Universität Zürich[UZH]

Bachelorarbeit

zur Erlangung des akademischen Grades

Bachelor of Arts

der Philosophischen Fakultät der Universität Zürich

Emotionsanalyse in literarischen Texten mittels Machine Learning

Verfasserin: Karin Thommen

Matrikel-Nr: 16-580-011

Betreuer: Dr. Manfred Klenner

Institut für Computerlinguistik

Abgabedatum: 01.06.2021

Abstract

Sentiment and emotion analysis has become an active field of research in recent years, especially regarding the detection of hate comments on social media or the classification and filtering of online ratings on the internet. However, there is little research that deals with machine emotion analysis in literary texts. Clearly identifying an emotion in a text passage is already a difficult task for humans and an even greater challenge for a machine learning-based system. Moreover, there are only few annotated data that form a basis for such a system.

However, emotions are of great importance in literary texts, as they are fundamentally related to and influence the plot, but also determine how a reader perceives a story.

The aim of this work is to investigate whether and with what quality literary texts can be classified according to emotions using the Bidirectional Encoder Representations for Transformers (BERT), which represents a current state-of-the-art in language models. The aim is to train and develop a system that is capable of reproducing the emotion distributions of books.

Zusammenfassung

Die Sentiment- und Emotionsanalyse wurde in den vergangenen Jahren zu einem aktiven Forschungsfeld, vor allem was die Erkennung von Hasskommentaren auf Social Media betrifft oder die Einordnung und Filterung von Online Bewertungen im Internet. Es gibt jedoch wenige Forschungen, die sich mit der maschinellen Emotionsanalyse bei literarischen Texten befassen. Die eindeutige Identifikation einer Emotion in einer Textstelle stellt bereits für Menschen eine schwierige Aufgabe dar und ist für ein auf maschinellem Lernen basierten System eine noch grössere Herausforderung. Es existieren zudem erst wenige annotierte Datensätze, welche eine Grundlage für ein solches System bilden.

Emotionen sind in literarischen Texten jedoch von grosser Bedeutung, da sie grundlegend mit der Handlung zusammenhängen und diese beeinflussen, aber auch bestimmen, wie ein Leser oder eine Leserin eine Geschichte wahrnimmt.

In dieser Arbeit soll untersucht werden, ob und mit welcher Qualität mittels dem Bidirectional Encoder Representations for Transformers (BERT), welcher ein aktueller State-of-the-Art der Sprachmodelle darstellt, literarische Texte nach Emotionen klassifiziert werden können. Das Ziel ist das Trainieren und Ausarbeiten eines Systems, welches befähigt ist, die Emotionsverteilungen von Büchern wiederzugeben.

Danksagung

Ich möchte mich an dieser Stelle herzlich bei allen bedanken, die mich bei der Erstellung meiner Bachelorarbeit unterstützt haben. Ein besonderer Dank gilt Herrn Dr. Manfred Klenner, welcher sich als Betreuungsperson für meine Arbeit bereit erklärt hatte. Ich konnte mich jederzeit mit Fragen an ihn wenden und erhielt konstruktive Ratschläge und wertvolle Unterstützung bei der Realisierung dieser Arbeit. Weiter möchte ich Martin Knecht danken, welcher sich als zusätzlichen Annotierenden zur Verfügung gestellt hat, um die Stichprobe des Datensets zu überprüfen. Aber auch meiner Familie und meinen Freunden, die mich während meinem Bachelorstudium unterstützt und motiviert haben.

