Emotion Detection

Multi-Label-Klassifikation zur automatischen Erkennung von Emotionen in sozialen Medien

C. Corradini, V. Fuchs, C. Herzel, C. Heubach, L. Richter, A. Sänger, N. Schmalz, C. Schroth

Zusammenfassung

Die Emotionserkennung (*Emotion Detection*) ist ein zentraler Bestandteil der Analyse sozialer Medien, da Emotionen eine Schlüsselrolle in der menschlichen Kommunikation spielen und wertvolle Einblicke in öffentliche Stimmungen und Diskurse ermöglichen. Diese Arbeit untersucht die Klassifikation von Emotionen in Tweets, die im Zusammenhang mit der PEGIDA-Bewegung entstanden sind, wobei ein zweistufiger Ansatz verwendet wurde. Zunächst wurden die Tweets in emotional und nicht-emotional unterteilt, bevor die spezifischen Emotionen in den als emotional klassifizierten Tweets ermittelt wurden. Als Methodik kamen für die binäre Klassifikation der *Multinomial Naive Bayes (MNB)*- und der *CatBoost (CB)*-Algorithmus sowie ein *Voting Classifier (VC)* zum Einsatz. Für die Multi-Label-Klassifikation wurden eine *Support Vector Machine (SVM)* und *Random Forest (RF)* in Kombination mit TF-IDF als Feature verwendet. Die besten Ergebnisse erzielte der VC in der binären Klassifikation mit einem Makro-F1-Maß von ca. 0,76. In der Multi-Label-Klassifikation zeigte die SVM mit einem Makro-F1-Maß von knapp 0,64 die höchste Präzision.

Keywords Emotionserkennung, Tweets, binäre Klassifikation, Multi-Label-Klassifikation, Naive Bayes, TF-IDF, Support Vector Machine, Word2Vec.

1 Einführung

Emotionen sind ein wesentlicher Bestandteil menschlicher Kommunikation und spielen eine wichtige Rolle in Entscheidungsprozessen, Verhaltensweisen sowie sozialen Interaktionen. Sie sind ein relevantes Mittel, um die inneren Gefühle von Menschen auszudrücken, und beeinflussen oft, wie Menschen miteinander reagieren und kommunizieren. Mit der zunehmenden Bedeutung von Mensch-Maschine-Interaktionen und Anwendungen im Bereich künstlicher Intelligenz ist die Emotion Detection zu einem zentralen Forschungsgebiet geworden (Ameer et al., 2023; Nandwani & Verma, 2021). Ziel ist es, emotionale Zustände präzise zu identifizieren, indem unterschiedliche Datenquellen wie Sprache, Gesichtsausdrücke, Stimme oder physiologische Signale analysiert werden (Ekman, 1971).

Ein bedeutsamer Beitrag zur Emotionserkennung stammt von Paul Ekman, dessen Forschung zu den universellen Emotionen weltweit die Grundlage für heutige Klassifikationssysteme liefert (Ekman, 1971). Ekman identifizierte grundlegende Emotionen wie *joy*, *sadness*, *anger*, *fear*, *surprise* und *disgust*, die in allen Kulturen durch ähnliche Gesichtsausdrücke dargestellt und erkannt werden können. Diese universellen Emotionen dienen als Grundlage für Algorithmen, die darauf ausgerichtet sind, emotionale Reaktionen aus Texten, insbesondere aus sozialen Medien wie Twitter, zu extrahieren und zu klassifizieren. Die automatische Erkennung dieser Emotionen hat in den letzten Jahren zunehmend an Bedeutung gewonnen, da Texte aus sozialen Netzwerken wertvolle Einblicke in die Stimmungen und Ansichten von Nutzern geben (Ashraf, 2021). In dieser Arbeit wurden Tweets im Zusammenhang mit der PEGIDA-Bewegung analysiert. PEGIDA, ausgeschrieben als Patriotische Europäer gegen die Islamisierung des Abendlandes, ist 2014 entstanden und eine kontroverse Bewegung, die sich gegen die Islamisierung Europas und die Einwanderungspolitik richtet (Vorländer et al., 2016). Ihre Tweets sind geprägt von einer ablehnenden bis feindseligen Haltung gegenüber diesen Gruppen. Eine Untersuchung dieser bietet wertvolle Einblicke in emotionale Reaktionen auf polarisierende Themen und zeigt, wie Emotionen in sozialen Medien politische und gesellschaftliche Diskurse beeinflussen.

Obwohl Ekmans Theorie wichtige Impulse für die Emotionserkennung bietet, bleibt die Analyse von Texten, insbesondere aus sozialen Medien, eine besondere Herausforderung. Die Ausdrucksweise in diesen Netzwerken ist oft kurz, informell und durch Elemente wie Emojis, Slang oder Abkürzungen geprägt. Diese Eigenschaften erschweren eine präzise Emotionserkennung, da es naheliegt, dass sowohl die Vielfalt der Ausdrucksformen als auch der Kontext der Aussagen die Interpretation emotionaler Zustände beeinflussen. Zudem können Emotionen in Texten komplex und mehrdeutig geäußert werden, was die automatische Erkennung von Gefühlen zusätzlich erschwert.

Um diesen Herausforderungen zu begegnen, wird in dieser Arbeit ein zweistufiger Ansatz vorgestellt, bei dem die Tweets zunächst in emotionale und nicht-emotionale Inhalte unterteilt werden, bevor die spezifischen Emotionen in den emotionalen Tweets identifiziert werden. Ziel ist es, die Emotionen aller Tweets der zur Verfügung gestellten Datensätze möglichst genau zu klassifizieren.

Das Paper gliedert sich wie folgt: Zunächst wird im Literaturreview ein Überblick über die bestehende Forschung zur Emotionserkennung gegeben und die relevanten Ansätze zur Textklassifikation zusammengefasst. Anschließend werden im Kapitel 3 die verwendeten Daten beschrieben und im Kapitel 4 die eingesetzten Methoden, inklusive der Textvorverarbeitung, erklärt. Daran anknüpfend erfolgt in Kapitel 5 eine Darstellung und Diskussion der Ergebnisse. Hierbei werden auch aufgetretene Fehler und Probleme betrachtet. Zum Schluss werden im Kapitel 6 die wichtigsten Ergebnisse kurz zusammengefasst und mögliche zukünftige Verbesserungen und Forschungsansätze erörtert.

