Automatisierte Zielerkennung in politischen Twitter-Diskursen mittels Ensemble-Learning

https://github.com/KOLLGO/TargetDetector

Zusammenfassung

Die automatische Identifikation von Zielobjekten und -gruppen (Target Detection) in sozialen Medien ist eine zentrale Herausforderung, da die Informalität und Kontextabhängigkeit von Kurznachrichten eine eindeutige Zuordnung für die Analyse von Onlinediskursen erschwert. In dieser Arbeit wird ein System zur Multi-Klassen-Klassifikation (*individual, group, public*) vorgestellt, welches auf einem Datensatz von 14.470 manuell annotierten deutschsprachigen Tweets aus dem Pegida-Kontext trainiert wurde.
Dieses System vereint eine umfassende Vorverarbeitung mit Term-Frequenz-Inverse-Dokumentenfrequenz-(TF-IDF)-Vektoren als grundlegende Feature-Repräsentation sowie zusätzliche Features wie Pronomen und anonymisierte Mentions. Zur Klassifikation wurde ein Soft-Stacking-Ensemble eingesetzt, das die Vorhersagewahrscheinlichkeiten von Support Vector Classifier, Multinomial Naive Bayes, Logistischer Regression und Random Forest über ein Meta-Modell aggregiert. Die Evaluation mittels fünffacher Cross-Validation belegt die Robustheit des Ensembles, das mit einem Makro-F1-Score von 0,6694 sowie einer Makro-Precision von 0,7331 die Einzelmodelle übertrifft. Die Ergebnisse verdeutlichen, dass die Integration linguistischer Regeln in maschinelle Lernverfahren die Leistungsfähigkeit bei der Erkennung des Ziels von politischen Kurztexten entscheidend steigert.

Keywords: Target Detection, hybride Textklassifikation, Twitter, Ensemble Learning, Multilabel Klassifikation.

1 Einführung

Soziale Medien haben sich zu einem zentralen Schauplatz politischer Meinungsbildung, Mobilisierung und Polarisierung entwickelt (Yadav, 2024). Plattformen wie Twitter (heute bekannt als X) dienen als diskursive Räume, in denen diverse Akteure ihre Positionen aushandeln. Die Besonderheiten digitaler Kurznachrichten erschweren jedoch die Analyse dieser Diskurse. Slang, Ironie und implizite Referenzen prägen die informelle Sprache und fordern automatisierte Verfahren der Textanalyse heraus (Qaiser & Ali, 2018). In politischen Kontexten bleibt dabei oft unklar, an wen sich eine Äußerung konkret richtet, da Adressaten häufig nicht explizit benannt werden (Jiang et al., 2011). Die automatisierte Zielerkennung (Target Detection) ist daher essenziell, um öffentliche Onlinediskurse zu verstehen und Phänomene wie Polarisierung oder Hate Speech präzise zu erfassen. Die Arbeit behandelt diese Herausforderung durch die Entwicklung eines Systems zur Multi-Klassen-Klassifikation adressierter Akteure in deutschsprachigen Tweets. Auf Basis eines Datensatzes von 14.470 manuell annotierten Beiträgen aus dem Kontext der Pegida-Bewegung (2014–2016) wird die Unterscheidung zwischen den Kategorien *individual*, *group* und *public* vorgenommen. Um die Defizite rein statistischer Modelle bei kurzen Texten zu überwinden (Han et al. 2013), kombiniert der vorgestellte Ansatz spezifische linguistische Merkmale mit statistischen Wort-N-Grammen in einem Ensemble-Klassifikator. Um die Generalisierungsfähigkeit in diesen dynamischen Diskursen zu steigern, setzt die vorliegende Arbeit auf eine dezidierte Vorverarbeitung durch Anonymisierung und Pronomen-Filterung. Das entwickelte Stacking-Ensemble demonstriert, dass die gezielte Reduktion linguistischen Rauschens die Trennschärfe zwischen den Zielkategorien maßgeblich verbessert. Damit legt dieses System nicht nur eine robuste Basis für die automatisierte Beobachtung gesellschaftlicher Konfliktlinien, sondern eröffnet zugleich Perspektiven für künftige Erweiterungen: Die Einbeziehung von Thread-Kontexten sowie der Einsatz von Transformer-Architekturen stellen vielversprechende Wege dar, um auch implizite Adressierungen und komplexe Multi-Label-Konstellationen in der politischen Online-Kommunikation aufzulösen.
Die weitere Arbeit ist wie folgt aufgebaut: Nach einer Einordnung in den aktuellen Forschungsstand in Abschnitt 1 folgt in Abschnitt 2 eine Diskussion verwandter Arbeiten und eine Einordnung des vorgestellten Ansatzes in den bestehenden Forschungskontext. In Abschnitt 3 folgt die detaillierte Beschreibung der Datenbasis sowie der Methodik. Die experimentellen Ergebnisse werden in Abschnitt 4 präsentiert und in Abschnitt 5 kritisch diskutiert, bevor die Arbeit in Abschnitt 6 zusammengefasst wird und Perspektiven aufgezeigt werden.

2 Literaturdiskussion

Die automatische Erkennung von Zielobjekten und -gruppen in Texten, auch Target Detection genannt, gewann in der forensischen Sprachverarbeitung zunehmend an Bedeutung (Lemmens et al., 2021). Ziel war es, sprachliche Äußerungen, insbesondere in sozialen Medien, bestimmten Adressaten oder Gruppen zuzuordnen, um potenziell bedrohliche oder diskriminierende Inhalte gezielt identifizieren zu können. In der Forschung existierten dazu verschiedene Ansätze, die sich bezüglich der Methodik, Datenlage und Zieldefinition unterschieden (Amjad et al., 2021). Während die Target Detection häufig im Kontext von Hate-Speech-Analysen untersucht wurde, da beide Bereiche oft auf denselben Datengrundlagen (z. B. Tweets) basierten und ähnliche methodische Herausforderungen wie die Analyse informeller Sprache teilten, stellte sie ein eigenständiges Aufgabenfeld dar, das über die reine Erkennung von Hassrede hinausging.
Frühere Arbeiten zur Textklassifikation nutzten vor allem traditionelle Verfahren wie Support Vector Machines (SVM) oder Naive Bayes auf Basis von Bag-of-Words oder TF-IDF-Merkmalen (Adewoye und Ara, 2024). Diese Ansätze galten als interpretierbar und effizient, stießen jedoch bei komplexer Semantik oder verschleierten sprachlichen Mustern an ihre Grenzen (Adewoye und Ara, 2024). Neuere Untersuchungen belegten, dass tiefe neuronale Netze und insbesondere Transformer-Modelle wie Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) inzwischen einen Standard in der Textanalyse darstellen, da sie kontextuelle Abhängigkeiten wesentlich besser erfassten und damit die Klassifikations-

güte erhöhen konnten. Dennoch war die Übertragbarkeit von Ergebnissen aus anderen Bereichen der Textklassifikation auf die Target Detection in Kurznachrichten oft eingeschränkt. So untersuchten Adewoye und Ara (2024) beispielsweise die Klassifikation von Themen in Newsgroups, deren strukturelle Merkmale sich deutlich von der Identifikation adressierter Akteure in Tweets unterschieden.

