Evaluation computergestützter Verfahren der Emotionsklassifikation für deutschsprachige Dramen um 1800

Schmidt, Thomas
thomas.schmidt@ur.de
Lehrstuhl für Medieninformatik, Universität Regensburg

Dennerlein, Katrin
katrin.dennerlein@uni-wuerzburg.de
Institut für Deutsche Philologie, JMU Würzburg

Wolff, Christian
christian.wolff@ur.de
Lehrstuhl für Medieninformatik, Universität Regensburg

Einleitung

Transformerbasierte Sprachmodelle wie BERT (Devlin et al. 2018) und ELECTRA (Clark et al. 2020) gelten als state-of-the-art und Ausgangspunkt für zahlreiche Aufgaben des Natural Language Processing (NLP) (Shmueli / Ku 2019; Munikar et al. 2019; Cao et al. 2020; Dang et al. 2020; Gonzáles-Carvajal et al. 2021; Cortiz 2021). Als ein entscheidender Vorteil dieser Modelle hat sich die dynamische Repräsentation von Tokens in Abhängigkeit von ihrem Kontext herausgestellt. Der Großteil dieser Modelle wird jedoch mit zeitgenössischer Sprache, vor allem mit Sach- und Fachtexten aus dem Web (z.B. *Wikipedia*) trainiert. Dies stellt ein Problem für Forschungsbereiche wie die Digital Humanities (DH) dar, die mit literarischen Texten arbeiten. Literarische Texte unterscheiden sich entscheidend von Textsorten wie Wikipedia-Artikeln, weil sie fiktional sind und Sprache kreativ und ästhetisch motiviert verwenden. Mit literarischen Texten wird zudem häufig nicht explizit, sondern indirekt durch Bilder kommuniziert. Entwicklungen im Bereich der Domänenadaptation ermöglichen jedoch auch die Optimierung transformerbasierter Modelle auf spezielle Domänen, was Projekte auch im deutschsprachigen Bereich bereits gewinnbringend nutzen konnten (Labusch et al. 2019; Schweter / Baiter 2019; Brunner et al. 2020; Schweter / März 2020). Für die Aufgabe der Emotionsklassifikation findet man im englischsprachigen Bereich Studien, die derartige Methoden für zeitgenössische Texte explorieren (Shmueli / Ku 2019; Acheampong et al. 2020; Cao et al. 2020).

In den Digital Humanities (DH) werden Sentiment-Analyse (die Einteilung, ob ein Text eher positiv/negativ konnotiert ist) und Emotionsklassifikation (die Erkennung bzw. Zuordnung distinkter Emotionskonzepte in Texten) in den letzten Jahren immer populärer. Sie werden verwendet, um moderne Textsorten wie Songtexte (Schmidt et al. 2020a), Filmtexte (Schmidt et al. 2020b) und Texte aus den sozialen Medien zu analysieren (Moßburger et al. 2020; Schmidt et al. 2020c; 2020d) finden aber auch Einsatz für literarische Genres wie beispielsweise Märchen (Alm / Sproat 2005; Mohammad 2011), Romane (Kakkonen / Kakkonen 2011; Mohammad et al. 2011; Reagan et al. 2016; Zehe et al. 2016) oder Dramen (Mohammad 2011; Schmidt / Burghardt 2018; Schmidt et al. 2018a; 2018b; Schmidt 2019; Schmidt et al. 2019a; 2019b; 2019c; Yavuz 2020; Schmidt et al. 2021). Die Ziele variieren dabei von der Exploration von Sentiment- und Emotionsverläufen in einzelnen Werken bis zu Gruppenvergleichen (siehe Kim / Klinger 2019). Die steigende Popularität ist wenig überraschend, da die hermeneutische Analyse von Emotionen eine lange Tradition in der Literaturwissenschaft hat, z.B. in der Dramenanalyse (Pikulik 1965; Wiegmann 1987; Anz 2011; Schonlau 2017).

Im folgenden Proposal präsentieren wir eine Studie aus dem DFG-Projekt *Emotions in Drama*[1] zur Evaluation von Methoden transformerbasierter Emotionsklassifikation für ein annotiertes Korpus historischer deutschsprachiger Dramentexte. Unser Ziel ist, es die Leistung verschiedener Verfahren zu vergleichen und Impulse für Optimierungen auf dieser Textsorte zu sammeln. Im nächsten Kapitel wird dazu zunächst in das verwendete Annotationsschema sowie das annotierte Goldstandard-Korpus eingeführt.[2] Danach werden die verwendeten Klassifikationsverfahren erläutert. Aktuelle Verfahren werden dabei mit bekannten Baseline-Methoden verglichen und für verschiedene Kategorienmodelle evaluiert. Abschließend werden die Ergebnisse der Evaluation präsentiert.

Annotation und Goldstandard-Erstellung

Zur Evaluation und zum Training von Algorithmen wurde ein Goldstandard für ein Sub-Korpus unseres Gesamtkorpus‘ annotiert.

Definitionen und Annotationsschema

Emotion wird definiert als der Bewusstseinszustand einer Figur, wie sie sich auch in Text ausdrückt. Annotiert wird die eigene oder zugeschriebene Emotion von Figuren in Abhängigkeit von Kontext und Interpretation. Das Schema hebt sich von üblichen Schemata, die meist von der Psychologie inspiriert sind (Wood et al. 2018a; 2018b) ab, um literarische Interessen zu integrieren. Es besteht aus 13 *Sub-Emotionen*, die sich in sechs *Hauptklassen* unterteilen lassen und weiter in die *Polarität* (positiv/negativ) auf höchster Ebene. Abbildung 1 (Kapitel *Annotationsergebnisse*) illustriert die einzelnen Konzepte.