2 Literaturreview

In diesem Kapitel werden aktuelle Forschungsansätze und -ergebnisse zur Emotionserkennung zusammengefasst und diskutiert, die für die vorliegende Arbeit und ähnliche Problemstellungen relevant sind. Es wird insbesondere auf Herausforderungen eingegangen, die sich aus der Verarbeitung deutschsprachiger Tweets sowie der Erkennung von komplexeren Emotionen wie *envy* oder *jealousy* ergeben. Diese umfassen differenzierte Gefühle mit starkem sozialen und kontextuellem Bezug und wurden in bisherigen Studien häufig vernachlässigt.

Die Emotionserkennung wird mit einer Vielzahl von Techniken adressiert, die von lexikonbasierten Ansätzen bis hin zu *Deep-Learning (DL)* (Nandwani & Verma, 2021) und Hybrid-

methoden (Ashraf, 2021) reichen. Bei Nandwani und Verma (2021) werden häufige Probleme angesprochen, die mitunter auch für diese Arbeit von Bedeutung sind: Web-Slang, sarkastische und ironische Ausdrücke, implizite und mehrdeutige Emotionsäußerungen sowie die Verwendung von inkorrekter Grammatik.

Klinger (2020) beschreibt in seiner Arbeit die Anwendung von Klassifikatoren zur Sentiment- und Emotionsanalyse wie *Long Short-Term Memory (LSTM)* (Hochreiter & Schmidhuber, 1997), *Bidirectional Encoder Representations from Transformers (BERT)* (Devlin et al., 2019) und SVMs zur präzisen Erfassung von Emotionen und deren Ursachen. Insbesondere Ironie und Sarkasmus erschweren die Emotionserkennung, weshalb für multimodale Ansätze sowie die Erweiterung annotierter Korpora in weniger dokumentierten Sprachen plädiert wird.

Diese Thematik wurde von Bostan und Klinger (2018) vertieft, bei der die Autoren die Eignung verschiedener annotierter Korpora für die Emotionserkennung untersuchten. Sie identifizierten domänenspezifische Unterschiede und uneinheitliche Annotationen, welche die Generalisierbarkeit von Modellen erschweren und die Modellleistung beeinflussen. Auch in dieser Arbeit zeigt sich diese Problematik, da die ungleiche Verteilung der Daten die Leistung des Modells beeinträchtigt.

Während diese Analyse grundlegende Herausforderungen bei der Datengrundlage aufzeigt, untersuchten Sharupa et al. (2020) die Emotionserkennung in Tweets mit dem MNB-Algorithmus, der auch hier für die Klassifikation verwendet wurde. Nach der manuellen Annotation und Vorverarbeitung von 40.000 Tweets erreichten die Autoren eine Präzision von bis zu 70,25 % und betonten die Bedeutung der Vorverarbeitung für die Genauigkeit der Emotionserkennung. Andere Ansätze legten den Schwerpunkt auf sprachliche Mehrdeutigkeiten, wie das Modell von Qamar und Ahmad (2015), das Fuzzy-Logik nutzt. Bei diesem Ansatz werden anstelle binärer Entscheidungen graduelle Werte zwischen 0 und 1 verwendet, um Unsicherheiten und Mehrdeutigkeiten in der Sprache besser zu verarbeiten und Emotionen präziser zu klassifizieren. Dies ist relevant für diese Arbeit, da komplexe Emotionen wie *envy* und *jealousy* in einem mehrsprachigen Kontext präzisere Modelle erfordern. Auch wurde die Gesamtstimmung einer Person, unter Berücksichtigung der Emotionsintensität und zeitlichen Abnahme dieser, bestimmt. Trotz verschiedener Ansätze der eben genannten Studien bleibt die sprachliche Komplexität eine zentrale Herausforderung, die vor allem durch moderne neuronale Netzwerke adressiert wird.

Haryadi und Kusuma (2019) verwendeten verschachtelte LSTM-Netzwerke und erreichen mit über einer Million Trainingsdaten eine Genauigkeit von 99,2 %, was die Leistungsfähigkeit komplexer neuronaler Netzwerke zeigt. Allerdings stellen die erforderliche Rechenleistung und die Modelloptimierung in der Trainingsphase im Rahmen dieser Arbeit eine Herausforderung dar.

Ein weiterer Ansatz, der über klassische LSTM-Netzwerke hinausgeht, ist die Nutzung wissensbasierter Transformer-Modelle. Ein Beispiel hierfür ist der *Topic-Driven Knowledge-Aware Transformer* (Zhu et al., 2021), der themenbasierte Sprachverarbeitung mit „*Common-Sense*"-Wissen kombiniert. Durch ein tieferes Kontextverständnis, welches durch die Integration von thematischen Kontexten und allgemeinem Wissen zustande kommt, verbesserte das Modell die Emotionserkennung und übertraf frühere Ansätze, auch wenn Anpassungen für mehrsprachige Dialoge erforderlich sind.

Ein ähnlicher Fokus auf die Verbesserung der Modellleistung durch Kontextualisierung zeigt sich in der sequentiellen Klassifikation von Sätzen in Dokumenten. Cohan et al. (2019) präsentierten ein BERT-basiertes Modell, das kontextuelle Abhängigkeiten zwischen Sätzen ohne hierarchische Modelle und *Conditional Random Fields* erfasst. Getestet auf wissenschaftlichen Artikeln, die im Vergleich zu Tweets weniger sprachliche Abweichungen und keinen Web-Slang aufweisen, erzielte es überlegene Leistungen bei Aufgaben wie der Dokumentklassifikation und Textzusammenfassung. Jedoch bleibt es in der Generalisierbarkeit auf unterschiedliche Texttypen begrenzt. Für die Multi-Label-Klassifikation könnte das Modell adaptiert werden, erfordert jedoch erhebliche Anpassungen für informelle und mehrsprachige Texte wie Tweets.

Zur Verbesserung der Generalisierbarkeit bietet sich ein universeller Transfer-Learning-Ansatz wie *Universal Language Model Fine-tuning for Text Classification* (Howard & Ruder, 2018) an, der diskriminatives *Fine-Tuning* und *Slanted Triangular Learning Rates* kombiniert. Ein vortrainiertes Sprachmodell wird dabei gezielt für spezifische Aufgaben angepasst, was besonders bei kleineren Datensätzen leistungsstark ist und das *Catastrophic Forgetting* verhindert, wodurch der Lernprozess effizienter wird.

