# GermEval 2021: Transformator-basierte Sprache verwenden Modelle zur Identifizierung toxischen, anregenden und faktenorientierter Kommentare

Kevin Fitkau
Fachhochschule Südwestfalen

***Zusammenfassung*—Die Forschung konzentriert sich darauf, wie ich bei der GermEval 2021 die Erkennung von toxischen, anregenden und faktenorientierten Kommentaren angegangen sind. Hierfür griff ich auf frei verfügbare Transformer-basierte Modelle aus dem Huggingface Model Hub zurück. Durch Feinabstimmung mit verschiedenen Hyperparametern der Trainingsdaten bewertete ich die Leistung dieser Modelle. Die Vorhersagen der beiden besten Modelle wurden kombiniert und eingereicht. Als Grundlage diente der Datensatz der GermEval 2021, der über 3.000 Kommentare umfasst. Nach sorgfältiger Evaluation erwies sich Electra als das effektivste Modell, das in der ersten Aufgabe eine F1-Score von 66% und in der zweiten von 68% erreichte. Besonders effektiv erwies sich dieser Ansatz bei der Teilaufgabe 3, wo ich einen beeindruckenden F1-Score von 0.75 erzielten.**

## I. Einleitung

Mit dem Aufkommen und der weit verbreiteten Nutzung sozialer Medien sind nutzergenerierte Inhalte zu einem wesentlichen Bestandteil der Online-Kommunikation geworden. Gleichzeitig haben Plattformen mit der zunehmenden Verbreitung anstößiger Inhalte zu kämpfen. Die Identifizierung toxischer Sprache in sozialen Medien ist zu einem zentralen Anliegen geworden und wird voraussichtlich weiter an Bedeutung gewinnen. Die Forschung zu diesem Thema hat sich auf die Entwicklung von Modellen konzentriert, die verschiedene Formen negativer Inhalte wie Hassrede, Cybermobbing und Missbrauch erkennen können. Im Rahmen von GermEval2021 liegt der Fokus auf der Identifizierung verschiedener Arten von Kommentaren in sozialen Medien. Die gemeinsame Aufgabe des diesjährigen Wettbewerbs ist in drei verschiedene Klassifizierungen unterteilt: toxisch, anregend und faktenbezogen. Basierend auf Erfahrungen aus früheren GermEvals kann davon ausgegangen werden, dass die diesjährigen Aufgaben ähnliche Klassifikationsansätze erfordern. Insbesondere die Identifizierung von Kommentaren, die Leser zur Teilnahme an Diskussionen ermutigen, steht im Mittelpunkt. Daher ist die präzise Klassifizierung dieser Kommentare von entscheidender Bedeutung, da Plattformen kontinuierlich nutzergenerierte Inhalte überwachen und validieren müssen. Zur Lösung dieser Aufgabe wurden BERT und ELECTRA feinabgestimmt. In diesem Dokument werden die Feinabstimmungsmethoden sowie die Ergebnisse für GermEval21 vorgestellt und die Leistung der Transformermodelle anhand des GermEval-2021-Datensatzes bewertet.

## II. Themenbezogene Arbeiten

Die Thematik der Identifizierung anstößiger Sprache in Online-Diskussionen hat in den letzten Jahren stark an Bedeutung gewonnen. Während sich der Großteil der Forschung auf englische Daten konzentriert hat, bedingt durch die Verfügbarkeit annotierter Datensätze, wurden auch Datensätze für anstößige Sprache in anderen Sprachen erstellt und untersucht. Forscher weltweit setzen sich mit dem Problem beleidigender Inhalte in sozialen Medien auseinander, sei es für das Griechische [1], Arabische [2] oder Italienische [3]. Die bisherigen Ansätze zur Bewältigung dieses Problems erstrecken sich von traditionellen maschinellen Lernklassifikatoren wie logistischer Regression und SVMs bis hin zu verschiedenen Deep-Learning-Modellen.