Im Bereich der Target Detection zeigten aktuelle Studien, dass spezialisierte Modellarchitekturen die Zielidentifikation verbesserten. Amjad et al. (2021) kombinierten in einem zweistufigen Ansatz Bedrohungserkennung und Target-Identifikation für Urdu-Tweets. Sie erzielten damit eine durchschnittliche F1-Score-Steigerung von rund sechs Prozent gegenüber konventionellen Modellen, worunter in diesem Kontext Baselines ohne die Kombination dieser zwei Aufgabenstufen zu verstehen waren. Lemmens et al. (2021) erweiterten die Hate-Speech-Erkennung um metaphorische Sprachmerkmale und verbesserten so die Erkennungsrate zielgerichteter Aussagen um bis zu vier Prozent, wobei sie die Zielgruppen nach spezifischen Opferkategorien definierten. Rashid et al. (2025) verwendeten einen Multi-Stage- und Multi-Label-Ansatz für toxische Sprache und zeigten, dass eine schrittweise Klassifikation robuster gegenüber Rauschen und Klassenungleichgewicht war. Thapa et al. (2025) demonstrierten in einer Shared Task für Devanagari-Sprachen, dass die gemeinsame Modellierung von Hate Speech, Sprachidentifikation und Target Detection die Transferleistung zwischen Sprachen erhöhte, was für forensische Systeme mit gemischten Sprachumgebungen relevant war.

Neben überwachten Verfahren gewannen unüberwachte oder semi-überwachte Methoden an Bedeutung. Ollagnier et al. (2023) nutzten Multi-View-Clustering, um Zielgruppen und Opfergemeinschaften in multilingualen Hate-Speech-Datensätzen ohne Labels zu identifizieren. Diese Ansätze waren besonders relevant, wenn keine manuell annotierten Daten vorhanden waren – ein Fall, der im forensischen Alltag häufig auftritt, da bei tagesaktuellen Ermittlungen oft große Datenmengen ohne vorherige Kennzeichnung gesichtet werden müssen. Einen weiteren Fortschritt stellte das Multi-Task-Learning dar. Rajamanickam et al. (2024) zeigten, dass das gleichzeitige Training von Hate-Speech-, Emotions- und Target-Erkennung die Generalisierungsfähigkeit neuronaler Modelle verbesserte.

Verwandte Forschungsrichtungen lieferten zusätzliche Erkenntnisse: Xing und Tsang (2024) modellierten kontextuelle Zielattribute mithilfe von Graphenstrukturen. Olusegun et al. (2023) belegten, dass hybride Convolutional Neural Network(CNN)- und Long Short-Term Memory(LSTM)-Architekturen effektiv zwischen emotionalen Nuancen unterscheiden konnten und damit für implizite Beziehungen zwischen Zielgruppen nutzbar sind. Zudem zeigten Volkovs et al. (2020), dass Sprachmodelle wie Multilingual BERT (mBERT) in Social-Media-Kontexten erfolgreich eingesetzt werden konnten, um Beziehungen zwischen Autor und Adressat zu erkennen. Methoden aus dem Bereich des Ähnlichkeitslernens zeigten ebenfalls Potenzial für Target-Similarity-Analysen (Ghawi & Pfeffer, 2019).

Wesentliche methodische Herausforderungen bestehen weiterhin in der Behandlung von Klassenungleichgewichten und in der Evaluation der Ergebnisse. Nti et al. (2021) zeigten, dass die Wahl der K-Fold-Parameter einen signifikanten Einfluss auf die Varianz von Performance-Schätzungen hatte. Da Texte häufig mehrere Zielgruppen betrafen, war zudem die Multi-Label-Klassifikation entscheidend, für die nach Tsoumakas und Katakis (2017) spezielle Verfahren wie Classifier Chains erforderlich sind. Neben technischen Aspekten ist auch der kommunikative Kontext relevant. Basenko et al. (2022) betonen, dass der sogenannte Adressatenfaktor, also die soziale Rolle und Beziehung zwischen Sender und Empfänger, für das Verständnis von Sprache essenziell ist. In der forensischen Anwendung bedeutet dies, dass Target-Detection-Systeme nicht nur Textinhalte, sondern auch Kontext- und Interaktionsstrukturen berücksichtigen sollten, um Fehlklassifikationen zu vermeiden.

Zusammenfassend lässt sich feststellen, dass spezialisierte Ansätze und erste Versuche mit Transformer-Modellen in der Zielerkennung vielversprechende Ergebnisse liefern(Basenko et al. 2022). Dennoch bestehen offene Herausforderungen hinsichtlich der Kombination von überwachten und unüberwachten Verfahren, der Integration adressatenbewusster Kontextmodelle und dem Bias, wodurch die Minderheitsklasse bei Klassenungleichgewicht schlechter erkannt wird, bei Klassenungleichgewicht, unter Klassenungleichgewicht. Die vorliegende Arbeit adressiert diese Lücken, indem sie ein hybrides System entwickelte, das statistische N-Gramme mit gezielten linguistischen Merkmalen adressiert. Im Gegensatz zu rein tiefenlernbasierten Modellen setzte dieser Ansatz auf die Integration von Domänenwissen durch spezifische Filterregeln, um die Robustheit in informellen politischen Diskursen zu erhöhen und die im Forschungsstand identifizierte Schwierigkeit der Klassenabgrenzung zwischen Einzelpersonen, Gruppen und der Öffentlichkeit methodisch zu bewältigen.

3 Daten und Methoden

In diesem Kapitel werden die zugrundeliegenden Daten sowie die methodische Vorgehensweise zur Entwicklung des Systems vorgestellt. Die Prozesskette reicht von der Vorverarbeitung über die Merkmalsextraktion bis hin zur Klassifikation mittels eines Stacking-Ensembles.