Inhaltsverzeichnis

Abstract ..... **I**

Zusammenfassung ..... **I**

Danksagung ..... **II**

Inhaltsverzeichnis ..... **III**

Abbildungsverzeichnis ..... **V**

Abkürzungsverzeichnis ..... **VII**

Sprachkürzel VII

1 Einleitung **1**

1.1 Motivation 1

1.2 Fragestellung 1

1.3 Aufbau der Arbeit 2

2 Theoretische Grundlagen **3**

2.1 Sentiment- und Emotionsanalyse 3

2.2 Affekt und Emotion 4

2.2.1 Basisemotionen nach Ekman 5

2.2.2 Emotionsmodell nach Plutchik 5

2.2.3 Circumplexmodell nach Posner, Russell und Peterson 5

2.3 Stand der Forschung 6

3 Daten und Methoden **7**

3.1 Emotionsklassifizierung 7

3.1.1 Lexikonbasierter Ansatz 7

3.1.2 Machine Learning Ansatz 7

3.2 Korpora und Lexika 9

3.2.1 Annotierter Goldstandard 9

3.2.2 NRC Lexikon 12

3.2.3 Halbautomatisch annotierter Silver Standard 13

3.3 Klassifikationssystem ... 14
3.3.1 Preprocessing des REMAN Korpus ... 14
3.3.2 Generierung des Silver Standards ... 15
3.3.3 BERT ... 18
3.3.4 Emotion Predictor – Erweiterung von BERT ... 19
4 Evaluation ... 21
4.1 Überprüfung der Stichprobe ... 21
4.2 BERT und der Gold Standard aus dem REMAN Korpus ... 23
4.3 BERT und der Silver Standard ... 29
4.4 BERT mit dem Gold und Silver Standard ... 31
4.5 Evaluation des Emotion Predictors ... 34
5 Diskussion ... 37
5.1 Herausforderungen der Emotionsanalyse ... 37
5.1.1 Fehlende Kategorien bei den Emotionen ... 37
5.1.2 Entscheidung für eine Emotion ... 38
5.1.3 Abhängigkeit von Blickwinkel ... 39
5.1.4 Verneinungen, Modalverben ... 40
5.1.5 Ironie und Sarkasmus ... 40
5.1.6 Veränderung der Sprache über die Zeit ... 41
5.2 Fazit ... 41
5.3 Ausblick ... 43
Literatur- und Quellenverzeichnis ... i
Anhang ... iii
Daten und Skripte für Preprocessing und Generierung des Silver Filter ... iii
Dateien zur Evaluation ... iv
BERT Modell ... iv
Readme REMAN Korpus ... iv
Verwendete Bücher aus dem Gutenberg Projekt ... v

Abbildungsverzeichnis

Abbildung 1 Emotionsmodell nach Plutchik (Wikipedia, Zugriff: 2. März 2021) 5
Abbildung 2 Circumplexmodell nach Posner, Russell und Peterson (2005), Bild und Übersetzung von (Klinger, 2020) 5
Abbildung 3 Beispiel der Annotation aus dem Annotationshandbuch des REMAN Korpus (Screenshot aus Guidlines.pdf, Downloadordner REMAN Korpus) (Kim & Klinger, Who Feels What and Why? Annotation of a Literature Corpus with Semantic Roles of Emotions, 2018) 10
Abbildung 4 Auszug aus der CSV-Datei basierend auf dem REMAN Korpus 11
Abbildung 5 Emotionsverteilung der verarbeiteten Daten (REMAN) als Diagramm 11
Abbildung 6 Absolute Zahlen der Emotionsverteilung (REMAN) 11
Abbildung 7 Ausschnitt aus dem Intensitätslexikon für die Emotion «joy» (de: Freude) 12
Abbildung 8 Emotionsverteilung im Intensitätslexikon 12
Abbildung 9 Absolute Zahlen der Emotionsverteilung im Intensitätslexikon 13
Abbildung 10 Auszug aus der XML-Datei des REMAN Korpus 15
Abbildung 11 Darstellung der Annotation durch den lexikonbasierten Ansatz des Filters zur Generierung des Silver Standards 16
Abbildung 12 Darstellung einer weiteren Annotation durch den lexikonbasierten Ansatz des Filters zur Generierung des Silver Standards 17
Abbildung 13 Verteilung der Emotion der vollständigen Version des Silver Standards (10702 Sätze) 17
Abbildung 14 Emotion Predictor Output nach einem Dateiupload 19
Abbildung 15 Emotion Predictor Output nach eingegebenem Text 20
Abbildung 16 Anzahl der annotierten Emotionen pro Annotierenden (Annotator 1 ist der Silver Filter beziehungsweise der Silver Standard, Annotator 2 und 3 sind menschliche Annotierende) 22
Abbildung 17 Übereinstimmung zwischen den Annotierenden. A1 steht für Annotator 1 usw. Für die Prozentzahlen in der rechten Spalte wurden jeweils die Summe durch das Total der oberen Tabelle dividiert, um die Genauigkeit bei den Emotionen zu berechnen. Die unterste Spalte zeigt jeweils das Verhältnis der totalen Übereinstimmungen der Annotierenden zur Gesamtzahl von 150. 75% ist der Durchschnitt über alle Prozentzahlen der Übereinstimmungen. 22