Die Einführung von BERT [4] hat die Nutzung vortrainierter Transformatormodelle zur Klassifizierung anstößiger Sprache vorangetrieben. Die Anwendung vortrainierter BERT-Modelle sowie BERT-basierter Modelle hat gezeigt, dass sie in Wettbewerben konkurrenzfähige Leistungen erbringen können. Es wurden auch sprachspezifische und mehrsprachige Modelle entwickelt, um die NLP-Forschung in verschiedenen Sprachen zu unterstützen, wie zum Beispiel gBERT [5] oder das Modell von Philip May, einer der Autoren von ‘german-nlpgroup/electra-base-german-uncased‘, hat mehrere Modelle mit dem GermEval 2018-Datensatz für das Deutsche evaluiert, AraBERT für Arabisch [6] und das mehrsprachige XLM-R [7], das erfolgreich zur Identifizierung beleidigender Sprache eingesetzt wurde.

## III. GermEval Dataset

Die Modelle wurden auf einem Datensatz deutschsprachiger Tweets trainiert, der im Rahmen des GermEval21 zur Klassifizierung bereitgestellt wurde. Dieser Datensatz umfasst über 3000 Facebook-Kommentare, die von einer Seite einer politischen Talkshow eines deutschen Fernsehsenders extrahiert wurden. Die Trainingsdatensätze bestehen aus etwa 3244 Einträgen, darunter 1074 Datensätze ohne toxische, faktenbezogene oder fesselnde Inhalte. Ein Beispiel für die Trainingsdaten ist in Abbildung 1 dargestellt. Jeder Eintrag in den Trainingsdaten enthält eine Kommentar-ID (comment_id) zur Identifizierung

| comment_id | comment_text | Sub1_Toxic | Sub2_Engaging | Sub3_FactClaiming |
|---|---|---|---|---|
| 1 | Ziemlich traurig diese Kommentare zu lesen. Ihr kÃ¶nnt euch zwar beiÃ¤gen, dass es den v | 0 | 0 | 0 |
| 2 | Sag ich doch, wir befeuern den Klimawandel. Raucher kÃ¶nnen ihr Lebensende meiner Mei | 0 | 1 | 1 |
| 3 | Schublade auf, Schublade zu. Zu mehr Denkleistung reicht es wohl bei dir nicht. | 1 | 0 | 0 |
| 4 | Dummerweise haben wir in der EU und in der USA einen viel hÃ¶heren CO2 FuÃ¤abdruck al | 0 | 0 | 1 |
| 5 | "So lange Gewinnmaximierung Vorrang hat, wird sich das nur schleppend Ändern" Da gebe | 0 | 0 | 0 |

Abbildung 1. Beispiel Tabelle der Datensätze des GermEval21

sowie den Trainingskommentar selbst. Zusätzlich gibt es drei Spalten (Toxic, Engaging und FactClaiming), die jeweils 0 oder 1 enthalten. Eine Eins in einer dieser Spalten zeigt an, dass der Kommentar zu dem entsprechenden Bereich gehört, während eine Null das Gegenteil angibt.

## IV. Aufgabenbeschreibung

Die Teilnehmer durften an einer, zwei oder allen drei Teilaufgaben teilnehmen und maximal drei Läufe pro Aufgabe einreichen. Die Aufgaben bestehen aus „Toxic Comment“-Klassifizierung, „Engaging Comment“-Klassifizierung und die letzte Aufgabe ist „Fact-Claming Comment“-Klassifizierung

### *A. „Toxic Comment“-Klassifizierung*

Die erste Teilaufgabe bestand darin, die Identifizierung von giftigen oder toxischen Kommentaren in Online-Diskussionen, die potenziell Leser beleidigen oder verletzen könnten. Diese Unteraufgabe baut auf früheren GermEval-Aufgaben zur Erkennung anstößiger Sprache auf und wurde in dieser Studie weitergeführt

| message | Sub1_Toxic |
|---|---|
| *“Na, welchem tech riesen hat er seine Eier verkauft..?”* | 1 |
| *“Ich macht mich wütend, dass niemand den Schülerinnen Gehör schenkt”* | 0 |