Vielversprechende Ergebnisse bieten auch kombinierte Klassifikationsansätze. Coletta et al. (2014) verbesserten die Sentimentanalyse von Tweets durch die Kombination von SVM und Clustering mithilfe des C3E-SL-Algorithmus, der Klassifikations- und Clustering-Modelle integriert und mit der *Squared Loss* Funktion optimiert wird. Dieser Ansatz zeigte eine höhere Genauigkeit und bessere F1-Werte für positive und negative Klassen als eine reine SVM. Zur Optimierung wurden Clustering-Algorithmen wie *Latent Dirichlet Allocation* oder aktive Lernmethoden vorgeschlagen, die die Komplexität der Emotionserkennung adressieren.

Neuere Modelle verfeinern diese Ansätze durch tiefere Kontextintegration. So entwickelten Zhang et al. (2020) ein mehrschichtiges Modell, das soziale und zeitliche Muster integriert, um gleichzeitig mehrere Emotionen im Kontext eines Nutzers zu erkennen. Basierend auf dem Ekman-Modell (Ekman, 1971) analysiert es emotionale Korrelationen und verwendet einen Multi-Label-Ansatz zur Genauigkeitssteigerung.

Ein weiteres Konzept zur Kontextintegration ist das *Multi-Task Learning*, wie bei Rajamanickam et al. (2020), das Emotionserkennung mit der Detektion missbräuchlicher Sprache verknüpfte. Durch die Nutzung mehrerer bidirektionaler LSTMs und eines *Attention*-Mechanismus wurden wichtige Textabschnitte gezielt gewichtet, um die Reihenfolge und den Kontext der Wörter besser zu erfassen. Ein größerer Kontext könnte jedoch die Emotionserkennung noch weiter verbessern und Fehlklassifikationen reduzieren.

Ein innovativer Ansatz zur Erkennung mehrerer Emotionen in Texten wurde von Ameer et al. (2023) vorgestellt. Hierbei wurden Transfer-Learning-Methoden in Kombination mit *Pretrained Language Models* wie *eXtreme Language Model* (Yang et al., 2019) und *Robustly optimized BERT approach* (Beltagy et al., 2019) eingesetzt, die in Tests mit Datensätzen wie SemEval-2018 (Mohammad et al., 2018) hohe Genauigkeiten erzielten. Das Fine-Tuning für komplexere Emotionen bleibt jedoch eine Herausforderung.

Zudem untersuchten Cordelia und Kokatnoor (2024) die emotionale Bedeutung von Emojis in Nutzerbewertungen. Überwachte Machine-Learning-Modelle wie SVM und *Naive Bayes (NB)* erreichten hohe Genauigkeiten bei der Stimmungsanalyse, wobei eine stärkere Berücksichtigung kultureller Unterschiede in der Emoji-Nutzung empfohlen wird.

Ein weiterer Fortschritt wurde durch DL-Techniken erzielt, wie sie Guo (2022) zur Analyse großer Textmengen vorschlägt. Dafür wurde eine *Deep-Learning-gestützte semantische Textanalyse* eingeführt, die auf *Word-Embeddings* und *Natural Language Processing (NLP)*-Algorithmen aufbaut, um Emotionen in großen Textmengen präzise zu identifizieren. Hier wurde eine Emotionserkennungsrate von 97,2 % und eine Klassifizierungs-

genauigkeit von 98 % erzielt. Dies verdeutlicht das Potenzial von DL zur Analyse großer Datensätze, jedoch bleiben Anpassungen für mehrsprachige Texte und komplexere Emotionen erforderlich.

Trotz fortschrittlicher neuronaler Netzwerke und Transformer-Modelle bleibt eine Forschungslücke in der präzisen Erkennung komplexer Emotionen wie *envy* und *jealousy*. Besonders die Herausforderungen bei kulturellen und kontextuellen Unterschieden sowie die Verbesserung der Modellübertragbarkeit in mehrsprachigen und informellen Texten wie Tweets wurden bisher noch nicht ausreichend thematisiert.

3 Daten

Als Datengrundlage dienten zwei als csv-Dateien zur Verfügung gestellte Datensätze (emo.csv und ems.csv), die mehrsprachige Tweets von Anhängern der PEGIDA-Bewegung beinhalten, welche im Zeitraum von Dezember 2014 bis Juli 2016 gecrawlt wurden.

Der Datensatz für die binäre Klassifikation (emo.csv) umfasst 9097 Tweets, davon 5637 emotionale und 3460 nicht-emotionale. Der Datensatz für die Multi-Label-Klassifikation (ems.csv) beinhaltet 5637 Tweets. Ein Überblick über die Verteilung der einzelnen Emotionen lässt sich aus der Abbildung 1 entnehmen. Die Emotion *anger* ist mit Abstand am häufigsten vorzufinden. Im Gegensatz dazu sind die Emotionen *envy* und *jealousy* kaum in den Tweets vorhanden.

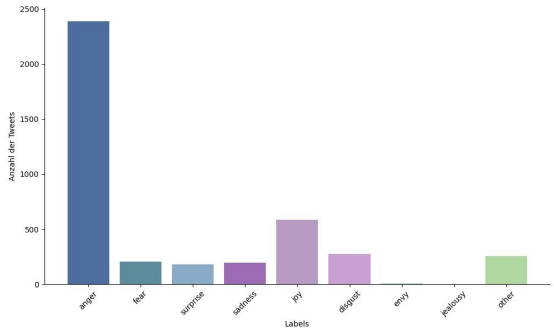

Abbildung 1: Verteilung der Emotionen in der ems.csv-Datei

Bei der Annotation der Tweets ist zunächst entschieden worden, ob eine Emotion zum Ausdruck gebracht worden ist. Ist dies der Fall, wurde weiter in die Basisemotionen *anger*, *fear*, *surprise*, *sadness*, *joy*, *disgust* sowie die Emotionen *envy* und *jealousy* differenziert. Wenn der Tweet eine andere Emotion als die oben genannten ausdrückte, sollte er mit dem Label *other* gekennzeichnet werden.

Zusätzlich wurde ein deutschsprachiger Datensatz (Jikeli et al., 2023) verwendet, der 1675 Tweets für die Multi-Label-Klassifikation umfasst. Die Tweets wurden im Format *<id,description,anger,fear,surprise,sadness,joy,disgust,envy, jealousy,other>* in einer csv-Datei gespeichert. Für jede im Tweet vorkommende Emotion wurde der Wert mit *TRUE* gekennzeichnet, andernfalls mit *FALSE*.