3.1 Datengrundlage

Die Grundlage dieser Arbeit bildet ein Korpus von 14.470 deutschsprachigen Tweets und Kommentaren, die zwischen Dezember 2014 und Juli 2016 im Kontext der Pegida-Bewegung erhoben wurden (Felser et al., 2025). Bei der Annotation des Datensatzes wurden drei Zielkategorien unterschieden: *individual* (Einzelpersonen), *group* (Organisationen/Gruppen) und *public* (Allgemeinheit). Aufgrund der wesensbedingten Unausgewogenheit der Klassenverteilung (siehe Tabelle 1) erfordert die Evaluation spezifische Metriken, um Verzerrungen durch dominante Klassen zu vermeiden (Nti et al., 2021).

Zum Schutz sensibler Informationen werden Mentions (Erwähnungen von Personen und Organisationen) durch funktionale Platzhalter ersetzt. Diese Anonymisierung dient nicht nur dem Datenschutz (Vogel et al., 2019), sondern erhält gleichzeitig strukturelle Informationen, da die Platzhalter als informative Marker für die Target-Bestimmung im Datensatz verbleiben.

Tabelle 1: **Klassenverteilung des Datensatzes**

TAR	Anzahl	Anteil in %
public	8721	60,27
individual	3289	22,73
group	2460	17,00

3.2 Vorverarbeitung

Die Vorverarbeitung der Textdaten dient der Vereinheitlichung der Textrepräsentation sowie der Reduktion von Rauschen, was für die Analyse kurzer Social-Media-Texte als entscheidend gilt (Lubis & Nasution, 2023). Alle Tweets werden zunächst

in Kleinschreibung konvertiert (Lowercasing), um semantisch identische Token einheitlich zu behandeln (Qaiser & Ali, 2018). Da Hyperlinks, Ziffern sowie Satz- und Sonderzeichen häufig keinen inhaltlichen Mehrwert besitzen, werden diese anschließend einer regelbasierten Textnormalisierung unterzogen und somit entfernt (Wang et al., 2011; Han et al., 2013). Das @-Symbol wird bewusst beibehalten, um die Integrität der für die Target Detection essenziellen anonymisierten Mentions zu wahren (Jiang et al., 2011). Zur Standardisierung der Zeichenkodierung werden zudem deutsche Umlaute durch ihre jeweiligen Digraphen (z. B. „ä“ zu „ae“) sowie das Eszett („ß“ zu „ss“) ersetzt.
Ein zentraler Bestandteil der Vorverarbeitung ist die Handhabung von Stoppwörtern unter Verwendung des Sprachmodells spaCy, um inhaltlich wenig aussagekräftige Wörter zu eliminieren (Felden et al., 2006). Da Pronomen für die Target Detection eine zentrale Rolle spielen, werden relevante Pronomen über eine Whitelist gezielt beibehalten, wie es in früheren Arbeiten zur zielabhängigen Textanalyse vorgeschlagen wird (Jiang et al., 2011). Da sich die Pronomen der ersten Person Singular (z. B. „ich“, „mir“) überwiegend auf den Autor beziehen und keinen direkten Aufschluss über das adressierte Target geben, wodurch diese ein hinderliches Rauschen darstellen, wurden diese konsequent entfernt (Pradana & Hayaty, 2019).

3.3 Feature-Engineering

Zur Repräsentation der Texte wird ein Feature-Satz verwendet, der statistische Muster mit gezielten linguistischen Merkmalen kombiniert. Die primäre Textrepräsentation basiert auf TF- IDF-gewichteten Wort-N-Grammen (Uni-, Bi- und Trigramme), wobei die relevantesten Merkmale unter Nutzung von sublinearer Skalierung und L2-Normalisierung extrahiert werden (Qaiser & Ali, 2018). N-Gramme sind insbesondere für kurze Texte geeignet, da sie lokale Kontextinformationen abbilden können (Schonlau & Guenther, 2020).
Ergänzend wurden linguistische Zählfeatures integriert, darunter eine zusammengestellte Liste generischer Begriffe, welche Gruppen repräsentieren (z. B. „Leute“, „alle“), sowie eine Personal- und Possessivpronomen der zweiten und dritten Person, die für die Identifikation von Zielgruppen essenziell sind. Zusätzlich wurde die Häufigkeit der Anonymisierungs-Tags ([*@IND*], [*@GRP*], etc.) als numerisches Merkmal erfasst, da das direkte Adressieren einer Entität ein Schlüsselfaktor für die Bestimmung des Targets ist (Vo & Zhang, 2015). Der Einsatz solcher linguistischen Merkmale wird in mehreren Studien als sinnvoller Zusatz zu rein statistischen Features beschrieben (Demus et al., 2023; Jiang et al., 2011).

3.4 Klassifikationssystem

Zur Klassifikation wurde ein Ensemble-Ansatz in Form eines Soft-Stacking-Klassifikators eingesetzt, um die Stärken unterschiedlicher Lernverfahren zu kombinieren und die Schwächen einzelner Modelle auszugleichen (Ghawi & Pfeffer, 2019). Die Basisschicht des Ensembles setzt sich aus vier diversen Modellen zusammen: einer Support Vector Machine (SVM) mit Radial Basis Function (RBF)-Kernel, einem Multinomial-Naive-Bayes-Klassifikator (MNB) mit Alpha-Glättung, einer logistischen Regression (LR) und einem Random-Forest-Klassifikator (RF). Diese Verfahren haben sich in früheren Vergleichsstudien als robust für Textklassifikationsaufgaben erwiesen (Pranckevičius & Marcinkevičius, 2017). Bei den Modellen SVC, RF und LR wurde eine gewichtete Klassifizierung angewendet, um dem Klassenungleichgewicht entgegenzuwirken.
Die Basismodelle lieferten jeweils Wahrscheinlichkeitsvorhersagen, die als Eingabe für einen Meta-Klassifikator dienen. Als Meta-Modell wurde eine logistische Regression verwendet, da diese eine stabile Kombination der Basismodellvorhersagen erlaubt (Wolpert, 1992; Ghawi & Pfeffer, 2019). Die Hyperparameter der Basismodelle wurden systematisch mittels einer verschachtelten Grid Search (dreifold innerer Split) optimiert, was das Risiko optimistischer Verzerrungen reduziert (Varma & Simon, 2006).