Ein Sonderfall des Schemas ist *emotionale Bewegtheit*, die verwendet wird, um unspezifische emotionale Erregungen zu markieren. Zusammen mit den Klassen *negativ*/ *positiv* bezeichnen wir diese Sammlung an Oberkategorien als *Dreifach-Polarität*. Es werden sowohl Repliken (einzelne Sprechakte von Figuren) als auch Regieanweisungen annotiert, sofern Annotator*innen dort Emotionen erkennen. Annotator*innen können variable Textlängen pro Einheit annotieren, also einzelne Wörter, Satzteile und mehrere Sätze. Annotationen können sich zudem überlappen. Obwohl es Vorteile hat, feste Annotationseinheiten festzulegen, wurde dieser variable Annotationsstil basierend auf der Erfahrung von Pilotstudien bestimmt.

Annotiertes Teilkorpus

Das zu analysierende Hauptkorpus unseres Gesamtprojektes setzt sich aus unterschiedlichen Dramenkollektionen für die Jahre 1650-1815 aus TextGrid[3], GerDracor (Fischer et al. 2019) und anderen Quellen zusammen. Für die vorliegende Studie wurde eine repräsentative Menge von Dramen, gemessen an Sprache und Genre für die Zeit um 1800, gewählt: *Minna von Barnhelm* (1767, Lessing, Komödie), *Kabale und Liebe* (1784, Schiller, Tragödie), *Kasperl' der Mandolettikrämer* (1789, Eberl, Komödie), *Menschenhass und Reue* (1790, Kotezbue, Komödie), *Faust. Eine Tragödie* (1807, Goethe, Tragödie).

Annotationsprozess

Für die Annotation wurde das Tool CATMA (Gius et al. 2020) verwendet. Die Dramen wurden vollständig von Anfang bis Ende annotiert. Die Lektüre des gesamten Dramas ist notwendig, da kontextabhängig annotiert wird. Je zwei studentische Hilfskräfte haben jedes Werk unabhängig voneinander annotiert. Die Hilfskräfte wurden vor der Annotation mittels Pilotstudien von einer Expertenannotatorin trainiert und hatten Zugriff auf eine Annotationsanleitung. Je nach Länge des Textes hatten die Annotator*innen 1-2 Wochen Zeit pro Drama.

Annotationsergebnisse

Der Goldstandard besteht insgesamt aus 6.596 Emotionsannotationen (Abbildung 1).

Hauptklassen und Sub-Emotionen	absolut	%	Avg: tokens	Min: tokens	Max: tokens	Std: tokens
HK: Emotionen der Zuneigung	**1 266**	**19**	**24,05**	**1**	**326**	**28,61**
Lust (-)	50	1	23,22	4	83	16,49
Liebe (+)	783	12	26,16	1	326	33,67
Freundschaft (+)	127	2	22	1	120	18,66
Verehrung (+)	306	5	19,63	1	96	16,36
HK: Emotionen der Freude	**1 051**	**16**	**23,21**	**1**	**223**	**23,86**
Freude (+)	850	13	22,78	1	223	24,3
Schadenfreude (+)	201	3	25,02	1	121	21,89
HK: Emotionen der Angst	**706**	**11**	**22,42**	**1**	**206**	**24,32**
Angst (-)	424	7	16,87	1	173	17,45
Verzweiflung (-)	282	4	30,78	1	206	30,15
HK: Emotionen des Leids	**2 196**	**33**	**23,87**	**1**	**302**	**26,27**
Leid (-)	998	15	26,12	1	302	28,91
Mitleid (-)	318	5	21,61	1	156	21,87
Ärger (-)	880	13	22,14	1	261	24,35
HK: Abscheu (-)	**614**	**9**	**25,05**	**1**	**167**	**26,19**
HK: Emotionale Bewegtheit	**763**	**12**	**24,4**	**1**	**313**	**32,74**
Gesamt	**6 596**	**100**	**23,82**	**1**	**326**	**24,08**

Abb. 1: Verteilung der Annotationsklassen. Nach den jeweiligen Hauptklassen (HK) folgen die Sub-Emotionen. + markiert positive Polarität, - negative Polarität (Avg=Mittelwert, Std=Standardabweichung). Die Aufteilung für die Polarität ist: 3.566 absolut, 54% für negativ, 2.267, 34% positiv und 763, 12% Emotionale Bewegtheit. Alle Prozentangaben sind gerundet.

Auf Polaritätsebene sind die meisten Annotationen negativ (56%), 34% positiv und 11% mit der Klasse „emotionale Bewegtheit" markiert. Einige Kategorien (z.B. Lust und Freundschaft) wurden selten markiert. Die Token-Statistiken verdeutlichen die Varianz in den Annotationslängen: im Schnitt besteht eine Annotation aber aus 25 Tokens für alle Kategorien.

Da Texteinheiten von variabler Länge und überlappende Texteinheiten annotiert werden können, muss zur Berechnung von Übereinstimmungsmetriken eine Festlegung auf eine Texteinheit getroffen werden. Dazu wird folgende Heuristik angewendet: Für jede Replik oder Regieanweisung wird pro Annotator*in diejenige Annotation markiert, die am meisten (gemessen an der Zahl an annotierten Token) markiert wurde. Keine Annotation pro Replik/Regieanweisung wird als zusätzliche Klasse markiert und dann replikenweise Übereinstimmungen kalkuliert (vgl. Abbildung 2).