Für die binäre Klassifikation der Tweets wurde ausschließlich die emo.csv-Datei verwendet. Bei der Multi-Label-Klassifikation wurden die annotierten Daten aus dem zusätzlichen Datensatz nur den Trainingsdaten, nicht jedoch den Testdaten, hinzugefügt. Im Baseline-System und SVM sind diese Daten nicht integriert. Der einzige Tweet in der ems.csv-Datei, der das Label *jealousy* enthält, wurde manuell auf insgesamt 33 Einträge vervielfältigt. Diese Erweiterung des Datensatzes ermöglichte die Durchführung der k-fold Cross-Validation.

Durch die Cross-Validation mit 10 Folds ergab sich ein Verhältnis von 90 % Trainingsdaten und 10 % Testdaten, da in jedem Fold neun Teile der Daten für das Training und ein Teil für das Testen verwendet wurden. Dieses Verfahren wurde einheitlich sowohl für binäre als auch für Multi-Label-Klassifikation angewendet.

4 Methoden

Im Folgenden wird die Vorgehensweise dieser Arbeit genauer beschrieben, beginnend mit der Textvorverarbeitung, bei der die Daten für die eigentliche Klassifikation aufbereitet wurden. Darauffolgend wird das Baseline-System zur binären Klassifikation vorgestellt, das eine Unterscheidung von emotionalen und nicht-emotionalen Tweets vornahm. Anschließend wurden die emotionalen Tweets mithilfe des Multi-Label-Klassifikators jeweils allen der neun Labels zugeordnet, die auf diese zutrafen.

4.1 Textvorverarbeitung

Zunächst wurden die Rechtschreibfehler korrigiert, um sicherzustellen, dass fehlerhafte Wörter die Erkennung bestimmter Begriffe nicht erschweren (Ashraf, 2021). Danach wurden die Satzzeichen entfernt, da sie oft wenig zur Bedeutung des Textes beitragen. Weil nur wenige Emojis in den Tweets vorhanden waren und deren Interpretation herausfordernd ist – insbesondere aufgrund der Schwierigkeit, zwischen Emotion und Sarkasmus zu unterscheiden (Bostan & Klinger, 2018) -, wurden sie ebenfalls entfernt. Links wurden durch Platzhalter wie „Link“ ersetzt und normalisiert, da sie keine relevanten Informationen für die Emotionserkennung enthalten (Mohammad et al., 2018). Die Tweets wurden anschließend in kleinere Einheiten, sogenannte „Token“, zerlegt. Dieser Schritt bildet die Grundlage für viele NLP-Modelle und legt die Basis für weitere Verarbeitungsschritte (Ameer et al., 2023). Nach der Tokenisierung erfolgte die Entfernung häufig vorkommender, aber wenig aussagekräftiger Wörter wie Konjunktionen und Artikel. Dies reduziert die Datenmenge und erhöht die Relevanz der verbleibenden Wörter für die Emotionserkennung (Ashraf, 2021). In einem weiteren Schritt wurden die Wörter auf ihre Grundform zurückgeführt, um die Erkennungsgenauigkeit zu verbessern. Schließlich wurden alle Buchstaben in Kleinbuchstaben umgewandelt, um eine einheitliche Darstellung sicherzustellen und die Anzahl einzigartiger Wörter zu verringern. Dadurch wurde die Analyse vereinfacht (Cohan et al., 2019). Zuletzt wurden die vorverarbeiteten Tweets in der Form *<Tweet>, <Label(s)>* in einer Datei gespeichert, um eine weitere Nutzung und Analyse zu ermöglichen.

4.2 Binäre Klassifikation

Für die binäre Klassifikation wurde ein MNB in Kombination mit *Bag-of-Words (BOW)* als Feature eingesetzt, um den Datensatz in emotionale und nicht-emotionale Tweets zu unterteilen. Dieses Baseline-System wurde aufgrund seiner einfachen Implementierung und gleichzeitig guten Leistung gewählt und dient als Grundlage zur Erstellung von Vergleichswerten.

Verbesserungen erfolgten durch den Einsatz des CB-Algorithmus, ebenfalls mit BOW als Feature. CB ist ein fortschrittlicher Gradient Boosting-Algorithmus und besonders leistungsfähig bei der Verarbeitung kategorialer Daten. Er baut iterativ Entscheidungsbäume auf, die die Fehler vorheriger Modelle korrigieren (Dorogush et al., 2018).

Die Ergebnisse wurden zusätzlich durch den Einsatz eines VC mit TF-IDF als Feature verbessert. Diese Ensemble-Methode kombiniert die Stärken von RF, MNB und SVM. Die Auswahl der Modelle begründete sich durch ihre jeweiligen Leistungen:

NB ist simpel und schnell mit guter Baseline-Leistung (Hand et al., 2001), RF erkennt nichtlineare Beziehungen zwischen Daten effektiv (Breiman, 2001) und SVM zeigt auch bei kleinen Datensätzen gute Ergebnisse (Joachims, 1998).

Da die Anzahl der Tweets mit Emotionen (5637) diejenigen ohne Emotionen (3460) übersteigt, wurde zur gleichmäßigen Repräsentation der beiden Klassen *Oversampling* mittels der *Synthetic Minority Over-sampling Technique (SMOTE)*-Methode genutzt (Chawla et al., 2002). SMOTE balanciert das Ungleichgewicht durch künstliche Generierung neuer Datenpunkte in der Klasse mit weniger Instanzen aus und trägt so zu einer verbesserten Trainingsgrundlage bei.

Zur Evaluation wurden die Makro-Precision, der Makro-Recall und das Makro-F1-Maß mittels einer 10-fold Cross-Validation berechnet. Dabei wurde in jedem Fold evaluiert und anschließend der Durchschnitt der Ergebnisse ermittelt.

4.3 Multi-Label-Klassifikation

Als Baseline-System wurde, wie bei der binären Klassifikation, ein MNB verwendet, jedoch mit TF-IDF als Feature. Durch den Einsatz eines *One-vs-Rest*-Klassifikators konnte der MNB einfach auf Multi-Label-Probleme angewendet werden (Hand et al., 2001). Bei diesem Ansatz wurden die Labels unabhängig voneinander analysiert, wobei für jedes Label ein separater NB-Klassifikator trainiert wurde. Diese wurden mitsamt ihrer Vorhersagen am Ende kombiniert. Das Konzept der One-vs-Rest-Klassifikation wurde sowohl beim Baseline-System, als auch bei der SVM, RF und SC angewandt.

Verbesserungen wurden durch den Einsatz einer SVM mit TF-IDF erzielt. Für jede Emotion wurde ein separater SVM-Klassifikator trainiert, um die einzelnen Emotionen präzise und unabhängig voneinander zu lernen.