3.5 Evaluation

Die Evaluation des Systems erfolgte mittels einer fünffachen stratifizierten Cross-Validation, was eine robuste Abschätzung der Modellleistung sicherstellt (Nti et al., 2021). Dieses Verfahren gewährleistet, dass jedes Dokument genau einmal als Testinstanz verwendet wird und die Stratifizierung das ursprüngliche Klassengewicht in jedem Teildatensatz proportional erhält (Lo et al., 2015; Sulistiana & Muslim, 2020). Aufgrund der ungleichen Klassenverteilung werden makro-gemittelte Precision-, Recall- und F1-Werte berechnet, da diese Metriken jede Klasse gleichgewichtet berücksichtigen und verhindern, dass dominante Klassen die Gesamtleistung verzerren (Nti et al., 2021; Powers, 2011).

4 Ergebnisse

Die Evaluation des entwickelten Systems mittels einer fünffachen Cross-Validation verdeutlicht, dass der eingesetzte Ensemble-Ansatz eine robuste und leistungsfähige Lösung für die Target Detection in deutschsprachigen Kurztexten darstellt. Für alle Tests wurde der Zufalls-Seed auf 42 gesetzt, um reproduzierbarkeit zu gewährleisten. Der Soft-Stacking-Klassifikator erzielt konsistent höhere makro-gemittelte F1-Werte als die einzelnen Basismodelle und erreicht einen durchschnittlichen Makro-F1-Score von 0,6669 bei einer Makro-Precision von 0,7331 sowie einem Makro-Recall von 0,6409. Die geringe Varianz über alle Folds hinweg deutet auf eine hohe Robustheit gegenüber unterschiedlichen Trainings- und Testaufteilungen hin, was vergleichbare Beobachtungen zur Stabilität von Ensemble-Ansätzen in früheren Studien stützt (Ghawi & Pfeffer, 2019).
Die Analyse der Precision-Werte verdeutlicht die Überlegenheit des Ensemble-Verfahrens gegenüber den Einzelklassifikatoren (siehe Abbildung 1). Der Stacking-Klassifikator erzielt die höchste Exaktheit und zeigt über alle fünf Folds hinweg eine konsistente Performance mit einem Makro-Mittelwert von 0,7331 (siehe Abbildung 1). Den zweithöchsten Wert erreichte der MNB mit einer durchschnittlichen Precision von 0,7000 (siehe Abbildung 1). Der Support Vector Classifier (SVC) weist im Gegensatz dazu mit 0,6543 die geringste Präzision im Testfeld auf (siehe Abbildung 1). Dies lässt darauf schließen, dass der Stacking-Ansatz besonders effektiv darin ist, die Rate der Falsch-Positiv-Klassifizierungen zu minimieren, indem er die Stärken der Basismodelle kombiniert (siehe Abbildung 1).
Bei der Betrachtung des Recalls ergibt sich eine Verschiebung der Leistungsverhältnisse (siehe Abbildung 2). Hier erzielen die Logistische Regression (0,6594) und der SVC (0,6565) die besten Ergebnisse, was auf eine höhere Sensitivität dieser Modelle bei der Identifizierung relevanter Instanzen hindeutet (siehe Abbildung 2). Ein Leistungsabfall ist hingegen bei Betrachtung des MNB zu beobachten, welcher mit einem Wert von 0,6163 deutlich hinter die anderen Modelle zurückfällt (siehe Abbildung 2). Der Stacking CLF ordnet sich mit einem Recall von 0,6409 im Mittelfeld ein (siehe Abbildung 2).
Der F1-Score bietet als harmonisches Mittel zwischen Precision und Recall die fundierteste Grundlage für die Bewertung der Gesamtperformance (siehe Abbildung 3). Der Stacking-CLF bestätigt hier seine Position als leistungsfähigstes Modell mit einem durchschnittlichen F1-Score von 0,6694 (siehe Abbildung