Drama	Polarität (κ)	Polarität (%)	Hauptklasse (κ)	Hauptklasse (%)	sub-Emotion (κ)	sub-Emotion (%)
Faust	0,44	67,853	0,345	59,399	0,342	58,064
Kabale und Liebe	0,382	58,908	0,325	50,313	0,312	47,992
Menschenhass und Reue	0,402	75,28	0,347	72,331	0,347	71,91
Minna von Barnhelm	0,406	74,619	0,377	72,752	0,356	71,23
Kasperl'der Mandolettikrämer	0,42	70,83	0,344	65,34	0,312	62,72
Gesamt	0,41	69,498	0,3476	64,027	0,333	62,383

Abb. 2: Übereinstimmungsmetriken für jedes Drama und insgesamt (κ=Cohen's κ; %=prozentuelle Übereinstimmung der Annotator*innen).

Zur Interpretation von Cohen's κ werden im Folgenden in Klammern die Wertebereiche für einzelne Intervalle gemäß Landis und Koch (1977) mitangegeben. Im Schnitt kann man für die Polarität eine moderate Übereinstimmung (laut Landis und Koch gilt moderat für 0,4<κ<=0,6) und für die anderen Kategorien eine schwache Übereinstimmung (0,1<κ<= 0,4) feststellen. Im Vergleich zu anderen Textsorten ist dies eine geringe Übereinstimmung (Wood et al. 2018a; 2018b), die jedoch vergleichbar mit anderen Sentiment- und Emotionsannotationsprojekten mit literarischen und/oder historischen Texten ist (Alm / Sproat 2005; Sprugnoli et al. 2016; Schmidt et al. 2018b; Schmidt et al. 2019b; 2019d). Mehr Erläuterungen und Ergebnisse zur Annotation findet man bei Schmidt et al. (2021c).

Trainings- und Evaluationsmaterial

Im Folgenden werden die Ergebnisse für denjenigen Fall präsentiert, bei dem als Trainings- und Evaluationsmaterial („Goldstandard") alle Annotationen des obigen Annotationskorpus' verwendet werden (also je zwei Annotationssätze pro Drama). Dadurch liegt folgende Besonderheit vor: Eindeutige und partielle Annotationswidersprüche werden nicht aufgelöst, sondern dem Modell mit als Trainingsmaterial übergeben. Je nach kategorialem System gibt es eine unterschiedliche Menge an partiellen und absoluten Widersprüchen (ca. 16% für Polarität, 14% für Dreifach-Polarität, 28% für Hauptklassen, 47% für Sub-Emotionen). Dieses Verfahren wurde dennoch gewählt, da aufgrund der variablen Annotationspraxis die Auflösung eindeutiger Annotationswidersprüche schwerfällt (siehe Kapitel *Diskussion* mit Anregungen, wie mit diesem Problem in künftigen Studien umzugehen ist). Für weitere Evaluationen mit anderen Korpusinstanzen siehe Schmidt et al. (2021b). Insgesamt besteht der „Goldstandard" aus 6.596 annotierten Textsequenzen variabler Länge. Nicht-annotiertes Textmaterial wurde dem Goldstandard nicht hinzugefügt. Auch diese Limitation wird in der Diskussion besprochen.

Verfahren der Emotionsklassifikation

Wir definieren die Emotionsklassifikation als single-label-Klassifikationsaufgabe für Textsequenzen variabler Länge für folgende Klassengruppen:

- Polarität (zwei Klassen: positiv vs. negativ; Emotionale Bewegtheit wird hierbei entfernt)
- Dreifach-Polarität (drei Klassen)
- Hauptklassen (sechs Klassen)
- Sub-Emotionen (13 Klassen)

Alle Verfahren wurden in *Python* implementiert. Für die Evaluation und klassische Methoden des maschinellen Lernens wurde *scikit-learn* (Pedregosa et al. 2011) verwendet, für die transformerbasierten Modelle die *Hugging-Face* library (Wolf et al. 2019) und *simpletransformers*[4].

Baseline-Methoden

Obschon die Leistung lexikonbasierter Sentiment-Analyse meist von *Machine Learning*-Verfahren übertroffen wird, wird sie in den DH häufig angewendet, da keine vorannotierten Trainingskorpora notwendig sind (siehe Kim / Klinger 2019 und Schmidt et al. 2021a). Das Verfahren ist regelbasiert und wird bei Taboada et al. (2011) beschrieben. Wir evaluieren zwei Ansätze: (1) das Lexikon *SentiWortschatz* (SentiWS) (Remus et al. 2010) ohne Vorverarbeitung (im Folgenden als *lb-sentiws* bezeichnet), (2) SentiWS kombiniert mit Methoden wie Lemmatisierung und Lexikonerweiterung (Schmidt / Burghardt 2018) (*lb-sentiws-optimized*). Letztere Methodik erzielte gute Ergebnisse in historischen deutschsprachigen Dramen (Schmidt / Burghardt 2018). Die gewählten Ansätze können nur für die Polarität angewendet werden, da keine differenzierten Emotionsannotationen in SentiWS vorhanden sind.

