hate Speech Detection using Natural Language Processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 1–10. https://doi.org/10.18653/v1/w17-1101

Schmidt, T., Fehle, J., Weissenbacher, M., Richter, J., Gottschalk, P., & Wolff, C. (2022). Sentiment Analysis on Twitter for the Major German Parties during the 2021 German Federal Election. In R. Schaefer, X. Bai, M. Stede & T. Zesch (Hrsg.), *Proceedings of the 18th Conference on Natural Language Processing KONVENS 2022* (S. 74–87). KONVENS 2022 Organizers.

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1715–1725.

Shazeer, N. (2020). *GLU Variants Improve Transformer* [arXiv preprint]. arXiv: 2002.05202 [cs.LG]. https://doi.org/10.48550/ARXIV.2002.05202

Sidarenka, U. (2016). PotTS: The Potsdam Twitter Sentiment Corpus. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (Hrsg.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (S. 1133–1141). European Language Resources Association (ELRA). https://aclanthology.org/L16-1181/

Siegel, M. (n. d.). nlp_de_basic/sentiment_words.py [GitHub].

Siegel, M., & Alexa, M. (2020). *Sentiment-Analyse deutschsprachiger Meinungsäußerungen: Grundlagen, Methoden und praktische Umsetzung*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-29699-5

Solovev, K., & Pröllochs, N. (2022). Hate Speech in the Political Discourse on Social Media: Disparities Across Parties, Gender, and Ethnicity. *Proceedings of the ACM Web Conference 2022*, 3656–3661. https://doi.org/10.1145/3485447.3512261

spaCy Doku. (n. d.). spaCy Industrial-strength Natural language processing in Python. Verfügbar 7. Juli 2025 unter https://spacy.io/

spaCy Usage Documentation. (n. d.). Language Processing Pipelines. Verfügbar 19. Juli 2025 unter https://spacy.io/usage/processing-pipelines

Spelmen, V. S., & Porkodi, R. (2018). A Review on Handling Imbalanced Data. *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, 1–11. https://doi.org/10.1109/icctct.2018.8551020

Srinivasa-Desikan, B. (2018). *Natural Language Processing and Computational Linguistics*. https://learning.oreilly.com/library/view/-/9781788838535

Stanford NLP Group. (n. d.). Stanza: A Python NLP Library for Many Human Languages README.md. Verfügbar 7. August 2025 unter https://github.com/stanfordnlp/stanza/blob/main/README.md

Stanford NLP Group. (2020). Pipeline and Processors. Verfügbar 7. August 2025 unter https://stanfordnlp.github.io/stanza/pipeline.html

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., & Liu, Y. (2024). RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, *568*, 127063. https://doi.org/10.1016/j.neucom.2023.127063

Suandi, F., Anam, M. K., Firdaus, M. B., Fadli, S., Lathifah, L., Yumami, E., Saleh, A., & Hasibuan, A. Z. (2024). Enhancing Sentiment Analysis Performance Using SMOTE and Majority Voting in Machine Learning Algorithms. In *Proceedings of the 7th International Conference on Applied Engineering (ICAE 2024)* (S. 126–138). Atlantis Press International BV. https://doi.org/10.2991/978-94-6463-620-8_10

Tabassum, A., & Patil, R. R. (2020). A survey on text pre-processing & feature extraction techniques in natural language processing. *International Research Journal of Engineering and Technology (IRJET)*, *7*(06), 4864–4867.

tabularisai, Gyamfi, S., Borisov, V., & Schreiber, R. H. (2025). multilingual-sentiment-analysis. https://doi.org/10.57967/HF/5968

tagesschau.de. (2008). Staat garantiert für private Spareinlagen. Verfügbar 31. Oktober 2025 unter https://www.tagesschau.de/wirtschaft/spareinlagen-ts-100.html

tagesschau.de. (2024). Bas zu Umgangston im Parlament Atmosphäre im Bundestag spürbar verändert. Verfügbar 20. April 2025 unter https://www.tagesschau.de/inland/baerbel-bas-debattenkultur-100.html

Takamura, H., Inui, T., & Okumura, M. (2005). Extracting semantic orientations of words using spin model. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, 133–140.

Tan, K. L., Lee, C. P., & Lim, K. M. (2023). A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research. *Applied Sciences*, *13*(7), 4550. https://doi.org/10.3390/app13074550

Tan, K. L., Lee, C. P., Lim, K. M., & Anbananthen, K. S. M. (2022). Sentiment Analysis With Ensemble Hybrid Deep Learning Model. *IEEE Access*, *10*, 103694–103704. https://doi.org/10.1109/access.2022.3210182

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models [arXiv preprint]. https://doi.org/10.48550/ARXIV.2302.13971

United Nations. (n. d.). Understanding hate speech. Verfügbar 19. Oktober 2025 unter https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech

VAGO solutions. (n. d.). Llama-3.1-SauerkrautLM-70b-Instruct. https://huggingface.co/VAGOsolutions/Llama-3.1-SauerkrautLM-70b-Instruct

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett (Hrsg.), *Advances in Neural Information Processing Systems* (Bd. 30). Curran Associates, Inc.

Vijayarani, S., Ilamathi, M. J., Nithya, M., et al. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, *5*(1), 7–16.

Waltinger, U., et al. (2010). GermanPolarityClues: A Lexical Resource for German Sentiment Analysis. *LREC*, 1638–1642.

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, *55*(7), 5731–5780. https://doi.org/10.1007/s10462-022-10144-1

Weber, K. (2022). Land unter: Die Jahrhundertflut an der Elbe. Verfügbar 31. Oktober 2025 unter https://www.ndr.de/geschichte/chronologie/Jahrhundertflut-an-der-Elbe-Die-Hochwasser-Katastrophe-2002,elbehochwasser165.html

Webster, J. J., & Kit, C. (1992). Tokenization as the initial phase in NLP. *COLING 1992 volume 4: The 14th international conference on computational linguistics*.

Weissenbacher, M., & Kruschwitz, U. (2024). Analyzing Offensive Language and Hate Speech in Political Discourse: A Case Study of German Politicians. In R. Kumar, A. K. Ojha, S. Malmasi, B. R. Chakravarthi, B. Lahiri, S. Singh & S. Ratan (Hrsg.), *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024* (S. 60–72). https://aclanthology.org/2024.trac-1.8/

Wenzek, G., Lachaux, M.-A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., & Grave, E. (2020). CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (Hrsg.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (S. 4003–4012). European Language Resources Association. https://aclanthology.org/2020.lrec-1.494/

Widmann, T., & Wich, M. (2022). Creating and Comparing Dictionary, Word Embedding, and Transformer-Based Models to Measure Discrete Emotions in German Political Text. *Political Analysis*, *31*(4), 626–641. https://doi.org/10.1017/pan.2022.15

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., ... Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Young, L., & Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. *Political communication*, *29*(2), 205–231. https://doi.org/https://doi.org/10.1080/10584609.2012.671234

Zhang, W., Deng, Y., Liu, B., Pan, S. J., & Bing, L. (2023). Sentiment Analysis in the Era of Large Language Models: A Reality Check [arXiv preprint]. https://doi.org/10.48550/ARXIV.2305.15005

Zheng, A. (2015). *Evaluating Machine Learning Models*. O'Reilly Media, Inc.