Abbildung 18 Konfusionmatrix nach Training und Testing von BERT. Daten inkl. «other»-Label 24
Abbildung 19 Konfusionmatrix nach Training und Testing von BERT. Daten ohne «other»-Label 24
Abbildung 20 Top 10 Losses von BERT mit Trainingsdaten aus dem REMAN Korpus 25
Abbildung 21 Output von predictor.explain() zur Klassifizierung eines Satzes durch BERT. 26
Abbildung 22 Output von predictor.explain() zur Klassifizierung eines weiteren Beispielsatzes durch BERT 27
Abbildung 23 Konfusionsmatrix von BERT mit ganzen Sätzen aus dem REMAN Korpus ... 28
Abbildung 25 BERT mit komplettem Silver-Datenset 29
Abbildung 24 BERT mit Selektion des Silver Standards 29
Abbildung 26 Top 10 Losses des Trainings mit dem Silver Standard 30
Abbildung 27 Gold und Silver Daten als Input für BERT 31
Abbildung 28 Konfusionsmatrix von BERT mit Clean Silver und Gold Daten 32
Abbildung 29 Top 10 Losses des besten Durchgangs mit BERT 33
Abbildung 30 predictor.explain des Satzes der ID 7501 33
Abbildung 31 Genres der getesteten Bücher und ihre Klassifizierung 34
Abbildung 32 Verteilung der Emotionen im Datenset für Training und Validation 35
Abbildung 33 Aggregierte Prozentzahlen der Emotionen in den Büchern. 35

Abkürzungsverzeichnis

BERT	Bidirectional Encoder Representations from Transformers
biLSTM	Bidirectional Long Short-Term Memory
CSV	Comma-Separated Values File
CNN	Convolutional Neural Networks
CRF	Conditional Random Fields
LSTM	Long Short-Term Memory
NLTK	Natural Language Tool-Kit
NLP	Natural Language Processing
POS	Part-of-Speech
REMAN	Relational Emotion Annotation
SVM	Support Vector Machines
XML	Extensible Markup Language

Sprachkürzel

DE / de	Deutsch
EN / en	Englisch

1 Einleitung

1.1 Motivation

Die Sentiment Analyse betrachtet die Meinungen und Gefühle von Personen in geschriebenen Texten, wobei grundsätzlich mittels Polaritätswerten entschieden wird, ob es sich um negative oder positive Gefühle handelt (Liu, 2015, S. 1-2). Literarische Texte sind bislang in wenigen Studien Inhalt der Sentiment Analyse, wie Schmidt, Burghardt und Dennerlein beschreiben (Schmidt, Burghardt, & Dennerlein, "Kann man denn auch nicht lachend sehr ernsthaft sein?" - Zum Einsatz von Sentiment Analyse-Verfahren für die quantitative Untersuchung von Lessings Dramen, 2018, S. 1). Denn Sentiment Analyse ist in den vergangenen Jahren vor allem in Bereichen von Produkt- und Film-Reviews oder der Filterung von Kommentaren und Posts auf Social Media ein aktives Forschungsfeld geworden (Schmidt, Burghardt, & Wolff, Herausforderungen für Sentiment Analysis-Verfahren bei literarischen Texten, 2018, S. 1).
Emotionen sind jedoch auch in literarischen Texten, wie Erzählungen und Romanen, von grosser Bedeutung und oft verantwortlich für die Handlung des Textes. Ausserdem interpretiert der Erzähler oder die Erzählerin einer Geschichte die Emotionen, damit er oder sie möglichst lebendig erzählen kann. Bei geschriebenen Texten muss die emotionale Bedeutung einer Textstelle zuerst entsprechend identifiziert werden, wobei bei mündlichen Erzählungen Emotionen unter anderem mit Tonhöhe und Intensität ausgedrückt werden (Alm, Roth, & Sproat, 2005, S. 1).