Wir evaluieren zudem zwei klassische Methoden des maschinellen Lernens: (1) Repräsentation über Termfrequenzen in einem bag-of-words-Modell und dem Lern-Algorithmus *Multinomial Naive Bayes* (*bow-mnb*) sowie (2) *Support Vector Machines* (SVM) (*bow-svm*) als Lern-Algorithmus. Methode (2) wurde mit dem rbf-kernel der SVC-Klasse von scikit-learn umgesetzt.[5] Für mehr Informationen über bag-of-words-Ansätze siehe Gonzáles-Carvajal et al. (2021). Die Algorithmen wurden in einer stratifizierten 5x5 Kreuzevaluation trainiert und evaluiert.

fastText

Statische Sprachmodelle repräsentieren Wörter als Vektoren in Vektorräumen, so dass geometrische Verhältnisse der jeweiligen Semantik entsprechen. Diese Repräsentationen (*word embeddings*) können als Input für neuronale Netze genutzt werden. Wir evaluieren das *word embedding fastText* (Bojanowski et al. 2017), da es im Vergleich zu anderen statischen Modellen gute Ergebnisse für deutsche Sprache erzielt (Schmitt et al. 2018). Wir nutzen deutschsprachige *fastText embeddings*[6] trainiert auf der deutschsprachigen Wikipedia sowie ein rekurrentes neuronales Netzwerk (RNN) zur Klassifikation (Cho et al. 2014). Bezüglich der Hyperparameter wird der empfohlene Default des FLAIR-frameworks gewählt (Akbik et al. 2019)[7] und je ein Modell in einem stratifizierten 5x5-Setting für 12 Epochen trainiert und evaluiert. Für alle Evaluationsmetriken wird der Mittelwert aus den Ergebnissen der fünf Modelle gebildet.

Transformerbasierte Sprachmodelle (zeitgenössische Sprache)

Als transformerbasierte Sprachmodelle werden dynamische *word embeddings* wie BERT (Devlin et al. 2018) oder ELECTRA (Clark et al. 2020) bezeichnet, die in Erweiterung zu statischen Modellen den Kontext eines Wortes in seiner Umgebung. Wir evaluieren einige der wichtigsten und (über die *Hugging Face*-Plattform[8]) frei verfügbaren Modelle, die auf zeitgenössischer Sprache trainiert wurden (Abbildung 3). Die gewählten Modelle erreichen state-of-the-art-Ergebnisse in standardisierten Evaluationen auf deutscher Sprache (Chan et al. 2020).

ID	Texte für das Vortraining	Zugehöriges Paper (wenn verfügbar) und Provider
bert-base-german-cased	Wikipedia, juristische Texte, News (~ 12 GB)	Deepset
dbmdz-bert-base-german-cased	Wikipedia, Bücher, Untertitel, Web-Texte, News (~ 16 GB)	MDZ Digital Library
electra-base-german-uncased	Wikipedia, Untertitel, News (~ 73 GB)	German-NLP-Group
gbert-large	Web-Texte, Wikipedia, Untertitel, Bücher, juristische Texte (~ 161 GB)	Deepset (Chan et al., 2020)
gelectra-large	Web-Texte, Wikipedia, Untertitel, Bücher, juristische Texte (~ 161 GB)	Deepset (Chan et al., 2020)

Abb. 3: Evaluierte transformerbasierte Modelle (vortrainiert mit zeitgenössischer Sprache).

Für die Klassifikationsaufgabe werden die Modelle in einem „Fine-Tuning"-Schritt mit dem Goldstandard trainiert. Für die konkrete Implementierung folgen wir den jeweiligen Empfehlungen für die gewählte Architektur (Devlin et al. 2018; Clark et al. 2020)[9] und nutzen die *Hugging Face*-Bibliothek (Wolf et al. 2020). Pro Sprachmodell und Klassifikationstask werden fünf Klassifikationsverfahren in einem stratifizierten 5x5-setting für je vier Epochen trainiert und Mittelwerte gebildet.

Transformerbasierte Sprachmodelle (historische/poetische Sprache)

Die Performanz von Klassifikations-Aufgaben kann verbessert werden, indem Texte der gleichen Domäne zum Vortraining von transformerbasierten Modellen genutzt werden (siehe Rietzler et al. 2020; Gururangan et al. 2020). Man kann entweder (1) selbst ein Modell von Grund auf mit domänennahen Texten erstellen oder (2) Modelle zeitgenössischer Sprache mit domänenspezifischen historischen Texten nachtrainieren. Beide Methoden wurden bereits erfolgreich im Kontext deutscher, historischer Sprache angewendet (Labusch et al. 2019; Schweter / Baiter 2019; Schweter / März 2020; Brunner et al. 2020).

ID	Vortrainierte Texte	Zeitraum	Zugehöriges Paper (wenn verfügbar) und Provider
bert-base-german-europeana-cased	Europeana-Zeitungen (51 GB)	18.-20. Jahrhundert	MDZ Digital Library (Schweter, 2020)
electra-base-german-europeana-cased-discriminator	Europeana-Zeitungen (51 GB)	18.-20. Jahrhundert	MDZ Digital Library (Schweter, 2020)
literary-german-bert	Basiert auf bert-base-german-dbmdz-cased weiter vortrainiert mit *Corpus of German-Language-Fiction (CGLF)* (~ 1 GB)	CGLF: hauptsächlich 1840-1930	Severin Simmler
bert-base-historical-german-rw-cased	Märchen, historische Zeitungen, narrative Texte, Texte von Projekt Gutenberg (genaue Größe unbekannt)	1840-1920	Brunner et al. (2020)

Abb. 4: Evaluierte transformerbasierte Modelle vortrainiert mit historischer Sprache.[10]

Auch hier evaluieren wir etablierte vortrainierte Modelle, die über die *Hugging Face*-Plattform frei verfügbar sind. Abbildung 4 fasst die Daten der Modelle zusammen. Alle Modelle nähern sich dem Kontext unserer Dramen-Texte auf historischer Ebene oder dadurch, dass narrative/poetische Texte genutzt werden, an. Des Weiteren wurde das Modell *bert-base-german-cased* noch mit den Texten des eigenen Korpus nachtrainiert, zum einen mit unserem Hauptkorpus GerDracor (*bert-base-german-cased-main-corpus*) und in einem zweiten Ansatz lediglich mit den annotierten Dramen (*bert-base-german-cased-annotated-texts).* Das Nachtraining wurde für 4 Epochen mit den default-settings der *simpletransformer*-library durchgeführt.[11] Das Implementierungs-, Trainings- und Evaluationsverfahren sowie die gewählten Hyperparameter für die Emotionsprädiktion sind äquivalent zum vorigen Kapitel.