Abbildung 2. Beispiel der toxischen Kommentare

### *B. „Engaging Comment“-Klassifizierung*

Zusätzlich zur Erkennung von giftiger Sprache zeigen Community-Manager und -Moderatoren zunehmendes Interesse daran, besonders wertvolle Inhalte von Nutzern zu identifizieren. Dies kann beinhalten, vernünftige, respektvolle und gegenseitige Kommentare hervorzuheben, um ihnen mehr Sichtbarkeit zu verleihen. Solche Kommentare ermutigen die Leser, sich an der Diskussion zu beteiligen, verbessern die positive Wahrnehmung der Diskussionsanbieter und fördern einen fruchtbareren und weniger gewalttätigen Austausch. Daher war die zweite Teilaufgabe darauf ausgerichtet, genau solche Kommentare zu identifizieren, die die Leser zur Teilnahme an den Gesprächen ermutigen könnten.

| message | Sub2_Engaging |
|---|---|
| *“Wie wär’s mit einer Kostenteilung. Schließlich haben beide Parteien (Verkäufer und Käufer) etwas von der Tätigkeit des Maklers. Gilt gleichermassen für Vermietungen. Die Kosten werden so oder soweit verrechnet, eine Kostenreduktion ist somit nicht zu erwarten.”* | 1 |
| *“Die aktuelle Situation zeigt vor allem eines: viele Kinder mussten erkennen, dass ihre Mutter bestenfalls das Niveau Grundschule, Klasse 3 haben.”* | 0 |

Abbildung 3. Beispiel der „Engaging Comment“-Kommentare

### *C. „Fact-Claiming Comment“-Klassifizierung*

Zusätzlich zur Gewährleistung nicht-feindseliger Debatten stehen Plattformen und Moderatoren vor der Herausforderung, der raschen Verbreitung von Fehlinformationen und Fake News entgegenzuwirken. Sie sind daher unter Druck, gepostete Informationen zu überprüfen und zu verifizieren, um ihrer Verantwortung als Informationsanbieter und -verbreiter gerecht zu werden. Aus diesem Grund beinhaltete die letzte Teilaufgabe die Identifizierung von faktenbezogenen Kommentaren. Dabei sollte jedoch nicht die Richtigkeit der Kommentare selbst überprüft werden.

| message | Sub3_FactClaiming |
|---|---|
| *“Kinder werden nicht nur seltener krank, sie infizieren sich wohl auch seltener mit dem Coronavirus als ihre Eltern - das ist laut Ministerpräsident Winfried Kretschmann (Grüne) das Zwischenergebnis einer Untersuchung der Unikliniken Heidelberg, Freiburg und Tübingen.”* | 1 |
| *“hmm...das kann ich jetzt nich nachvollziehen...”* | 0 |

Abbildung 4. Beispiel der faktenbezogenen Kommentare

### *D. Bewertungsmetriken*

Die Klassifizierungsleistung wird anhand der gängigen Bewertungsmaßstäbe Präzision, Recall und F-Score bewertet. Diese Maße werden für jede der einzelnen Klassen in den drei Teilaufgaben berechnet. Zusätzlich zu den klassenbezogenen Metriken werden für jede Aufgabe die makrodurchschnittliche „Präzision“, der „Recall“ und der

„F-Score“ sowie die Gesamtgenauigkeit berechnet. Beachten Sie jedoch, dass die Genauigkeit nicht im Bewertungsrang verwendet wird und daher nur anhand von WANDB-Informationen approximiert wird. Um die Daten zu evaluieren, musste eine ZIP-Datei mit den erforderlichen Teilaufgaben auf Codalab hochgeladen werden.

## V. Fine-Tuning gBERT and Electra

In diesem Abschnitt erläutere ich das Feintuning sowohl des deutschsprachigen BERT- als auch des Electra-Modells. Zudem präsentiere ich meine Einstellungen für diese Modelle.