Als weitere Verbesserung wurde ein RF-Klassifikator eingesetzt, der ebenfalls mit TF-IDF arbeitet und um den zusätzlichen Datensatz erweitert wurde. Da RF gut mit unausgeglichenen Datensätzen umgehen kann, ein geringes Risiko für *Overfitting* aufweist und die Möglichkeit bietet, Klassengewichte zu nutzen, wurde dieser Ansatz gewählt (Breiman, 2001).

Die Anwendung eines *StackingClassifiers (SC)*, der Word2Vec und TF-IDF kombiniert und den zusätzlichen Datensatz nutzt, verbesserte die Ergebnisse zusätzlich. Dabei wurde zunächst ein Word2Vec-Modell separat trainiert und im Anschluss mit der TF-IDF-Repräsentation zusammengefügt. Damit wurde durch Word2Vec der semantische Zusammenhang der Wörter erfasst, während TF-IDF die gewichtete Häufigkeit der Wörter berücksichtigte. Der SC kombiniert die zuvor genutzten Algorithmen, um ihre individuellen Stärken auszunutzen und eine Übereinstimmung zu bilden. Dabei trainieren mehrere Basis-Klassifikatoren mit den gleichen Eingabedaten, lernen jedoch auf unterschiedliche Weise, damit sie die Fehler der anderen Modelle ausgleichen können. Ein Meta-Modell, in diesem Fall ein RF, verarbeitet die Vorhersagen der Basis-Klassifikatoren, um eine finale Entscheidung zu treffen (Zhou, 2012). Ein Nachteil des SC besteht jedoch darin, dass das Training verhältnismäßig viel Zeit beansprucht, ohne dabei signifikant bessere Ergebnisse in der Leistung im Vergleich zu den anderen Systemen zu erzielen (siehe Kapitel 5.2).

Zur Evaluation wurden die Makro-Precision, der Makro-Recall, das Makro-F1-Maß und das Mikro-F1-Maß mittels einer 10-fold Cross-Validation berechnet. Das Mikro-F1-Maß eignet sich besonders für unausgeglichene Datensätze, wie den zusätzlichen Datensatz, da es die Häufigkeit der Labels berücksichtigt. Es wird global berechnet, indem die True-Positive-, False-Positive- und False-Negative-Ergebnisse über alle Klassen hinweg summiert werden, was eine Gewichtung entsprechend der Klassenhäufigkeit ermöglicht (Harbecke et al., 2022).

5 Ergebnisse und Diskussion

Um die Modellleistung zu bewerten, wurden die Metriken Makro-Precision, Makro-Recall und Makro-F1-Maß verwendet. Die Precision berechnet, wie genau die positiven Vorhersagen des Modells im Durchschnitt für jede Klasse sind, unabhängig von der Klassenhäufigkeit. Der Recall misst, wie viele der tatsächlich vorhandenen Emotionen im Durchschnitt korrekt erkannt wurden. Das Makro-F1-Maß kombiniert die Precision und den Recall als harmonischen Mittelwert, um ein ausgewogenes Maß der Modellleistung über alle Klassen hinweg zu bieten (Goutte & Gaussier, 2005).

5.1 Binäre Klassifikation

Das Ziel der binären Klassifikation von Tweets mit dem MNB und BOW war eine präzise Trennung zwischen emotionalen und nicht-emotionalen Tweets. Dies wurde anhand von NB, dem CB-Algorithmus und VC realisiert. Die Ergebnisse der einzelnen Systeme sind der Tabelle 1 zu entnehmen.

Tabelle 1: Konfusionsmatrizen der binären Klassifikation

Modell	Konfusionsmatrix
Naive Bayes	4460 1177 2787 2850
CatBoost	3599 2038 1115 4522
VotingClassifier	4253 1384 1267 4370

Das NB-Modell erzielte eine Precision von 66,1 %, einen Recall von 64,9 % und ein Makro-F1-Maß von knapp 0,64. Diese Ergebnisse dienen als Grundlage für die Bewertung der nachfolgenden Modelle. Es zeigte sich, dass NB eine relativ hohe Anzahl an Fehlklassifikationen aufwies. Zwar wurden 4460 Tweets korrekt als emotional und 2850 richtigerweise als nicht-emotional erkannt, aber 2787 Tweets ohne Emotion wurden fälschlicherweise als emotional eingestuft. Auf der anderen Seite wurden 1177 emotionale Tweets inkorrekt als nicht-emotional klassifiziert. Dies zeigt, dass der Klassifikator Schwierigkeiten hatte, zwischen Emotion und Nicht-Emotion zu differenzieren, was auf Einschränkungen des BOW-Ansatzes zurückzuführen sein könnte. Der BOW-Ansatz berücksichtigt weder Kontext noch semantische Beziehungen, was zu Fehlinterpretationen häufiger Wörter und Negationen führt (Zhixiang et al., 2013). Zudem könnte durch die Annahme der Unabhängigkeit von Wörtern die Modellgenauigkeit beeinträchtigt worden sein.

Durch den Einsatz des CB-Algorithmus konnten deutliche Leistungssteigerungen erzielt werden. Die Precision erhöhte sich auf 73,6 %, der Recall auf 72 % und das Makro-F1-Maß auf knapp 0,72. Der Algorithmus zeigte eine signifikant bessere Leistung als der NB, insbesondere in der präzisen Trennung der beiden Klassen. Es wurden nur 1115 Tweets fälschlicherweise als emotional eingestuft. Bei den emotionalen Tweets wurden jedoch mit 2038 deutlich mehr fälschlicherweise als nicht-emotional klassifiziert, was für einen konservativen Klassifikator spricht. Bei diesem werden Tweets nur bei einer hohen Sicherheit als emotional oder nicht-emotional kategorisiert. Dadurch sind die Ergebnisse nicht ideal. Im Vergleich zum NB-Modell zeigte der CB somit eine deutlich geringere Anzahl an Fehlklassifikationen, insbesondere bei den nicht-emotionalen Tweets.

Der VC erzielte eine zusätzliche Verbesserung der Precision auf 77,3 %, des Recalls auf 76,5 % und des Makro-F1-Maßes auf ca. 0,76. Im Vergleich zu CB klassifizierte der VC insgesamt mehr Tweets korrekt (8623 gegenüber 8121) und wies eine geringere Anzahl an False-Positives auf (1384 im Vergleich zu 2038),

was die gesteigerte Precision erklärt. Der Recall fiel hingegen geringfügig niedriger aus, da die Anzahl der False-Negatives mit 1267 gegenüber 1115 bei CB leicht erhöht war. Insgesamt stellt der VC eine robuste Alternative dar, insbesondere bei einem Schwerpunkt auf der Precision.