1.2 Fragestellung

In dieser Arbeit soll untersucht werden, ob und mit welcher Qualität mittels dem Sprachmodell BERT (Bidirectional Encoder Representations for Transformers) literarische Texte nach Emotionen klassifiziert werden können. BERT ist ein Sprachmodell, welches ein aktueller State-of-the-Art für maschinelles Lernen darstellt (Devlin, Chang, Lee, & Toutanova, 2019, S. 1). Das Ziel ist ein System, welches Sätze, aber ganze Bücher als Eingabe verwendet und die dominierende Emotion in der entsprechenden Textstelle ausgibt. Auf dem Weg zu diesem System müssen halbautomatisch zusätzliche Daten aus Büchern generiert werden, sodass das System für die Emotionsklassifizierung trainiert werden kann, was ebenfalls Teil der Arbeit ist.

1.3 Aufbau der Arbeit

Zu Beginn werden in Abschnitt 2 die theoretischen Grundlagen erklärt, die eine Basis für die Arbeit darstellen. Dazu gehören die Erklärung von Sentiment- und Emotionsanalyse, die Emotionstheorien und der Stand der Forschung. Im Kapitel der Daten und Methoden wird genauer auf die Emotionsklassifizierung mittels lexikonbasierter Ansätze und den Herangehensweisen über maschinelles Lernen eingegangen. Anschliessend werden die verwendeten Daten, die Korpora und Lexika genauer vorgestellt. Als letzter Teil des Abschnitts 3 werden die einzelnen Programme vorgestellt, welche im Rahmen der Arbeit verwendet oder erstellt wurden.
In der Evaluation werden die verschiedenen Ergebnisse der Emotionsanalyse vorgestellt und analysiert. Um bestmögliche Resultate zu erhalten wurden die Daten jeweils verändert oder gewisse Parameter des maschinellen Systems wurden angepasst, um allfällige Verbesserungen bei der Performance zu erzielen. Diese Versuche sind ebenfalls in der Evaluation festgehalten und genauer erklärt.
Die Diskussion stellt der Abschluss der Arbeit dar, wobei einerseits die Herausforderungen der maschinellen Emotionsanalyse beleuchtet werden, aber auch Ideen vorgestellt werden, wie man gewisse Probleme lösen oder verbessern könnte. Der letzte Teil der Diskussion ist das Fazit, welches die Arbeit zusammenfasst.
Zusätzlich sollte erwähnt werden, dass die Daten, welche die Grundlage für die Arbeit darstellen, englische Texte sind und die dazugehörigen Annotationen ebenfalls englische Begriffe sind. Da diese Arbeit auf Deutsch verfasst wurde, sind einige der Texte, welche für Beispiele oder weitere Erklärungen verwendet wurden, teilweise auf Deutsch übersetzt worden. Bei simplen Sätzen oder mehrfachen Wiederholungen derselben Emotionen wurde auf eine Übersetzung verzichtet.

2 Theoretische Grundlagen

2.1 Sentiment- und Emotionsanalyse

Die Sentiment Analyse ist ein Forschungsgebiet, welches die Meinungen, Gefühle und Einstellungen von Personen in geschriebenem Text analysiert (Liu, 2015, S. 1). Sentiment Analyse klassifiziert Texte grundsätzlich nach Aussagen, die positiver oder negativer Natur sind, wobei jedoch zusätzlich die neutralen Äusserungen, die keinem Gefühl zuzuordnen sind, relevant sind. In dieser Diskussion gibt es zudem die Begriffe des Affekts, der Emotion und der Stimmung, die psychologische Zustände darstellen (Liu, 2015, S. 2). Nicht zu vergessen sind die Zielgruppen der Emotionen, die von der positiven oder negativen Aussage betroffen oder damit gemeint sind (Liu, 2015, S. 2-3). Inwiefern die Zielgruppen relevant sind, wird im Kapitel 5.1.3 aufgezeigt.