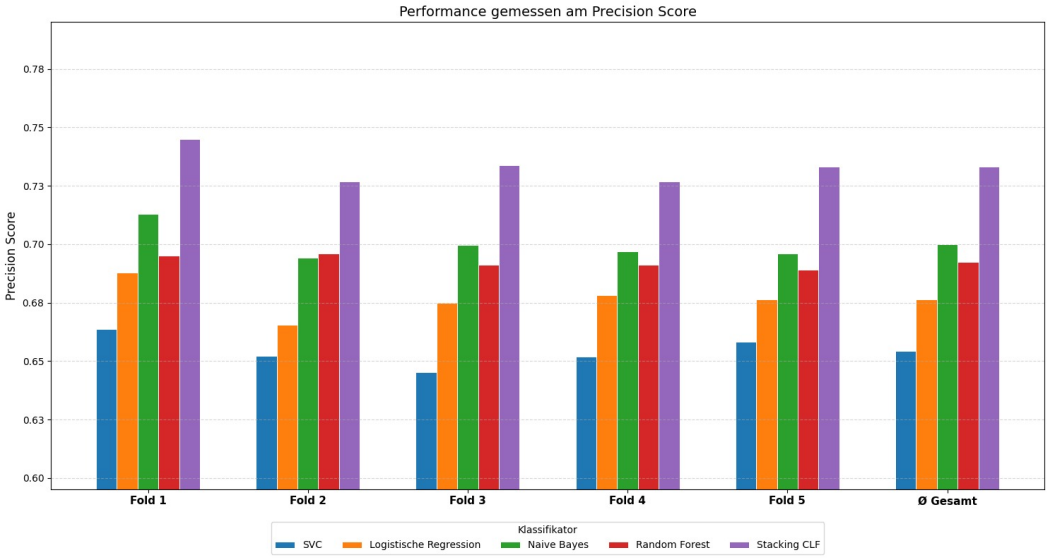

Abbildung 1: **Performance gemessen an Precision-Score**

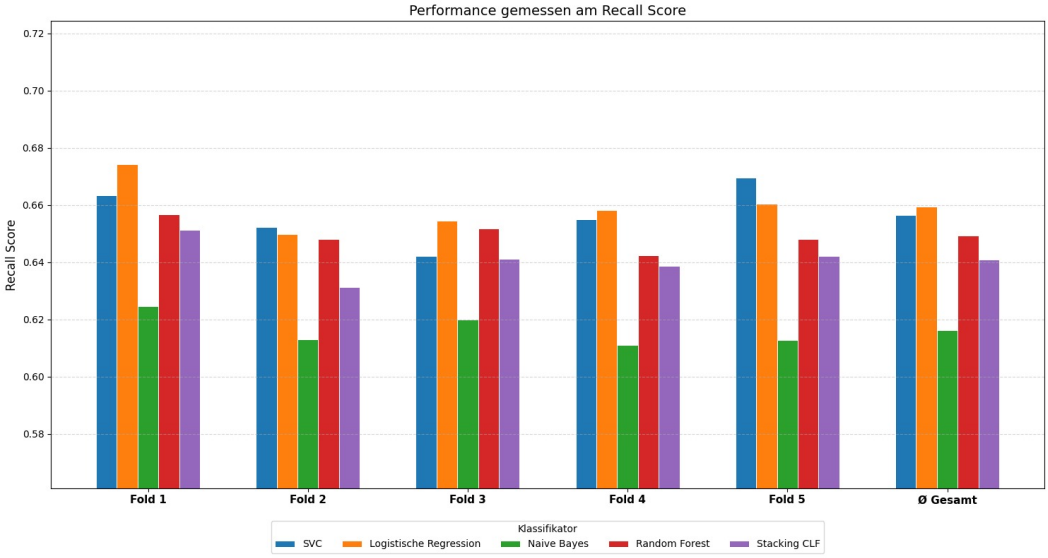

Abbildung 2: **Performance gemessen an Recall-Score**

3). Dicht darauf folgen der Random Forest (0,6665) und die Logistische Regression (0,6655), die eine sehr ähnliche Gesamtleistung erbringen (siehe Abbildung 3). Die geringe Standardabweichung, insbesondere beim Random Forest ($\pm$ 0,0037), unterstreicht die Robustheit der Klassifikatoren über verschiedene Folds hinweg (siehe Abbildung 3). Naive Bayes ist, trotz der guten Precision-Werte, aufgrund des schwachen Recalls mit einem F1-Score von 0,6420 der schwächste Klassifikator der Untersuchung. Zusammenfassend lässt sich festhalten, dass der Stacking-Ansatz die stabilste und ausgewogenste Gesamtleistung über alle Metriken hinweg bietet (siehe Abbildung 3).
Ergänzend zur Modellbewertung zeigen die Ergebnisse der Feature-Analyse, dass das gezielte Entfernen von Pronomen der ersten Person Singular zu einer messbaren Verbesserung der Makro-F1-Werte von 0,6630 auf 0,6690 führt.
Darüber hinaus bewirkt die Integration linguistischer Merkmale, insbesondere die Pronomen-Filterung, die Einbeziehung generischer Begriffe sowie anonymisierter Mentions, höhere Precision- und Recall-Werte im Vergleich zu rein statistischen N-Gramm-Modellen. Diese Merkmalskombination erweist sich insbesondere bei der Unterscheidung der Klassen *group* und *public* als vorteilhaft und führt zu verbesserten Klassifikationsergebnissen.
Um die Effektivität zu evaluieren, werden drei verschiedene Szenarien gegenübergestellt: Ein Basismodell ohne Vorverarbeitung und mit Berücksichtigung von Ich-Pronomen (siehe Tabelle 2), ein weiteres Basismodell mit Vorverarbeitung und mit Berücksichtigung von Ich-Pronomen (siehe Tabelle 2), sowie das optimierte Modell (siehe Tabelle 2).
Der Vergleich zeigt, dass das Modell ohne Vorverarbeitung zwar robust ist, jedoch in allen entscheidenden Metriken leicht hinter den Ergebnissen mit Preprocessing zurückbleibt (siehe Tabelle 2). Besonders deutlich wird dies beim Vergleich des Stacking Classifiers, der in beiden Testreihen als stärkstes Modell hervorging:
Während die Precision, also die Exaktheit der Vorhersagen, in allen Modellen auf einem sehr ähnlichen Niveau über rund 0,73 liegt, zeigen sich die größten Unterschiede beim Recall und dem

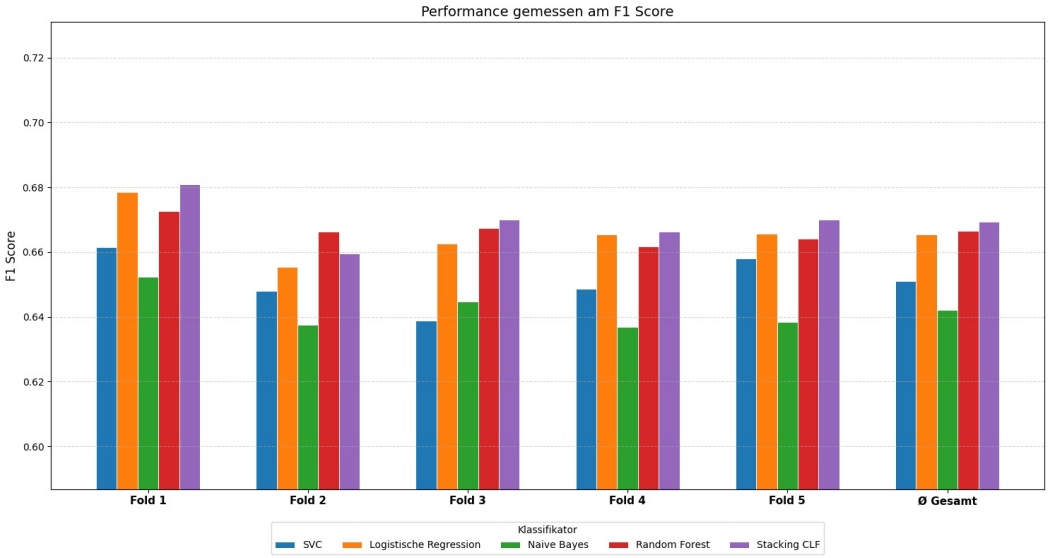

Abbildung 3: **Performance gemessen an F1-Score**

Tabelle 2: **Vergleich Leistungsmetriken des Stacking Classifiers nach Vorverarbeitungsstufen**

Metrik (Stacking CLF)	**ohne Preprocessing mit Ich-Pronomen**	**mit Preprocessing mit Ich-Pronomen**	**mit Preprocessing ohne Ich-Pronomen**
Precision (Makro-AVG)	0,7325	0,7321	0,7331
Recall (Makro-AVG)	0,6283	0,6406	0,6409
F1-Score (Makro-AVG)	0,6587	0,6690	0,6694

daraus resultierenden F1-Score. Das finale Modell mit vollständiger Verarbeitungs-Pipeline erzielt hier mit einem F1-Score von 0,6694 das beste Gesamtergebnis.