Zusammenfassend erzielten der CB-Algorithmus und der VC im Vergleich zum NB signifikant bessere Ergebnisse, da sie die Anzahl der Fehlklassifikationen, insbesondere bei emotionalen Tweets, deutlich reduzierten. Der VC wies zwar einen geringfügig höheren Anteil an falsch positiven Klassifikationen bei nicht-emotionalen Tweets auf, überzeugte jedoch durch seine starke Gesamtleistung, die auf der Kombination verschiedener Modelle basiert. Diese Ergebnisse verdeutlichen die Relevanz fortschrittlicher Modelle und deren Kombination zur Optimierung der Klassifikationsleistung. Zudem trug die Erhöhung der Anzahl der Folds in der k-fold Cross-Validation maßgeblich zur Stabilität und Zuverlässigkeit der Ergebnisse bei, da sie eine präzisere Schätzung der Modellleistung ermöglichte.

5.2 Multi-Label-Klassifikation

In Bezug auf die einzelnen Emotionen zeigten sich deutliche Schwankungen in der Modellleistung, wie auch aus Abbildung 2 hervorgeht.

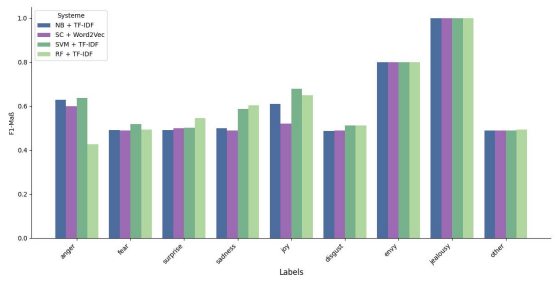

Abbildung 2: Makro-F1-Maße der einzelnen Emotionen

Während *envy* beim NB-Modell mit einem Makro-F1-Maß von knapp 0,80 vergleichsweise gut abschnitt, erreichten *fear*, *surprise*, *sadness*, *disgust* und *other* niedrigere Makro-F1-Maße von 0,48 bis 0,50. Dies weist auf mögliche Schwierigkeiten des Modells bei der Differenzierung dieser Klassen hin. Emotionen wie *anger* und *joy* zeigten mit einem Makro-F1-Maß von 0,63 bzw. 0,61 eine mittlere Leistung. Diese Ergebnisse verdeutlichen die eingeschränkte Leistungsfähigkeit des NB-Modells, insbesondere bei Klassen mit geringem Datenanteil und unausgewogenen Daten.

Durch den Einsatz des SC wurden, wie beim NB, ähnliche Durchschnittswerte für die einzelnen Labels erzielt. Lediglich *anger* und *joy* wiesen mit 0,60 bzw. 0,52 geringfügig schlechtere Makro-F1-Maße auf.

Beim SVM-Modell zeigten Emotionen wie *anger*, *sadness* und *joy* durchschnittliche Makro-F1-Maße von 0,59 bis 0,68, während *surprise* und *disgust* und *other* geringere Ergebnisse erzielten mit einem Makro-F1-Maß unter 0,51. Diese Varianz deutet auf Schwierigkeiten hin, bestimmte Emotionen richtig zu klassifizieren. Jedoch erreichte SVM hinsichtlich *anger*, *fear* und *joy* die beste Klassifikationsrate von allen Systemen.

Beim RF war die verbesserte Leistung des Makro-F1-Maßes bei *sadness* mit rund 0,60 und bei *joy* mit knapp 0,65 auffällig. Im Gegensatz dazu erreichte *anger* beim Makro-F1-Maß nur knapp 0,43. Ursache hierfür könnte der zusätzliche Datensatz sein, der bei der Klassifikation verwendet wurde. Dieser trug zur Verstärkung der Klassenungleichverteilung bei, da er nur von einer Person annotiert wurde.

Hervorzuheben ist, dass das Label *jealousy* in allen Systemen mit einem Makro-F1-Maß von 1,0 optimale Werte für Precision und Recall erzielte. Das System zeigte eine hohe Erkennungsrate dieser Emotion, was jedoch auf Overfitting zurückzuführen ist, da der betreffende Tweet im Datensatz dupliziert wurde. Die Aussagekraft des Ergebnisses ist daher eingeschränkt, da es sich um insgesamt 33 identische Tweets handelt. *Envy* schnitt mit einem hohen Makro-F1-Maß von 0,80 ebenfalls gut ab. Dies lässt darauf schließen, dass diese Emotion aufgrund ihrer Seltenheit im Datensatz vom System möglicherweise „auswendig gelernt“ wurde. Im Gegensatz zu anderen Emotionen sind bei *envy* und *jealousy* die Werte für Precision und Recall sowohl relativ hoch als auch nahezu identisch, was auf eine geringe Variabilität und eine starke Zuordnung der entsprechenden Tweets hinweist.

Tabelle 2: Durchschnittswerte der Multi-Label-Klassifikation

	Makro-P.	**Makro-R.**	**Makro-F1**	**Mikro-F1**
NB	0.6404	0.6125	0.6107	0.9293
SC	0.6400	0.6000	0.6000	0.9200
SVM	0.7196	0.6293	0.6360	0.9301
RF	0.6684	0.6056	0.6138	0.9280

Im Hinblick auf die Durchschnittswerte erreichte das NB-Modell solide Ergebnisse mit einem Makro-F1-Maß von ca. 0,61 und einem Mikro-F1-Maß von knapp 0,93 (siehe Tabelle 2). Durch den Einsatz des SC wurden, wie beim NB, ähnliche Durchschnittswerte für das gesamte System wie auch für die einzelnen Labels erzielt. Das SVM-Modell stach durch die besten Gesamtergebnisse hervor, mit einem Makro-F1-Maß von knapp 0,64 und einem Mikro-F1-Maß von ca. 0,93. Auch die Precision zeigte mit knapp 72 % eine gute Leistung. Das RF-Modell erzielte insgesamt leicht bessere Ergebnisse als das SC-Modell und schnitt durch die Verwendung des zusätzlichen Datensatzes bei den Labels *sadness* und *joy* ebenfalls gut ab. Jedoch blieb es in Bezug auf die Gesamtergebnisse hinter der SVM zurück.