Ergebnisse

Hauptmetrik zur Interpretation der Ergebnisse ist die *accuracy*, also der Anteil an korrekt erkannten Annotationen an allen Annotationen (siehe Abbildung 5). Weitere Details und Informationen zu den Ergebnissen der Studie findet man bei Schmidt et al. (2021d).

Methodengruppe	Methode / accuracy	Polarität	Dreifach-Polarität	Hauptkategorie	Sub-Emotion
-	random baseline	.500	.333	.167	.077
	majority baseline	.612	.541	.333	.151
Baseline-Methoden	lb-sentiws	.445	-	-	-
	lb-sentiws-optimized	.588	-	-	-
	bow-mnb	.742	.659	.451	.348
	bow-svm	.685	.603	.392	.284
Statisches Sprachmodell	fasttext	.714	.647	.404	.289
Zeitgenössische Transformer-Modelle	bert-base-german-cased	.804	.711	.512	.428
	dbmdz-bert-base-german-cased	.804	.716	.517	.430
	electra-base-german-uncased	.776	.690	.474	.358
	gbert-large	**.821**	**.740**	**.545**	**.467**
	gelectra-large	**.825**	**.748**	**.564**	**.460**
Historische Transformer-Modelle	bert-base-german-europeana-cased	.798	.718	**.528**	.420
	electra-base-german-europeana-cased-discriminator	.808	.722	.525	.416
	literary-german-bert	.799	.718	-	-
	bert-base-historical-german-rw-cased	**.813**	**.723**	.524	**.444**
Transformer-Modelle trainiert mit eigenem Korpus	bert-base-german-cased-main-corpus	.796	.714	.492	.379
	bert-base-german-cased-annotated-texts	.809	.709	.505	.425

Abb. 5: Klassifikationsergebnisse für alle Methoden (die drei besten Ergebnisse je Kategorie sind hervorgehoben).

Alle gewählten Methoden übertreffen in den einzelnen Settings die *random* und *majority*-baseline. Die Ergebnisse der lexikonbasierten Sentiment-Analyse bewegen sich auf einem ähnlichem Niveau für Evaluationen auf unterschiedlichen literarischen Texten (Fehle et al. 2021). Die beste Erkennungsrate für Polarität beträgt 83% und wird vom Modell *gelectra-large* erreicht. Gleiches gilt für die Dreifach-Polarität mit 75% sowie die Hauptklassen (55%). Das beste Modell für die Sub-Emotionen ist *gbert-large* mit jedoch lediglich 47% Erkennungsrate. Transformerbasierte Modelle erreichen im Schnitt wesentliche bessere Erkennungsraten als alle Baseline-Methoden oder fastText. Mit zunehmender Klassenzahl werden die Ergebnisse (trivialerweise) schlechter. Auch die Abstände zwischen bester und schlechtester Methode werden geringer. Die drei besten Modelle sind konsistent die zwei größten Modelle zeitgenössischer Sprache *gbert-large* und *gelectra-large* sowie das auf historische und narrative Sprache optimierte Modell *bert-base-historical-german-rw-cased*.

Diskussion

Obschon die Menge an annotiertem Material im Vergleich zu Studien auf der Basis anderer Textsorten limitiert ist, konnten wir erste Erkenntnisse für die Optimierung computergestützter Methoden sammeln. Für Polarität und Dreifach-Polarität erreichen die besten Modelle in ihren Default-Settings bereits Ergebnisse, die durchaus vergleichbar sind mit state-of-the-art-Resultaten für Sentiment- und Emotionsklassifikation in anderen Bereichen (Yang et al. 2019; Munikar et al. 2019; Cao et al. 2020; Dang et al. 2020). Die besten Ergebnisse erzielen grundsätzlich die derzeit größten transformerbasierten Modelle für die deutsche Sprache. Die Optimierung für historische oder poetische Sprache hat lediglich geringfügige Verbesserungen gegenüber den äquivalenten kontemporären Modellen aufgezeigt. Ein Grund dafür ist möglicherweise, dass die gewählten historischen Modelle noch zu viele Texte aus dem 19. und 20. Jahrhundert enthalten, die doch zu weit entfernt von unserer Zeitepoche sind. Wir befinden uns momentan im Prozess der Akquise großer Textmengen aus dem entsprechenden Zeitraum, um vortrainierte Modelle zu evaluieren, die noch stärker an unsere Domäne angepasst sind.

Für die mehrklassigen Kategoriensysteme können keine zufriedenstellenden Ergebnisse erzielt werden. Dies ist ohne größere Optimierung für derartige Klassifikationsverfahren nicht ungewöhnlich. Wir planen sowohl die Anwendung verschiedener empfohlener Verfahren, um mit dem Klassenungleichgewicht umzugehen (Buda et al. 2018) und die Optimierung von Hyperparametern als auch die Exploration des Einsatzes einer neutralen „Nicht-annotiert“-Klasse. Im Bereich der Annotation soll eine Expertenannotation eingefügt werden, welche die Entscheidungen der ersten beiden Annotationen berücksichtigt, aber eine eigenständig verwendbare, widerspruchsfreie Annotationsschicht darstellt. Evaluationsergebnisse mittels der Anwendung von manuellen Widerspruchsauflösungen findet man bei Schmidt et al. (2021b). Wir lassen derzeit weitere Texte annotieren und explorieren historische *word embeddings*, um akzeptable Ergebnisse für die Hauptkategorien zu erreichen und Emotionen in größeren Mengen unseres Korpus vorhersagen zu können.