Wie Schmidt, Burghardt und Wolff (2018) zusammengefasst haben, gibt es verschiedene Methoden, um Sentiment Analyse durchzuführen. Einerseits gibt es Sentiment Lexika, wobei Wörter mit einem positiven oder negativen Label erfasst sind oder Emotionslexika, in welchen diverse Wörter mit einer Emotion und einem Intensitätswert annotiert sind. Diese werden bei einem lexikonbasierten Ansatz der Sentiment- oder Emotionsanalyse verwendet. Anderseits gibt es verschiedenste Methoden im Bereich des maschinellen Lernens, wobei ein Korpus benötigt wird, um das System entsprechend zu trainieren (Schmidt, Burghardt, & Wolff, Herausforderungen für Sentiment Analysis-Verfahren bei literarischen Texten, 2018, S. 2).

Die Unterscheidung der Sentiment- von der Emotionsanalyse beläuft sich darauf, dass bei der Sentiment Analyse die negative und positive Polarität von Wörtern oder Texten angeschaut wird. Bei der Emotionsanalyse stehen, wie der Name sagt, die Emotionen, beispielsweise Freude oder Wut, im Zentrum. So werden in diesem Fall Textstellen nach ihren Emotionen und nicht nur nach der Polarität analysiert.

Mit der Sentiment Analyse können beispielsweise auf Social Media Hasskommentare herausgefiltert werden, aber auch Bewertungen auf Plattformen im Internet aufgrund ihrer Polarität analysiert und eingeordnet werden. Hierzu werden beispielsweise die Kommentare von einem maschinellen System durchgearbeitet und es wird mittels eines statistischen Ansatzes gemessen, wie viele Hasswörter im Text enthalten sind, was in den Grundsätzen einem lexikonbasierten Verfahren entspricht. Das System kann jedoch auch mit maschinellem Lernen trainiert sein und somit befähigt sein, unangebrachte, beleidigende Kommentare zu entdecken.

2.2 Affekt und Emotion

Wie erwähnt, werden bei der Sentiment- und der Emotionsanalysen Texte einer gewissen Polarität, beziehungsweise einer Emotion zugeordnet. Da diese Arbeit den Fokus auf die Emotionsanalyse legt, wird nun genauer auf den Begriff der Emotionen eingegangen.
Wie Harald C. Traue und Henrik Kessler im Buch zur Natur und Theorie der Emotion (2003) beschreiben, gibt es kaum eine einheitliche Theorie der Emotionen und keine interdisziplinär akzeptierte Definition. Der Grund dafür ist die Komplexität des emotionalen Verhaltens, was dazu führt, dass verschiedenste wissenschaftliche Disziplinen an der Forschung beteiligt sind (Traue & Kessler, 2003, S. 24). Die unterschiedlichen Disziplinen haben jedoch verschiedene Konzepte, die auf die Emotionen angewendet werden, was zu einer Heterogenität bei den Theorien führt (Traue & Kessler, 2003, S. 24).

Das Gabler Wirtschaftslexikon beschreibt Emotion als innere Empfindung, die angenehm oder unangenehm empfunden und mehr oder weniger bewusst erlebt wird, wie beispielsweise die Gefühle Freude, Angst, Kummer und Überraschung. Das Gabler Wirtschaftslexikon beschreibt weiter: «Die Emotion ist ein komplexes Muster aus physiologischen Reaktionen (z.B. Steigerung des Blutdrucks), Gefühlen (z.B. Liebe, Wut), kognitiven Prozessen (Interpretation, Erinnerung und Erwartung einer Person) sowie Verhaltensreaktionen (z.B. lachen, weinen)» (Gabler Wirtschaftslexikon, 2021).