Zusammenfassend zeigte die SVM die beste Gesamtleistung, insbesondere bei *joy*, wie in Abbildung 2 dargestellt. NB und SC lieferten vergleichbare Ergebnisse, wobei Word2Vec bei der Modellierung semantischer Beziehungen möglicherweise einen Vorteil bot. Es zeigte sich zudem, dass seltene Emotionen wie *jealousy* und *envy* besser erkannt wurden, was vermutlich auf eine spezifische Erfassung dieser Merkmale im Modell und möglicherweise auf Overfitting hindeutet.

5.3 Fehlerbetrachtung

Das Hauptproblem in dieser Arbeit war die schlechte Datengrundlage. Es gab insgesamt zu wenige Tweets und die Label-Verteilung war stark unausgeglichen. Insbesondere bei seltenen Emotionen wie *envy* und *jealousy* waren zu wenig Daten vorhanden (siehe Abbildung 1). Dieses Ungleichgewicht erschwerte die Multi-Label-Klassifikation, während die binäre Klassifikation bessere Ergebnisse lieferte.

Der Einsatz von Oversampling mittels SMOTE wurde bei der Multi-Label-Klassifikation zwar ausprobiert, zeigte aufgrund des starken Klassenungleichgewichts jedoch keine Wirkung. Es hätte vermutlich zu Overfitting geführt, da bestehende Tweets lediglich künstlich vervielfältigt worden wären. *Undersampling* war nicht möglich, da die ohnehin geringe Anzahl an Tweets dadurch weiter reduziert und Informationen verloren gegangen wären. Dies hätte zu einer Verringerung der Modellleistung geführt.

Ein weiterer relevanter Punkt ist der zusätzlich verwendete Datensatz (Jikeli et al., 2023), der beim SC und RF für

die Multi-Label-Klassifikation eingesetzt wurde. Dieser konzentriert sich im Gegensatz zur PEGIDA-Bewegung auf Antisemitismus und die Corona-Pandemie. Die Unterschiede in den thematischen Aussagen der Tweets könnten dazu beigetragen haben, dass die Klassifikation von *anger* mit RF vergleichsweise schlechter abschnitt. Insbesondere die variierenden Ausdrucksformen von *anger* in den Tweets könnten die Klassifikation zusätzlich erschwert haben.

Der Ansatz dieser Arbeit bestand darin, für jedes Label einen eigenen Klassifikator zu trainieren, der entscheidet, ob eine bestimmte Emotion in einem Tweet vorhanden ist oder nicht. Diese Vorhersagen wurden anschließend „kombiniert", sodass einem Tweet mehrere Labels zugewiesen werden konnten. Ein Problem, das hierbei jedoch auftrat, war die ungleiche Verteilung der Labels, wodurch für einige Emotionen zu wenige Tweets vorlagen und die Datenmenge nicht ausreichte. Darüber hinaus führte dieser Ansatz dazu, dass die Zusammenhänge zwischen den verschiedenen Labels verloren gingen.

6 Zusammenfassung und Ausblick

In dieser Arbeit wurde ein Ansatz zur Emotionserkennung in Tweets vorgestellt, der sowohl eine binäre Klassifikation (emotionale vs. nicht-emotionale Tweets) als auch eine Multi-Label-Klassifikation (Erkennung spezifischer Emotionen) umfasst. Die verwendeten Methoden, insbesondere MNB und BOW, lieferten trotz begrenzter Datenmenge gute Ergebnisse bei der binären Klassifikation. Allerdings zeigte sich, dass die Multi-Label-Klassifikation durch die unausgeglichene Verteilung der Emotionen sowie die geringe Datenmenge erheblich eingeschränkt wurde.

Zur Verbesserung der Modellleistung kommen mehrere Ansätze in Frage: So könnte *Hyperparameter-Tuning* die Leistung der Klassifikatoren steigern und Under- sowie Overfitting reduzieren. Eine Berücksichtigung emotions- und tweetspezifischer Merkmale wie Satzzeichen, Hashtags und Emojis könnte die Klassifikation verfeinern und die Zusammenhänge zwischen den Labels besser erfassen. Zudem wären durch den Einsatz komplexerer Modelle wie neuronaler Netze und emotionsspezifischer Merkmale wie beispielsweise Sentimentscores und externe Lexika, die zusätzliches Domänenwissen einbringen könnten, weitere Verbesserungen denkbar (Ameer et al., 2023). Eine Liste von Wörtern, die bestimmten Emotionen zugeordnet sind, könnte die Klassifikation unterstützen und zu besseren Ergebnissen führen. Ferner wäre es sinnvoll, die Tweets aus dem zusätzlichen Datensatz von mehr als einer Person annotieren zu lassen und vielfältigere Tweets zu verwenden, damit das Klassenungleichgewicht reduziert werden würde. Dies könnte dann auch bei weiteren Systemen zu einer besseren Leistung führen.

Für zukünftige Arbeiten könnten mehrere Ansätze zur Verbesserung der Ergebnisse beitragen. So könnte die Integration von fortschrittlicheren Modellen wie transformerbasierten Methoden wie beispielsweise BERT dazu beitragen, die semantischen und kontextuellen Beziehungen zwischen Wörtern besser zu erfassen. Ein weiterer Ansatz wäre die Nutzung alternativer Feature-Extraktionstechniken wie der Kombination von TF-IDF mit moderneren Embeddings, um die semantische Tiefe der Textrepräsentation zu erweitern. Hilfreich könnte auch die Erweiterung der Datenmenge durch gezieltes Crawlen und Annotieren sein, um insbesondere seltene Emotionen wie *envy* oder *jealousy* besser zu erfassen. Zusätzlich zu diesen technischen Verbesserungen könnte eine erweiterte Cross-Validation, die sowohl auf k-fold basiert als auch eine gleichmäßige Verteilung der Labels sicherstellt, die Stabilität der Ergebnisse weiter erhöhen.