Fußnoten

1. Das Projekt wird von der Deutschen Forschungsgemeinschaft (DFG) im Rahmen des Schwerpunktprogramms Computational Literary Studies (SPP 2207/1) gefördert https://dfg-spp-cls.github.io/projects_en/2020/01/24/TP-Emotions_in_Drama/

(Sachbeihilfen DE 2188/3-1 und WO 835/4-1, Projektnummer: 424207618).

2. Zur Definition des Emotionsbegriffs und zur Emotionsauswahl vgl. Dennerlein et al. 2022
3. https://textgrid.de
4. https://simpletransformers.ai/
5. https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC
6. https://fasttext.cc/docs/en/crawl-vectors.html
7. Lern-Rate: 0.1, Batch-Größe: 32
8. https://huggingface.co/
9. Lern-Rate: 0.00004; Batch-Größe: 32, maximale Sequenzenlänge: 128; Adam als Optimizer
10. Aus Architektur-Gründen wird das Modell literary-german-bert nur für Polarität/Dreifach-Polarität evaluiert
11. Siehe https://simpletransformers.ai/docs/lm-specifics/

Bibliographie

Acheampong, Francisca Adoma / Wenyu, Chen / Nunoo-Mensah, Henry (2020): "Text-based emotion detection: Advances, challenges, and opportunities.", in: *Engineering Reports* 2.7: e12189.

Akbik, Alan / Bergmann,Tanja / Blythe, Duncan / Rasul, Kashif / Schweter, Stefan / Vollgraf, Roland (2019): "FLAIR: An easy-to-use framework for state-of-the-art NLP.", in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*.

Alm, Cecilia Ovesdotter / Sproat, Richard (2005): "Emotional sequencing and development in fairy tales.", in: *International Conference on Affective Computing and Intelligent Interaction*. Springer, Berlin, Heidelberg.

Anz, Thomas (2011): "Todesszenarien: literarische Techniken zur Evokation von Angst, Trauer und anderen Gefühlen.", in: *Emotionale Grenzgänge.* Würzburg, pp. 54–59.

Bojanowski, Piotr / Grave, Edouard / Joulin, Armand / Mikolov, Tomas (2017): "Enriching word vectors with subword information.", in: *Transactions of the Association for Computational Linguistics* 5: 135-146.

Brunner, Annelen / Duyen Tanja Tu, Ngoc / Weimer, Lukas / Jannidis, Fotis (2020): "To BERT or not to BERT-Comparing Contextual Embeddings in a Deep Learning Architecture for the Automatic Recognition of four Types of Speech, Thought and Writing Representation.", in: *SwissText/KONVENS*.

Buda, Mateusz / Maki, Atsuto / Mazurowski, Maciej A. (2018): "A systematic study of the class imbalance problem in convolutional neural networks.", in: *Neural Networks* 106: 249-259.

Cao, Lihong / Peng, Sancheng / Yin, Pengfei / Zhou, Yongmei / Yang, Aimin / Li, Xinguang (2020): "A Survey of Emotion Analysis in Text Based on Deep Learning.", in: *2020 IEEE 8th International Conference on Smart City and Informatization (iSCI)*. IEEE.

Chan, Branden / Schweter, Stefan / Möller, Timo (2020): "German's Next Language Model.", in: *arXiv preprint arXiv:2010.10906*

Cho, Kyunghyun / Merriënboer, Bart van / Bahdanau, Dzmitry / Bengio, Yoshua (2014): "On the properties of neural machine translation: Encoder-decoder approaches.", in: *arXiv preprint arXiv:1409.1259*

Clark, Kevin / Luong, Minh-Thang / Le, Quoc V. / Manning, Christopher D. (2020): "Electra: Pre-training text encoders as discriminators rather than generators.", in: *arXiv preprint arXiv:2003.10555*

Cortiz, Diogo (2021): "Exploring Transformers in Emotion Recognition: a comparison of BERT, DistillBERT, RoBERTa, XLNet and ELECTRA." *arXiv preprint arXiv:2104.02041*

Dang, Nhan Cach / Moreno-García, María N. / De la Prieta, Fernando (2020): "Sentiment-Analyse based on deep learning: A comparative study.", in: *Electronics* 9.3 (2020): 483.

Dennerlein, Katrin / Schmidt, Thomas / Wolff, Christian (2022): "Emotionen im kulturellen Gedächtnis bewahren.", in: *Book of Abstracts, DHd2022*.

Devlin, Jacob / Chang, Ming-Wei / Lee, Kenton / Toutanova, Kristina (2018): "Bert: Pre-training of deep bidirectional transformers for language understanding.", in: *arXiv preprint arXiv:1810.04805*.

Fehle, Jakob / Schmidt, Thomas / Wolff, Christian (2021): "Lexicon-based Sentiment Analysis in German: Systematic Evaluation of Resources and Preprocessing Techniques", in; *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, 86-103.