### A. gBert

Ein Ansatz bestand darin, ein deutschsprachiges BERT-Modell zu verwenden, das die Groß- und Kleinschreibung berücksichtigt. Dieses Modell wurde mit 12 GB an Rohdaten aus dem deutschsprachigen Wikipedia-Dump, dem OpenLegalData-Dump und Nachrichtenartikeln trainiert. Es hat die gleiche Größe wie das englischsprachige BERT-Modell mit 12 Schichten, 768 Hidden Units und 12 Attention Heads. Das Modell umfasst insgesamt 110 Millionen Parameter [5].
Für das Feintuning von BERT wurde eine Batch-Größe von 8 verwendet. Der Adam-Optimizer mit einer Lernrate von 1e-5 wurde angewendet. Die Evaluationsschritte wurden alle 250 Schritte durchgeführt, und die maximale Sequenzlänge betrug 128 Token. Das Training wurde für eine bis fünf Epochen durchgeführt, wobei festgestellt wurde, dass bei mehr Epochen keine weiteren Verbesserungen erzielt wurden. Die maximale Anzahl der Trainingsdurchläufe wurde auf 1500 festgelegt. Es wurde festgestellt, dass bei höheren Einstellungen eine Verschlechterung des F1-Scores und der Genauigkeit auftrat. BERT verwendet tokenisierte Teile von Wörtern anstelle von tokenisierten Wörtern. Hier sind zwei Beispiele für die Tokenisierung:

Beispiel 1:

- Text: „Sag ich doch, wir befeuern den Klimawandel. Raucher können ihr Lebensende meiner Meinung nach auch gerne befeuern, nur hab ich daran kein Interesse.“
- Token: ['Sa', '##g', 'ich', 'doch', ',', 'wir', 'bef', '##euer', '##n', 'den', 'Klima', '##wandel', '.', 'Rauch', '##er', 'können', 'ihr', 'Lebens', '##ende', 'meiner', 'Meinung', 'nach', 'auch', 'gerne', 'bef', '##euer', '##n', ',', 'nur', 'hab', 'ich', 'daran', 'kein', 'Interesse', '.']

Dabei ist zu beobachten, dass das Tokenisierungssymbol ## die Zusammengehörigkeit der Wörter signalisiert und sie zu einem Token zusammenfasst.

Beispiel 2:

- Text: „Ziemlich traurig diese Kommentare zu lesen. Ihr kÃ¶nnt euch zwar belÃ¼gen, dass es den vom Menschen gemachten Klimawandel nicht gibt, nur kann man die Natur nicht belÃ¼gen. Wie viele Menschen mÃ¼ssen denn noch auf Grund des Klimawandels ihre Lebensgrundlage verlieren oder gar Sterben, bis ihr den ernst der Lage erkannt habt?“
- Token: ['Zie', '##mlich', 'traur', '##ig', 'diese', 'Kommentare', 'zu', 'lesen', '.', 'Ihr', '[UNK]', '[UNK]', 'n', '##nt', 'euch', 'zwar', '[UNK]', ',', 'dass', 'es', 'den', 'vom', 'Menschen', 'gemachten', 'Klima', '##wandel', 'nicht', 'gibt', ',', 'nur', 'kann', 'man', 'die', 'Natur', 'nicht', '[UNK]', '.', 'Wie', 'viele', 'Menschen', '[UNK]', 'denn', 'noch', 'auf', 'Grund', 'des', 'Klima', '##wandel', '##s', 'ihre', 'Lebens', '##grundlage', 'verlieren', 'oder', 'gar', 'Sterbe', '##n', ',', 'bis', 'ihr', 'den', 'ernst', 'der', 'Lage', 'erkannt', 'hab', '##t', '?']

Bitte beachten Sie, dass der zweite Satz einen Formatierungsfehler aufweist, der als unbekanntes Symbol behandelt wird. Dies kann dazu führen, dass der Klassifikator Wörter nicht korrekt erkennt. Darüber hinaus sollte der Tokenizer in der Lage sein, Wörter, die in Großbuchstaben geschrieben sind, korrekt zu behandeln. Abkürzungen und Großbuchstaben müssen jedoch vom Tokenizer erlernt werden, wie zum Beispiel „DE“ für Deutsch. Normalerweise sind solche Abkürzungen bereits im Modell des Modellerstellers integriert, da sie häufig in den Pre-Trainingsdaten für das deutsche BERT-Modell vorkommen.