Literatur

Ameer, I., Bölücü, N., Siddiqui, M. H. F., Can, B., Sidorov, G., & Gelbukh, A. (2023). Multi-label emotion classification in texts using transfer learning. *Expert Systems With Applications*, *213*, 118534. https://doi.org/10.1016/j.eswa.2022.118534

Ashraf, N. M. U. (2021). A Survey on Emotion Detection from Text in Social Media Platforms. *Lahore Garrison University Research Journal of Computer Science and Information Technology*, *5*(2), 48–61. https://doi.org/10.54692/lgurjcsit.2021.0502208

Beltagy, I., Lo, K., & Cohan, A. (2019, November). SciBERT: A Pretrained Language Model for Scientific Text. In K. Inui, J. Jiang, V. Ng & X. Wan (Hrsg.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (S. 3615–3620). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1371

Bostan, L. A. M., & Klinger, R. (2018). An Analysis of Annotated Corpora for Emotion Classification in Text. In E. M. Bender, L. Derczynski & P. Isabelle (Hrsg.), *Proceedings of the 27th International Conference on Computational Linguistics* (S. 2104–2119). Association for Computational Linguistics.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1214/aos/1015955400

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Cohan, A., Beltagy, I., King, D., Dalvi, B., & Weld, D. (2019). Pretrained Language Models for Sequential Sentence Classification. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3693–3699. https://doi.org/10.18653/v1/D19-1373

Coletta, L. F. S., da Silva, N. F. F., Hruschka, E. R., & Hruschka Jr, E. R. (2014). Combining Classification and Clustering for Tweet Sentiment Analysis. *2014 Brazilian Conference on Intelligent Systems*, 210–215. https://doi.org/10.1109/BRACIS.2014.46

Cordelia, A. S., & Kokatnoor, S. A. (2024). Emoji Sentiment Analysis of User Reviews on Online Applications Using Supervised Machine Learning. In S. Shukla, H. Sayama, J. Kureethara & D. Mishra (Hrsg.), *Data Science and Security. IDSCS 2023* (S. 257–267, Bd. 922). Springer. https://doi.org/10.1007/978-981-97-0975-5_23

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: Gradient Boosting with Categorical Features Support. *arXiv preprint arXiv:1810.11363*.

Ekman, P. (1971). Universals and cultural differences in facial expressions of emotion. *Nebraska Symposium on Motivation*, *19*, 207–283.

Goutte, C., & Gaussier, É. (2005). A Probabilistic Interpretation of Precision, Recall and F-Score, with Im-

plication for Evaluation. In D. E. Losada & J. M. Fernández-Luna (Hrsg.), *Advances in Information Retrieval. ECIR 2005* (S. 345–359, Bd. 3408). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-31865-1_25

Guo, J. (2022). Deep Learning Approach to Text Analysis for Human Emotion Detection from Big Data. *Journal of Intelligent Systems*, *31*, 113–126. https://doi.org/10.1515/jisys-2022-0001

Hand, D. J., Mannila, H., & Smyth, P. (2001). *A Study of Machine Learning Algorithms for Text Classification*. Cambridge University Press. https://doi.org/10.1017/CBO9780511755118

Harbecke, D., Chen, Y., Hennig, L., & Alt, C. (2022). Why only Micro-F1? Class Weighting of Measures for Relation Classification. *arXiv preprint arXiv:2205.09460*. https://doi.org/10.48550/arXiv.2205.09460

Haryadi, D., & Kusuma, G. P. (2019). Emotion Detection in Text using Nested Long Short-Term Memory. *International Journal of Advanced Computer Science and Applications*, *10*(6). https://doi.org/10.14569/IJACSA.2019.0100645

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Howard, J., & Ruder, S. (2018). Universal Language Model Fine-Tuning for Text Classification. In I. Gurevych & Y. Miyao (Hrsg.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (S. 328–339). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1031

Jikeli, G., Karali, S., Miehling, D., & Soemer, K. (2023, Oktober). A German Language Labeled Dataset of Tweets [Abgerufen am 12. Dezember 2024]. https://doi.org/10.5281/zenodo.10053509

Joachims, T. (1998). *Support Vector Machines for Text Classification* [Diss., PhD Thesis, University of Stuttgart]. https://doi.org/10.1.1.48.1469

Klinger, R. (2020). *Strukturierte Modellierung von Affekt in Text* [Diss.]. https://doi.org/https://doi.org/10.18419/opus-10994

Mohammad, S., Bravo-Marquez, F., Salameh, M., & Kiritchenko, S. (2018). SemEval-2018 Task 1: Affect in Tweets. *Proceedings of the 12th International Workshop on Semantic Evaluation*, 1–17. https://doi.org/10.18653/v1/S18-1001

Nandwani, P., & Verma, R. (2021). A Review on Sentiment Analysis and Emotion Detection from Text. *Social Network Analysis and Mining*, *11*, 81. https://doi.org/10.1007/s13278-021-00776-6

Qamar, S., & Ahmad, P. (2015). Emotion Detection from Text using Fuzzy Logic. *International Journal of Computer Applications*, *121*(3), 29–32. https://doi.org/10.5120/21522-4501

Rajamanickam, S., Mishra, P., Yannakoudakis, H., & Shutova, E. (2020, Juli). Joint Modelling of Emotion and Abusive Language Detection. In D. Jurafsky, J. Chai, N. Schluter & J. Tetreault (Hrsg.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (S. 4270–4279). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.394

Sharupa, N. A., Rahman, M., Alvi, N., Raihan, M., Islam, A., & Raihan, T. (2020). Emotion Detection of Twitter Post using Multinomial Naive Bayes. *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1–6. https://doi.org/10.1109/ICCCNT49239.2020.9225432

Vorländer, H., Herold, M., & Schäller, S. (2016). *PEGIDA: Entwicklung, Zusammensetzung und Deutung einer Empörungsbewegung* (1. Aufl.). Springer VS Wiesbaden. https://doi.org/https://doi.org/10.1007/978-3-658-10982-0

Yang, Z., Schaal, M., Huang, A., Manning, C. D., & Devlin, J. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Proceedings of the 2019 Conference on Neural Information Processing Systems (NeurIPS 2019)*. https://doi.org/10.5555/3454287.3454777

Zhang, X., Li, W., Ying, H., Li, F., Tang, S., & Lu, S. (2020). Emotion Detection in Online Social Networks: A Multilabel Learning Approach. *IEEE Internet of Things Journal*, *7*(9), 8133–8143. https://doi.org/10.1109/JIOT.2020.3004376

Zhixiang, X., Chen, M., Weinberger, K. Q., & Sha, F. (2013). An alternative text representation to TF-IDF and Bag-of-Words. *arXiv preprint arXiv:1301.6770*. https://doi.org/10.48550/arXiv.1301.6770

Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms*. CRC Press. https://doi.org/10.1201/b12207

Zhu, L., Pergola, G., Gui, L., Zhou, D., & He, Y. (2021). Topic-Driven and Knowledge-Aware Transformer for Dialogue Emotion Detection. In C. Zong, F. Xia, W. Li & R. Navigli (Hrsg.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (S. 1571–1582). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.125