Für die Emotionsanalyse muss zuerst festgelegt werden, auf welches psychologische Modell der Emotionen man sich stützt, um zu klären, wie Texte den Emotionen zugeordnet werden und welche Emotionen für die Analyse in Frage kommen.
Relevant hierfür ist das Verständnis davon, was Emotionen sind und wie die Menschen diese erleben. Dazu wird auf die Analyse von Khurshid Ahmad und Ann Devitt (2013) Bezug genommen. Für die kognitive Darstellung von Emotionen gibt es zwei primäre Ansätze, einerseits Abbildung der Emotionen als finite Kategorien und anderseits die Einordnung mittels Dimensionen: Das dimensionale Modell beschreibt die Emotionen nicht in Kategorien, sondern nach mehreren Dimensionen, auf welchen emotionale Zustände, Dispositionen und affektive Bewertungen eingeordnet werden können (Ahmad & Devitt, 2013, S. 477). Hierzu gibt es beispielsweise die drei Dimensionen: Die Dimension der Valenz (Gut vs. Schlecht), der Aktivierung und Intensität (Aktiv vs. Passiv) und der Dominanz (stark vs. Schwach) (Ahmad & Devitt, 2013, S. 477).

Der kategoriale Ansatz geht davon aus, dass es ein endliches Set von Basisemotionen gibt, welche sich allenfalls innerhalb von verschiedenen Kulturen unterscheiden können (Ahmad & Devitt, 2013, S. 477). Zu diesen Basisemotionen gehören Freude, Traurigkeit, Wut, Ekel und Angst (Ahmad & Devitt, 2013, S. 477).

2.2.1 Basisemotionen nach Ekman

Eine der bekannten Emotionstheorien ist diejenige von Paul Ekman, die an dem kategorialen Ansatz anknüpft (Klinger, 2020, S. 2-3). Die Basisemotionen von diesem Modell beschränken sich auf Freude, Wut, Ekel, Angst, Traurigkeit und Überraschung, wobei sich Ekman auf die Ausdruckskomponente fokussiert, sodass eine Unterscheidung der Emotionen mittels der Mimik möglich ist (Klinger, 2020, S. 2-3).

2.2.2 Emotionsmodell nach Plutchik

Ein weiteres Modell ist das Emotionsrad, welches von Plutchik entwickelt wurde. Bei diesem Modell liegen die Emotionen auf einem Rad, wobei die verschiedenen Basisemotionen auf die Blätter des Rades verteilt sind und die sekundären Emotionen dazwischen platziert wurden (Klinger, 2020, S. 3). Emotionen, die weiter im Zentrum des Kreises liegen, sind von stärkerer Intensität als die dazugehörigen Emotionen, die auf einem Blatt weiter aussen platziert sind (Klinger, 2020, S. 3).

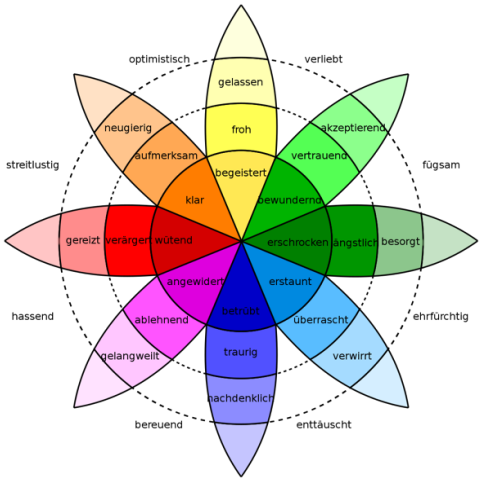

Abbildung 1 Emotionsmodell nach Plutchik (Wikipedia, Zugriff: 2. März 2021)

2.2.3 Circumplexmodell nach Posner, Russell und Peterson

Ein Beispiel für den dimensionalen Ansatz der Einordnung der Dimensionen ist das Modell von Posner, Russel und Peterson. Wobei sich dieser Ansatz grundsätzlich auf die Dimensionen der Valenz und der Aktivierung beschränkt (Kim & Klinger, A Survey on Sentiment and Emotion Analysis for Computational Literary Studies, 2018, S. 7-9). Mit diesem Modell ist es möglich auch subjektive Erfahrungen zu berücksichtigen, die nicht in die Kategorien eingeordnet werden können, wodurch jede Erfahrung als Punkt auf dem Circumplex dargestellt werden kann (Kim & Klinger, A Survey on Sentiment and Emotion Analysis for Computational Literary Studies, 2018, S. 7-9).