### B. Electra

Das Electra-Modell kann als Alternative zu BERT in den meisten Aufgaben verwendet werden. Es ist als erweiterte BERT-Klasse implementiert und wurde mit 12 GB Rohdaten aus dem deutschsprachigen Wikipedia-Dump, OpenLegalData-Dump und Nachrichtenartikeln trainiert. Das Modell weist die gleiche Größe wie das englisch- und deutschsprachige BERT-Modell auf, mit 12 Schichten, 768 versteckten Einheiten und 12 Attention-Heads. Es umfasst auch 110 Millionen Parameter wie BERT. Allerdings zielt Electra darauf ab, Rechenzeit und Ressourcen zu reduzieren, während gleichzeitig eine hohe Leistung beibehalten wird. Die Pre-Training-Aufgabe in Electra basiert auf der Erkennung ersetzter Token in der Eingabesequenz. Für diese Aufgabe werden zwei Transformer-Modelle benötigt: ein Generator und ein Diskriminator. Für die Feinab-

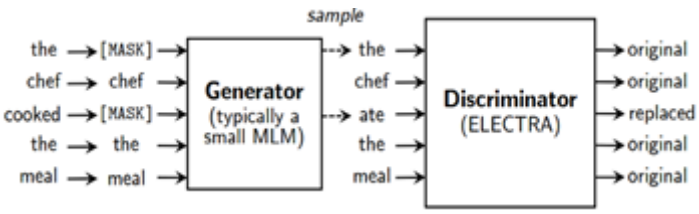


Abbildung 5. ELECTRA-Trainingsprozess

stimmung von Electra wurden die gleichen Einstellungen wie für gBERT verwendet. Die Batch-Größe betrug 8, der Adam-Optimizer mit einer Lernrate von 1e-5 wurde angewendet, und die Evaluationsschritte erfolgten alle 250 Schritte. Die maximale Sequenzlänge wurde auf 128 Token festgelegt. Das Training wurde für eine bis fünf Epochen durchgeführt, wobei festgestellt wurde, dass bei mehr Epochen keine weiteren Verbesserungen erzielt wurden. Die maximale Anzahl der Trainingsdurchläufe wurde auf

1500 festgelegt, wobei herausgefunden wurde, dass diese Anzahl bei einer Dauer von 5 Epochen ausreichen sollte. Es wurde ebenfalls festgestellt, dass höhere Einstellungen zu einer Verschlechterung des F1-Scores und der Genauigkeit führen.

Beispiel 1:

- Text: „Sag ich doch, wir befeuern den Klimawandel. Raucher können ihr Lebensende meiner Meinung nach auch gerne befeuern, nur hab ich daran kein Interesse.“
- Token: ['sag', 'ich', 'doch', ',', 'wir', 'befe', '##uern', 'den', 'klimawandel', '.', 'raucher', 'können', 'ihr', 'lebens', '##e', '##nde', 'meiner', 'meinung', 'nach', 'auch', 'gerne', 'befe', '##uern', ',', 'nur', 'hab', 'ich', 'daran', 'kein', 'interesse', '.']

Es ist zu bemerken, dass das Raute-Symbol die Zusammengehörigkeit der Wörter signalisiert, jedoch eine andere Zugehörigkeit als gBERT anzeigt.

Beispiel 2:

- Text: „Ziemlich traurig diese Kommentare zu lesen. Ihr kÃ¶nnt euch zwar belÃ¼gen, dass es den vom Menschen gemachten Klimawandel nicht gibt, nur kann man die Natur nicht belÃ¼gen. Wie viele Menschen mÃ¼ssen denn noch auf Grund des Klimawandels ihre Lebensgrundlage verlieren oder gar Sterben, bis ihr den ernst der Lage erkannt habt?“
- Token: ['ziemlich', 'traurig', 'diese', 'kommentare', 'zu', 'lesen', '.', 'ihr', 'k', '##ã', '¶', 'nn', '##t', 'euch', 'zwar', 'bel', '##ã', '##¼', '##gen', ',', 'dass', 'es', 'den', 'vom', 'menschen', 'gemachten', 'klimawandel', 'nicht', 'gibt', ',', 'nur', 'kann', 'man', 'die', 'natur', 'nicht', 'bel', '##ã', '##¼', '##gen', '.', 'wie', 'viele', 'menschen', 'm', '##ã', '##¼', '##ssen', 'denn', 'noch', 'auf', 'grund', 'des', 'klimawandels', 'ihre', 'lebens', '##grundlage', 'verlieren', 'oder', 'gar', 'sterben', ',', 'bis', 'ihr', 'den', 'ernst', 'der', 'lage', 'erkannt', 'habt', '?']