5 Diskussion

Die Ergebnisse dieser Arbeit bestätigen die Annahme, dass die Target Detection in politischen Tweets eine anspruchsvolle Klassifikationsaufgabe darstellt, bei der die Kombination statistischer Wort-N-Gramme mit linguistisch motivierten Merkmalen einen messbaren Mehrwert bietet. Insbesondere Pronomen, generische Begriffe und anonymisierte Mentions erwiesen sich als hochgradig informative Indikatoren für adressierte Akteure, was frühere Befunde zur Bedeutung sprachlicher Referenzen in Social-Media-Texten stützt (Jiang et al., 2011; Demus et al., 2023).
Ein wesentlicher Erfolgsfaktor ist dabei die gezielte Vorverarbeitung: Die Entscheidung, Pronomen der ersten Person Singular zu entfernen, steigert die Generalisierungsfähigkeit des Modells. Da diese Ausdrücke primär Informationen über die Haltung des Verfassers preisgeben und selten Aufschluss über das eigentliche Ziel einer Äußerung geben, kann durch deren Eliminierung das semantische Rauschen im hochdimensionalen Merkmalsraum reduziert werden. Der gewählte Ensemble-Ansatz zeigt deutliche Vorteile hinsichtlich Robustheit und Varianzreduktion, indem er die komplementären Stärken der Basismodelle effektiv nutzt. Die Support Vector Machine und die logistische Regression reagieren sensibler auf Rauschen im Merkmalsraum. Der Stacking-Klassifikator kompensiert diese modellindividuellen Schwächen teilweise und glättet Leistungsschwankungen über verschiedene Folds hinweg. Dies deckt sich mit Beobachtungen aus vergleichbaren Studien, in denen Meta-Klassifikatoren stabilere Ergebnisse liefern als isolierte Lernverfahren (Ghawi & Pfeiffer, 2019).
Hinsichtlich des Recalls erweisen sich die Logistische Regression (0,6594) und der SVC (0,6565) als am sensitivsten, während der MNB mit 0,6163 deutlich abfällt (siehe Abbildung 2). Dieser Effekt verdeutlicht einen klassischen Trade-off: Während Naive Bayes präzise Vorhersagen trifft, übersieht das Modell im Vergleich zu den anderen Modellen eine größere Anzahl tatsächlich positiver Fälle (siehe Abbildung 2). Der Stacking Classifier ordnet sich mit 0,6409 im Mittelfeld ein, ohne die Sensitivität der besten Einzelmodelle zu erreichen.
Trotz der robusten Gesamtleistung verdeutlicht die Analyse auch die wesensbedingten Grenzen rein textbasierter Ansätze. Eine zentrale Herausforderung bleibt die Identifikation impliziter Targets, bei denen Akteure ohne namentliche Nennung oder eindeutige Markierungen adressiert werden. In solchen Fällen stößt die oberflächenbasierte Merkmalsextraktion an ihre Grenzen, da Modelle auf kontextuelles oder weltwissenbasiertes Verständnis zurückgreifen müssen, das mit klassischen Feature-basierten Methoden nur eingeschränkt abbildbar ist. Diese Komplexität spiegelt sich auch in der menschlichen Urteilsbildung wider, da Targets selbst für Experten nicht immer eindeutig klassifizierbar sind (Jiang et al., 2011). Zudem ist die Abgrenzung zwischen den Klassen *group* und *public* konzeptuell unscharf, was die Modellleistung bei diesen selteneren Kategorien zusätzlich erschwert. Technisch betrachtet ergeben sich spezifische Limitationen bei der Merkmalsgenerierung, die in zukünftigen Iterationen adressiert werden sollten. Da anonymisierte Mentions (wie [*@GRP*]) aktuell nicht als „Word of Interest“ in die N-Gramm-Filterung einbezogen werden, bleiben potenziell aussagekräftige Wortkombinationen wie „[*@IND*] ist“

ungenutzt. Ein weiterer kritischer Punkt betrifft den Umgang mit Artikeln. Da diese nicht in der finalen Whitelist aufgenommen wurden, gehen potenziell relevante syntaktische Hinweise auf die Numerus-Kategorisierung (Singular oder Plural) verloren, welche zur Unterscheidung von *individual* und *group* hätten beitragen können (Lo et al., 2015).
Schließlich unterstreicht das Scheitern von Oversampling-Techniken zur Behebung des Klassenungleichgewichts, dass für die Identifikation seltener Zielstrukturen in kurzen, informellen Texten die Qualität der linguistischen Merkmale schwerer wiegt als rein quantitative Datenmanipulationen. Insgesamt verdeutlicht die Analyse, dass eine erfolgreiche Target Detection von einer sich ergänzenden Kombination aus linguistischem Domänenwissen, präziser Bereinigung und Ensemble-Strategien profitiert.
Die Robustheit des Modells, auch ohne Vorverarbeitung, resultiert primär aus der methodischen Konfiguration: Die TF-IDF-Vektorisierung mit den Top 10.000 Features schließt Ausreißer aus, während die sublineare Skalierung den Einfluss dominanter Begriffe dämpft und die L2-Normalisierung unterschiedliche Tweet-Längen ausgleicht. Eine fünffache Cross-Validation sichert dabei die Stabilität der Ergebnisse gegen zufällige Datenverteilungen ab. Dass die optimierte Pipeline dennoch überlegen ist, ersichtlich am Anstieg des F1-Scores von 0,6587 auf 0,6694, liegt an der tieferen semantischen Erfassung durch das Preprocessing. Maßnahmen wie Lowercasing zur Vereinheitlichung von Schreibweisen, das Entfernen von störenden Sonderzeichen sowie eine präzisere Tokenisierung ermöglichen es dem Modell, relevante Merkmale besser zu isolieren, was insbesondere die Trefferquote (Recall) im Vergleich zum Basismodell deutlich steigert.