Fischer, Frank / Börner, Ingo / Göbel, Mathias / Hechtl, Angelika / Kittel, Christopher / Milling, Carsten / Trilcke, Peer (2019): "Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama." Zenodo. <https://doi.org/10.5281/zenodo.4284002>

Evelyn, Gius / Meister, Jan Christoph / Petris, Marco / Meister, Malte / Bruck, Christian / Jacke, Janina / Schumacher, Mareike / Flüh, Marie / Horstmann, Jan (2020): "CATMA." Zenodo. <https://doi.org/10.5281/zenodo.4353618.>

González-Carvajal, Santiago / Garrido-Merchán, Eduardo C. (2020): "Comparing BERT against traditional machine learning text classification.", in: *arXiv preprint arXiv:2005.13012*

Gururangan, Suchin / Marasović, Ana / Swayamdipta, Swabha / Lo Kyle / Beltagy, Iz / Downey, Doug et al. (2020): "Don't stop pretraining: adapt language models to domains and tasks.", in: *arXiv preprint arXiv:2004.10964*

Kakkonen, Tuomo / Kakkonen, Gordana Galić (2011): "SentiProfiler: creating comparable visual profiles of sentimental content in texts.", in: *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage*.

Kim, Evgeny / Klinger, Roman (2019): "A survey on sentiment and emotion analysis for computational literary studies.", in: *Zeitschrift für digitale Geisteswissenschaften.*

Labusch, Kai / Neudecker, Clemens / Zellhofer, David (2019): "BERT for Named Entity Recognition in Contemporary and Historical German.", in: *Proceedings of the 15th Conference on Natural Language Processing, Erlangen, Germany*.

Landis, J. Richard / Koch, Gary G. (1977): "The measurement of observer agreement for categorical data.", in: *biometrics* (1977): 159-174.

Mohammad, Saif (2011): "From once upon a time to happily ever after: Tracking emotions in novels and fairy tales.", in: *arXiv preprint arXiv:1309.5909*

Moßburger, Luis / Wende, Felix / Brinkmann, Kay / Schmidt, Thomas (2020): "Exploring Online Depression Forums via Text Mining: A Comparison of Reddit and a Curated Online Forum", in: *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, 70-81.

Munikar, Manish / Shakya, Sushil / Shrestha, Aakash (2019): "Fine-grained sentiment classification using bert." *2019 Artificial Intelligence for Transforming Business and Society (AITB)*. Vol. 1. IEEE.

Pedregosa, Fabian et al. (2011): "Scikit-learn: Machine learning in Python", in: *Journal of machine Learning research*, 12, 2825-2830.

Pikulik, Lothar (1966): *"Bürgerliches Trauerspiel" und Empfindsamkeit*. Köln, Graz.

Reagan, Andrew J. / Mitchell, Lewis / Kiley, Dilan / Danforth, Christopher M. / Dodds, Peter Sheridan (2016): "The emotional arcs of stories are dominated by six basic shapes." *EPJ Data Science* 5.1: 1-12.

Remus, Robert / Quasthoff, Uwe / Heyer, Gerhard (2010): "SentiWS-A Publicly Available German-language Resource for Sentiment-Analyse.", in: *LREC*.

Rietzler, Alexander / Stabinger, Sebastian / Opitz, Paul / Engl, Stefan (2019): "Adapt or get left behind: Domain adaptation through bert language model finetuning for aspect-target sentiment classification.", in: *arXiv preprint arXiv:1908.11860*

Schmidt, Thomas (2019): "Distant Reading Sentiments and Emotions in Historic German Plays", in: *Abstract Booklet, DH_Budapest_2019*. Budapest, Hungary, 57-60.

Schmidt, Thomas / Burghardt, Manuel (2018): "An Evaluation of Lexicon-based Sentiment Analysis Techniques for the Plays of Gotthold Ephraim Lessing", in: *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Santa Fe, New Mexico: Association for Computational Linguistics, 139-149.

Schmidt, Thomas / Burghardt, Manuel / Dennerlein, Katrin (2018a): ""Kann man denn auch nicht lachend sehr ernsthaft sein?" – Zum Einsatz von Sentiment Analyse-Verfahren für die quantitative Untersuchung von Lessings Dramen", in: *Book of Abstracts, DHd 2018.*

Schmidt, Thomas / Burghardt, Manuel / Dennerlein, Katrin (2018b): "Sentiment Annotation of Historic German Plays: An Empirical Study on Annotation Behavior.", in: Sandra Kübler, Heike Zinsmeister (eds.), *Proceedings of the Workshop on Annotation in Digital Humanities (annDH 2018)*. Sofia, Bulgaria, 47-52.

Schmidt, Thomas / Burghardt, Manuel / Dennerlein, Katrin / Wolff, Christian (2019a): "Katharsis - A Tool for Computational Drametrics", in: *Book of Abstracts, Digital Humanities Conference 2019 (DH 2019)*. Utrecht, Netherlands.

Schmidt, Thomas / Burghardt, Manuel / Dennerlein, Katrin / Wolff, Christian (2019b): "Sentiment Annotation in Lessing's Plays: Towards a Language Resource for Sentiment Analysis on German Literary Texts", in: *2nd Conference on Language, Data and Knowledge (LDK 2019)*. LDK Posters. Leipzig, Germany.

Schmidt, Thomas / Burghardt, Manuel / Wolff, Christian (2019c): "Towards Multimodal Sentiment Analysis of Historic Plays: A Case Study with Text and Audio for Lessing's Emilia Galotti.", in: *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference (DHN 2019)*. Copenhagen, Denmark, 405-414.

Schmidt, Thomas / Winterl, Brigitte / Maul, Milena / Schark, Alina / Vlad, Andrea / Wolff, Christian (2019d): "Inter-Rater Agreement and Usability: A Comparative Evaluation of Annotation Tools for Sentiment Annotation", in: Draude, C., Lange, M. & Sick, B. (Hrsg.), *INFORMATIK 2019: 50 Jahre Gesellschaft für Informatik – Informatik für Gesellschaft (Workshop-Beiträge)*. Bonn: Gesellschaft für Informatik e.V., 121-133. DOI: 10.18420/inf2019_ws12

Schmidt, Thomas / Bauer, Marlene / Habler, Florian / Heuberger, Hannes / Pilsl, Florian / Wolff, Christian (2020a): "Der Einsatz von Distant Reading auf einem Korpus deutschsprachiger Songtexte", in *Book of Abstracts, DHd 2020*, 296-299.