## VI. Ergebnisse

Die Ergebnisse unterteile ich in zwei Kategorien: Testergebnisse, die auf WANDB hochgeladen und verarbeitet wurden, sowie Codalab-Ergebnisse, die zum GermEval2021-Wettbewerb gehören und dort evaluiert wurden. WANDB steht für „Weights & Biases“ und ist eine Plattform zur Verfolgung von Experimenten und zur Visualisierung von Modelltrainings. Es ermöglicht Forschern und Entwicklern, ihre Experimente zu organisieren, Metriken zu verfolgen, Modelle zu vergleichen und Ergebnisse zu visualisieren. Codalab ist eine Plattform für wissenschaftliche Wettbewerbe, auf der Forscher ihre Modelle hochladen und auf standardisierten Datensätzen testen können. Es wird oft für NLP- und Machine-Learning-Wettbewerbe verwendet, bei denen Teilnehmer ihre Modelle auf einem vorgegebenen Testdatensatz evaluieren und vergleichen können. Für die privaten Ergebnisse habe ich WANDB genutzt, während ich für die öffentlichen Ergebnisse auf Codalab zurückgegriffen habe.

### A. WANDB Ergebnisse

Bevor ich meine Ergebnisse bewertete, habe ich meine Läufe mit WANDB hochgeladen, um die Evaluation für das finale Ergebnis zu testen. Dabei habe ich die Testergebnisse sowohl mit dem gBERT-Modell als auch mit dem Electra-Modell evaluiert. Dabei stellte ich fest, dass meine besten Ergebnisse mit dem Electra-Modell erzielt wurden: Der F1-Score betrug 70% für Engaging-, 57% für Toxic- und 70% für Fact-Claiming-Klassifizierung. Allerdings war das Fact-Claiming-Ergebnis mit 70% beim Electra-Modell schlechter als beim gBERT-Modell. Letzteres schnitt insgesamt nur leicht schlechter ab als das Electra-Modell mit den gleichen Einstellungen. Es erreichte einen F1-Score von 65% für Engaging-, 54% für Toxic- und 71% für die Fact-Claiming-Klassifizierung. Die Genauigkeit der beiden Modelle lag zwischen 78% und 83%, während die Präzision zwischen 83% und 89% variierte. Der Recall bewegte sich im gleichen Bereich wie der F1-Score bei den unterschiedlichen Läufen.

### B. Codalab Ergebnisse

Im Folgenden präsentiere ich die Ergebnisse der Klassifizierungsaufgaben unter Verwendung des Electra-Modells für die deutsche Sprache. Aufgrund der überlegenen Leistung des Electra-Modells im Vergleich zum gBERT-Modell habe ich nur die Ergebnisse des Electra-Modells in der CSV-Datei hochgeladen. Nachdem ich die Antwortdatei (answer.CSV) kurz vor Ablauf der Abgabefrist erfolgreich eingereicht hatte, erhielt ich innerhalb weniger Tage die Evaluierungsergebnisse zurück. Mein herausragendes Ergebnis erzielte ich bei der abschließenden Klassifizierungsaufgabe, bei der faktenbezogene Kommentare klassifiziert wurden. Dabei erreichte ich einen F1-Score, eine Präzision und einen Recall von jeweils 75%. Dies ist in Abbildung 6 oben dargestellt.

| Sub1_F1 | Sub1_P | Sub1_R | Sub2_F1 | Sub2_P | Sub2_R | Sub3_F1 | Sub3_P | Sub3_R |
|---|---|---|---|---|---|---|---|---|
| 0,6601 | 0,6778 | 0,6433 | 0,6849 | 0,6791 | 0,6909 | 0,7507 | 0,7495 | 0,7519 |