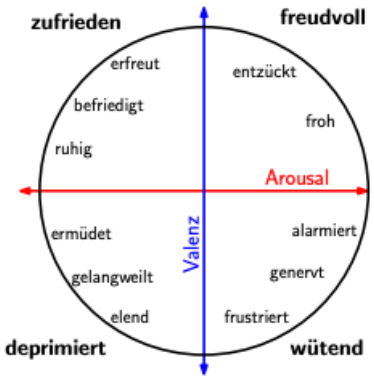

Abbildung 2 Circumplexmodell nach Posner, Russell und Peterson (2005), Bild und Übersetzung von (Klinger, 2020)

2.3 Stand der Forschung

Bis ins Jahr 2000 wurde kaum im Feld der Sentiment Analyse geforscht, da Texte mit Meinungen bis zu diesem Zeitpunkt kaum in einer digitalen Form existierten (Liu, 2015, S. 3). Dies obwohl gesprochener und geschriebener Text auch in der Geschichte der Menschen nicht frei von Emotionen und Meinungen war (Liu, 2015, S. 3). Doch mit der Entwicklung und dem Aufschwung von Internet und Social Media gibt es immer mehr Daten für Sentiment Analyse in digitaler Form, was die Forschung in diesem Gebiet vorantreibt (Liu, 2015, S. 3).

Es ist das Konzept von Social Media, dass Menschen dort frei ihre Meinungen äussern und sich frei unterhalten können (Liu, 2015, S. 3). Ausserdem sind für uns Menschen die Meinungen von andern von Bedeutung und spielt für unseren Entscheidungsfindungsprozess eine zentrale Rolle (Liu, 2015, S. 3). Genau deshalb erlebt die Industrie, sowie die Anwendungen rund um Sentiment Analyse in den letzten Jahren einen Aufschwung (Liu, 2015, S. 3).

Im Bereich der literarischen Texte gibt es hauptsächlich Forschungen bei Märchen und Romanen, wie Schmidt, Burghard und Wolff in ihrer Analyse beschreiben (Schmidt, Burghardt, & Wolff, Herausforderungen für Sentiment Analysis-Verfahren bei literarischen Texten, 2018, S. 2). Hierbei nutzen viele Studien lexikonbasierte Verfahren und erst wenige beziehen sich auf Methoden im Feld des maschinellen Lernens, obwohl Emotions- und Sentiment Analysen bessere Qualität aufweisen, wenn sie mit einem Machine-Learning Ansatz durchgeführt werden (Schmidt, Burghardt, & Wolff, Herausforderungen für Sentiment Analysis-Verfahren bei literarischen Texten, 2018, S. 2).

Wie in diesen Arbeiten gezeigt wurde, gibt es wenige Studien, die sich mit Sentiment Analyse oder Emotionsanalyse in literarischen Texten befassen und für die Klassifikation maschinelle Verfahren anwenden. Aus diesem Grund setzt diese Bachelorarbeit an dieser Stelle an und versucht mittels einem aktuellen State-of-the-Art im Bereich des maschinellen Lernens, mit Bidirectional Encoder Representations from Transformers (BERT), ein System zu trainieren, welches versucht, literarische Texte nach Emotionen zu klassifizieren. Die Arbeit zeigt ausserdem weitere mögliche Gründe und Schwierigkeiten auf, weshalb es auf diesem Gebiet erst wenige Studien gibt.