6 Zusammenfassung und Ausblick

Das in dieser Arbeit vorgestellte System stellt eine robuste Lösung für die automatisierte Target Detection in deutschsprachigen Kurztexten dar. Durch die Kombination einer sorgfältigen Vorverarbeitung, insbesondere der gezielten Anonymisierung und der strategischen Pronomen-Filterung, mit linguistisch motivierten Feature-Engineering-Ansätzen und einem Stacking-Ensemble können konsistente und ausbalancierte Ergebnisse erzielt werden. Die Evaluation verdeutlicht, dass die bewusste Reduktion von Rauschen, etwa durch das Entfernen selbstreferenzieller Bezüge, die Generalisierungsfähigkeit des Modells in einem hochgradig informellen und polarisierten Diskursumfeld maßgeblich verbessert. Mit einem Makro-F1-Score von 0,6694 zeigt der vorgeschlagene Ansatz eine hohe Stabilität und Vorhersagekraft, wobei der Ensemble-Klassifikator modellbedingte Schwächen einzelner Basismodelle effektiv ausgleicht, wie in verwandten Arbeiten zu Meta-Klassifikatoren berichtet wird (Ghawi & Pfeffer, 2019).
Aus den eigenen Ergebnissen lässt sich schließen, dass die gezielte Kombination aus statistischen N-Grammen und linguistischen Features, insbesondere Pronomen und generische Begriffe, die Klassifikationsleistung entscheidend verbessert. Diese Merkmalskombination erweist sich vor allem bei der schwierigen Abgrenzung der Klassen *group* und *public* als vorteilhaft und unterstreicht die Bedeutung einer durchdachten Feature-Auswahl für die automatisierte Target Detection. Gleichzeitig macht die Analyse die unvermeidbaren Herausforderungen dieser Aufgabe deutlich. Tweets ohne explizite Zielmarkierungen oder mit stark kontextabhängigen Referenzen bleiben auch für das vorgestellte System schwer zu klassifizieren. Zudem ist die konzeptuelle Trennschärfe zwischen den Klassen *group* und *public* nicht immer eindeutig, was sowohl die manuelle Annotation als auch die automatisierte Klassifikation erschwert. Diese Befunde deuten darauf hin, dass die erreichbare Modellleistung teilweise durch die Ambiguität der Daten selbst begrenzt ist.
Für zukünftige Forschungsarbeiten ergeben sich aus den identifizierten Limitationen verschiedene Anknüpfungspunkte. Basierend auf den beobachteten Grenzen bei der Erkennung impliziter Targets können zukünftige Ansätze Kontextinformationen aus Tweet-Threads oder begleitende Metadaten integrieren, da die Einbeziehung vorangegangener Beiträge in früheren Studien als hilfreich für die Auflösung impliziter Referenzen beschrieben wird (Jiang et al., 2011).
Technologisch stellt der Einsatz moderner, kontextsensitiver Sprachmodelle (beispielsweise Transformer-Architekturen wie BERT) einen logischen nächsten Schritt dar, wie er in aktuellen Arbeiten zur Textklassifikation diskutiert wird (Demus et al., 2023). Solche Modelle können zukünftig prüfen, ob kontextualisierte Embeddings die expliziten linguistischen Features ergänzen oder gar ersetzen können. Des Weiteren bietet sich eine Erweiterung der Aufgabe zur Multi-Label-Klassifikation an, da Tweets in komplexen Debatten potenziell mehrere Targets gleichzeitig adressieren können (beispielsweise eine Gruppe und ein Individuum). Langfristig bietet das vorgestellte System eine solide Grundlage für die Einbettung in umfassendere Analysepipelines zur Untersuchung politischer Online-Diskurse (Vogel et al., 2019). In Kombination mit Modellen zur Erkennung von Hate-Speech oder zur Identifikation von Angriffen auf die demokratische Grundordnung ist Target Detection ein gutes Mittel zur automatisierten Beobachtung und Analyse gesellschaftlicher Konfliktlinien in sozialen Medien.

7 Literaturverzeichnis

Adewoye, M. B., & Ara, S. (2024). Comprehensive Review of Multiclass Text Classification using the 20 Newsgroup Dataset. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 10*(6), 1193-1212. https://doi.org/10.32628/CSEIT241061166

Amjad, M., Ashraf, N., Zhila, A., Sidorov, G., Zubiaga, A., & Gelbukh, A. (2021). Threatening language detection and target identification in Urdu tweets. *IEEE Access, 9*, 128302-128313. https://doi.org/10.1109/ACCESS.2021.3112500

Basenko, G., Revyakina, N., & Sakharova, E. (2022). The addressee factor in modern communication. *E3S Web of Conferences, 363*, 04041. https://doi.org/10.1051/e3sconf/202236304041

Demus, C., Labudde, D., Pitz, J., Probol, N., Schütz, M., & Siegel, M. (2023). Automatische Klassifikation offensiver deutscher Sprache in sozialen Netzwerken. In S. Jaki & S. Steiger (Hrsg.), *Digitale Hate Speech*, 65-88. Springer. https://doi.org/10.1007/978-3-662-65964-9

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding.* arXiv. https://doi.org/10.48550/arXiv.1810.04805

Felden, C., Bock, H., Gräning, A., Molotowa, L., Saat, J., Schaefer, R., Schneider, B., Steinborn, J., Voecks, J., & Wörle, C. (2006). *Evaluation von Algorithmen zur Textklassifikation* (Freiberg Working Papers Nr. 2006, 10). Fakultät für Wirtschaftswissenschaften, Technische Universität Bergakademie Freiberg. http://www.alexandria.unisg.ch/Publikationen/67273

Felser, J., Spranger, M., & Siegel, M. (2025). Overview of the GermEval 2025 shared task on harmful content detec-

tion. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, 306–319. HsH Applied Academics.

Ghawi, R., & Pfeffer, J. (2019). Efficient hyperparameter tuning with grid search for text categorization using KNN approach with BM25 similarity. *Open Computer Science, 9*, 160-180. https://doi.org/10.1515/comp-2019-0011

Han, B., Cook, P., & Baldwin, T. (2013). Lexical normalization for social media text. *ACM Transactions on Intelligent Systems and Technology, 4*(1), 1-27. https://doi.org/10.1145/2414425.2414430

Henning, S., Beluch, W., Fraser, A., & Friedrich, A. (2023). *A survey of methods for addressing class imbalance in deep learning-based NLP.* arXiv. https://doi.org/10.48550/arXiv.2210.04675

Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. (2011). Target-dependent Twitter sentiment classification. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 151-160. https://aclanthology.org/P11-1016

Lemmens, J., Markov, I., & Daelemans, W. (2021). Improving hate speech type and target detection with hateful metaphor features. *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, 7-16. https://doi.org/10.18653/v1/2021.nlp4if-1.2

Li, G., Wang, Z., Zhao, M., Song, Y., & Lan, L. (2022). Sentiment analysis of political posts on Hong Kong local forums using fine-tuned mBERT. *2022 IEEE International Conference on Big Data (Big Data)*, 6763-6765. https://doi.org/10.1109/BigData55660.2022.10020704