Abbildung 6. Ergebnise GermEval2021

## VII. Fazit

In diesem Beitrag präsentiere ich meine Teilnahme an GermEval-2021, einer Veranstaltung, die den Teilnehmern die Möglichkeit bietet, computergestützte Modelle zur Identifizierung von toxischen, anregenden und faktenbezogenen Kommentaren zu testen. In meinem Experiment habe ich verschiedene neuronale Transformatormodelle wie gBert und Electra eingesetzt. Dabei konnte ich feststellen, dass Electra in Bezug auf toxische und anregende Kommentare eine bessere Leistung als gBert erzielte. Bei faktenbezogenen Kommentaren war hingegen gBert überlegen. Für zukünftige Forschungen sollten weitere Modelle getestet werden, da sich auch andere Transformatormodelle stetig verbessern. Außerdem könnten alternative Ansätze zur Datenverarbeitung die Leistung weiter steigern. Es wäre auch interessant, ähnliche Aufgaben in anderen Sprachen

zu bearbeiten und Modelle zu testen, die mit umfangreichen Testdaten trainiert wurden. Ebenso lohnt es sich, die Entwicklung von Transformatormodellen in anderen Ländern aufmerksam zu verfolgen.

## Referenzen


[1] Z. Pitenis, M. Zampieri und T. Ranasinghe, „Offensive Language Identification in Greek," English, in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache u. a., Hrsg., Marseille, France: European Language Resources Association, Mai 2020, S. 5113–5119, ISBN: 979-10-95546-34-4. Adresse: https://aclanthology.org/2020.lrec-1.629.

[2] H. Mubarak, A. Rashed, K. Darwish, Y. Samih und A. Abdelali, „Arabic Offensive Language on Twitter: Analysis and Experiments," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, N. Habash, H. Bouamor, H. Hajj u. a., Hrsg., Kyiv, Ukraine (Virtual): Association for Computational Linguistics, Apr. 2021, S. 126–135. Adresse: https://aclanthology.org/2021.wanlp-1.13.

[3] P. Chiril, F. Benamara Zitoune, V. Moriceau, M. Coulomb-Gully und A. Kumar, „Multilingual and Multitarget Hate Speech Detection in Tweets," in *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume II : Articles courts*, E. Morin, S. Rosset und P. Zweigenbaum, Hrsg., Toulouse, France: ATALA, Juli 2019, S. 351–360. Adresse: https://aclanthology.org/2019.jeptalnrecital-court.21.

[4] J. Devlin, M.-W. Chang, K. Lee und K. Toutanova, „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *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)*, J. Burstein, C. Doran und T. Solorio, Hrsg., Minneapolis, Minnesota: Association for Computational Linguistics, Juni 2019, S. 4171–4186. DOI: 10.18653/v1/N19-1423. Adresse: https://aclanthology.org/N19-1423.

[5] B. Chan, S. Schweter und T. Möller, „German's Next Language Model," in *Proceedings of the 28th International Conference on Computational Linguistics*, D. Scott, N. Bel und C. Zong, Hrsg., Barcelona, Spain (Online): International Committee on Computational Linguistics, Dez. 2020, S. 6788–6796. DOI: 10.18653/v1/2020.coling-main.598. Adresse: https://aclanthology.org/2020.coling-main.598.

[6] W. Antoun, F. Baly und H. Hajj, „AraBERT: Transformer-based Model for Arabic Language Understanding," English, in *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, H. Al-Khalifa, W. Magdy, K. Darwish, T. Elsayed und H. Mubarak, Hrsg., Marseille, France: European Language Resource Association, Mai 2020, S. 9–15, ISBN: 979-10-95546-51-1. Adresse: https://aclanthology.org/2020.osact-1.2.

[7] A. Conneau, K. Khandelwal, N. Goyal u. a., *Unsupervised Cross-lingual Representation Learning at Scale*, 2020. arXiv: 1911.02116 [cs.CL].