Schmidt, Thomas / Engl, Isabella / Halbhuber, David / Wolff, Christian (2020b): "Comparing Live Sentiment Annotation of Movies via Arduino and a Slider with Textual Annotation of Subtitles", in: *Post-Proceedings of the 5th Conference Digital Humanities in the Nordic Countries (DHN 2020)*, 212-223.

Schmidt, Thomas / Hartl, Philipp / Ramsauer, Dominik / Fischer, Thomas / Hilzenthaler, Andreas / Wolff, Christian (2020c): "Acquisition and Analysis of a Meme Corpus to Investigate Web Culture", in: *15th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2020, Conference Abstracts.* Ottawa, Canada.

Schmidt, Thomas / Kaindl, Florian / Wolff, Christian (2020d): "Distant Reading of Religious Online Communities: A Case Study for Three Religious Forums on Reddit", in: *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference (DHN 2020).* Riga, Latvia.

Schmidt, Thomas / Dangel, Johanna / Wolff, Christian (2021a): "SentText: A Tool for Lexicon-based Sentiment Analysis in Digital Humanities", in: Schmidt, Thomas / Wolff, Christian (Eds.), *Information between Data and Knowledge. Information Science and its Neighbors from Data Science to Digital Humanities. Proceedings of the 16th International Symposium of Information Science (ISI 2021).* Glückstadt: Verlag Werner Hülsbusch, 156-172. DOI: 10.5283/epub.44943

Schmidt, Thomas / Dennerlein, Katrin / Wolff, Christian (2021b): "Emotion Classification in German Plays with Transformer-based Language Models Pretrained on Historical and Contemporary Language", in: *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 67-79.

Schmidt, Thomas / Dennerlein, Katrin / Wolff, Christian (2021c): "Towards a Corpus of Historical German Plays with Emotion Annotations", in: *3rd Conference on Language, Data and Knowledge (LDK 2021).* Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Schmidt, Thomas / Dennerlein, Katrin / Wolff, Christian (2021d): "Using Deep Neural Networks for Emotion Analysis of 18th and 19th century German Plays", in: *Fabrikation von Erkenntnis: Experimente in den Digital Humanities.* Melusina Press. DOI:10.26298/melusina.8f8w-y749-udlf

Schmitt, Martin / Steinheber, Simon / Schreiber, Konrad / Roth, Benjamin (2018): "Joint aspect and polarity classification for aspect-based Sentiment-Analyse with end-to-end neural networks.", in: *arXiv preprint arXiv:1808.09238*

Schonlau, Anja (2017): *Emotionen im Dramentext: eine methodische Grundlegung mit exemplarischer Analyse zu Neid und Intrige 1750-1800*. De Gruyter.

Schweter, Stefan (2020): *Europeana BERT and ELECTRA models*. <https://doi.org/10.5281/zenodo.4275044>

Schweter, Stefan / Baiter, Johannes (2019): "Towards robust named entity recognition for historic german.", in: *arXiv preprint arXiv:1906.07592* (2019).

Schweter, Stefan / März, Luisa (2020): "Triple E-Effective Ensembling of Embeddings and Language Models for NER of Historical German.", in: *CLEF (Working Notes)*.

Shmueli, Boaz / Lun-Wei Ku (2019): "Socialnlp emotionx 2019 challenge overview: Predicting emotions in spoken dialogues and chats.", in: *arXiv preprint arXiv:1909.07734*

Sprugnoli, Rachele / Tonelli, Sara / Marchetti, Alessandro / Moretti, Giovanni (2016): "Towards Sentiment-Analyse for historical texts.", in: *Digital Scholarship in the Humanities* 31.4: 762-772.

Wiegmann, Hermann (Hrsg.) (1987): *Die ästhetische Leidenschaft: Texte zur Affektenlehre im 17. und 18. Jahrhundert*.

Wolf, Thomas / Debut, Lysandre / Sanh, Victor / Chaumond, Julien / Delangue, Clement / Moi, Anthony et al. (2019): "Huggingface's transformers: State-of-the-art natural language processing." *arXiv preprint arXiv:1910.03771*

Wood, Ian / McCrae, John / Andryushechkin, Vladimir / Buitelaar, Paul (2018a): "A comparison of emotion annotation schemes and a new annotated data set.", in: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Wood, Ian / McCrae, John / Andryushechkin, Vladimir / Buitelaar, Paul (2018b): "A comparison of emotion annotation approaches for text.", in *Information* 9.5: 117.

Yang, Kisu / Lee, Dongyub / Whang, Taesun / Lee, Seolhwa / Lim, Heuiseok (2019): "Emotionx-ku: Bert-max based contextual emotion classifier.", in: *arXiv preprint arXiv:1906.11565*

Yavuz, Mehmet Can (202 "Analyses of Character Emotions in Dramatic Works by Using EmoLex Unigrams.", in: *CLiC-it*.

Zehe, Albin / Becker, Martin / Hettinger, Lena / Hotho, Andreas / Reger, Isabella / Jannidis, Fotis (2016): "Prediction of happy endings in German novels based on sentiment information.", in: *3rd Workshop on Interactions between Data Mining and Natural Language Processing, Riva del Garda, Italy*.