Lo, S. L., Chiong, R., & Cornforth, D. (2015). Using support vector machine ensembles for target audience classification on Twitter. *PLOS ONE, 10*(4), Artikel e0122855. https://doi.org/10.1371/journal.pone.0122855

Lubis, A. R., & Nasution, M. K. M. (2023). Twitter data analysis and text normalization in collecting standard word. *Journal of Applied Engineering and Technological Science, 4*(2), 855-863. https://doi.org/10.37385/jaets.v4i2.1991

Nti, I. K., Nyarko-Boateng, O., & Aning, J. (2021). Performance of machine learning algorithms with different K values in K-fold cross-validation. *International Journal of Information Technology and Computer Science, 13*(6), 61-71. https://doi.org/10.5815/ijitcs.2021.06.05

Ollagnier, A., Cabrio, E., & Villata, S. (2023). Unsupervised fine-grained hate speech target community detection and characterisation on social media. *Social Network Analysis and Mining, 13*(1), Artikel 58. https://doi.org/10.1007/s13278-023-01061-4

Olusegun, R., Oladunni, T., Audu, H., Houkpati, Y., & Bengesi, S. (2023). Text mining and emotion classification on Monkeypox Twitter dataset: A deep learning-natural language processing (NLP) approach. *IEEE Access, 11*, 49882-49894. https://doi.org/10.1109/ACCESS.2023.3277868

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *International Journal of Machine Learning Technology, 2*(1), 37-63. https://doi.org/10.48550/arXiv.2010.16061

Pradana, A. W., & Hayaty, M. (2019). The effect of stemming and removal of stopwords on the accuracy of sentiment analysis on Indonesian-language texts. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, 4*(4), 375-380. https://doi.org/10.22219/kinetik.v4i4.912

Pranckevičius, T., & Marcinkevičius, V. (2017). Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression classifiers for text reviews classification. *Baltic Journal of Modern Computing, 5*(2), 221-232. https://doi.org/10.22364/bjmc.2017.5.2.05

Qaiser, S., & Ali, R. (2018). Text mining: Use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications, 181*(1), 25-29. https://doi.org/10.5120/ijca2018917395

Rajamanickam, S., Muralidharan, A., & Singh, R. (2022). Multi-task learning for hate speech and target identification. *Knowledge-Based Systems, 247*, Article 108749. https://doi.org/10.1016/j.knosys.2022.108749

Rashid, A., Mahmood, S., Inayat, U., & Zia, M. F. (2025). Urdu toxicity detection: A multi-stage and multi-label classification approach. *AI, 6*(8), 194. https://doi.org/10.3390/ai6080194

Schonlau, M., Guenther, N., & Sucholutsky, I. (2017). Text mining using n-gram variables. *The Stata Journal, 17*(4), 866-881. https://doi.org/10.2139/ssrn.2759033

Sulistiana, & Muslim, M. A. (2020). Support Vector Machine (SVM) optimization using grid search and unigram to improve e-commerce review accuracy. *Journal of Soft Computing Exploration, 1*(1), 8-15. https://doi.org/10.52465/joscex.v1i1.3

Thapa, S., Rauniyar, K., Jafri, F. A., Adhikari, S., Sarveswaran, K., Bal, B. K., Veeramani, H., & Naseem, U. (2025). Natural language understanding of Devanagari script languages: Language identification, hate speech and its target detection. *In Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHIPSAL 2025)*, 71-82. Association for Computational Linguistics. https://aclanthology.org/2025.chipsal-1.7

Tokhtakhunov, I., Altaibek, A., & Nurtas, M. (2025). Optimizing similar audience search in targeted advertising: Effectiveness of Siamese networks for autoencoder-based user embeddings. *Engineering, Technology & Applied Science Research, 15*(1), 23367-23375. https://doi.org/10.48084/etasr.10527

Tsoumakas, G., Katakis, I., & Vlahavas, I. (2009). Mining multi-label data. In O. Maimon & L. Rokach (Hrsg.), *Data mining and knowledge discovery handbook*, 667-685. Springer. https://doi.org/10.1007/978-0-387-09823-4_34

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics, 7*, Artikel 91. https://doi.org/10.1186/1471-2105-7-91

Vo, D.-T., & Zhang, Y. (2015). Target-dependent twitter sentiment classification with rich automatic features. *In Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15)*, 1347-1353. AAAI Press. https://dl.acm.org/doi/10.5555/2832415.2832437

Vogel, I., Regev, R., & Steinebach, M. (2019). Automatisierte

Analyse radikaler Inhalte im Internet. *In INFORMATIK 2019-50 Jahre Gesellschaft für Informatik*, 233-245. Gesellschaft für Informatik e.V. https://doi.org/10.18420/inf2019_27

Volkovs, M., Cheng, Z., Ravaut, M., Yang, H., Shen, K., Zhou, J. P., Wong, A., Zuberi, S., Zhang, I., Frosst, N., Ngo, H., Chen, C., Venkitesh, B., Gou, S., & Gomez, A. N. (2020). Predicting Twitter engagement with deep language models. *In Proceedings of the Recommender Systems Challenge 2020 (RecSysChallenge '20)*, 38-43. Association for Computing Machinery. https://doi.org/10.1145/3415959.3416000

Wang, X., Wei, F., Liu, X., Zhou, M., & Zhang, M. (2011). Topic sentiment analysis in Twitter: A graph-based hashtag sentiment classification approach. *In Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM '11)*, 1031-1040. Association for Computing Machinery. https://doi.org/10.1145/2063576.2063726

Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. (2012). A system for real-time Twitter sentiment analysis of 2012 U.S. presidential election cycle. *In Proceedings of the ACL 2012 System Demonstrations*, 115-120. Association for Computational Linguistics. https://aclanthology.org/P12-3020

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks, 5*(2), 241-259. https://doi.org/10.1016/S0893-6080(05)80023-1

Xing, B., & Tsang, I. W. (2024). Exploiting contextual target attributes for target sentiment classification. *Journal of Artificial Intelligence Research, 80*, 1243-1280. https://doi.org/10.1613/jair.1.14947

Yadav, V. (2024). Sentiment analysis on Twitter leveraging the power of machine learning methodologies. *International Journal for Multidisciplinary Research, 6*(1), Artikel 12249. https://doi.org/10.36948/ijfmr.2024.v06i01.12249