3 Daten und Methoden

3.1 Emotionsklassifizierung

Für die Sentiment Analyse oder die Emotionsklassifizierung gibt es einerseits den lexikonbasierten Ansatz und andererseits das Verfahren mittels maschinellen Lernens (Schmidt, Burghardt, & Dennerlein, "Kann man denn auch nicht lachend sehr ernsthaft sein?" - Zum Einsatz von Sentiment Analyse-Verfahren für die quantitative Untersuchung von Lessings Dramen, 2018, S. 2). Innerhalb dieser beiden Kategorien gibt es zahlreiche Ansätze für die Sentiment- oder Emotionsanalyse (Schmidt, Burghardt, & Wolff, Herausforderungen für Sentiment Analysis-Verfahren bei literarischen Texten, 2018, S. 2).

3.1.1 Lexikonbasierter Ansatz

Lexikonbasierte Verfahren sind regelbasiert und haben Lexika als Grundlage, die auf Crowdsourcing basieren oder von Experten kuratiert sind (Bostan & Klinger, 2018, S. 2106). Solche Lexika sind Listen von Wörtern, die entweder mit den Emotionen oder mit den Angaben zur Polarität annotiert sind. Problematisch ist dieser Ansatz bei Texten, die sich je nach Genre oder Domäne stark unterscheiden (Schmidt, Burghardt, & Wolff, Herausforderungen für Sentiment Analysis-Verfahren bei literarischen Texten, 2018, S. 2). Wie Schmidt, Burghardt und Wolff beschreiben, sind für Klassifikationen bei literarischen Texten die lexikonbasierten Methoden verbreiteter als Machine Learning-Ansätze. Ein Grund dafür ist, dass es kaum Korpora gibt, wo literarische Texte mit Emotionen oder Werten zur Polarität annotiert sind, wobei diese Korpora die Grundlage wären, um einen Algorithmus mittels maschinellen Lernens zu trainieren. (Schmidt, Burghardt, & Wolff, Herausforderungen für Sentiment Analysis-Verfahren bei literarischen Texten, 2018, S. 2).

3.1.2 Machine Learning Ansatz

Wie Liu im Buch «Sentiment Analysis: Mining Opinion, Sentiments and Emotions» beschreibt, ist Sentiment Klassifikation ein Problem im Bereich Textklassifizierung, wobei dort oft Texte verschiedenen Themen zugeordnet werden, wie Politik oder Sport, statt wie bei der Sentiment Analyse verschiedenen Emotionen oder Polaritäten (Liu, 2015, S. 49). Bei Textklassifizierungen sind oft die themenspezifischen Wörter relevant, bei Sentiment Analyse sind Wörter von Bedeutung, die Meinungen und Gefühle implizieren (Liu, 2015, S. 49).
Die im Abschnitt 3.1.1 erklärten lexikonbasierten Ansätze liefern in diesen Bereichen jedoch schlechtere Resultate als superversierte Lernansätze (Bostan & Klinger, 2018, S. 2107). Dies

zeigen auch Schmidt, Burghardt und Wolff, die ausserdem erklären, dass für maschinelle Ansätze Korpora benötigt werden, was im Bereich der literarischen Texte eine Herausforderung ist, da es wenige Korpora gibt, wo die Texte mit ihrer Polarität oder den Emotionen annotiert worden sind (Schmidt, Burghardt, & Wolff, Herausforderungen für Sentiment Analysis-Verfahren bei literarischen Texten, 2018, S. 2).

Da in dieser Arbeit mit dem Bidirectional Encoding Representations from Transformers (BERT) gearbeitet wird, was den maschinellen Verfahren zugeordnet werden kann und da, wie oben erklärt, die Ansätze mit maschinellem Lernen oftmals bessere Resultate liefern, wird folgend kurz erklärt, welche Möglichkeiten es in diesem Bereich gibt.

Die Erkennung von Emotionen ist eine Art der Textklassifizierung, was wiederum ins Themenfeld des Natural Language Processing (NLP) gehört (Gonzáles-Carvajal & Garrido-Merchán, 2020, S. 1). Eine Lösung für Klassifikationsprobleme ist ein linguistischer Ansatz, welcher auf den Eigenschaften des Textes beruht (feature-basiert) (Gonzáles-Carvajal & Garrido-Merchán, 2020, S. 1).
Es gibt verschiedene Methoden des überwachten, feature-basierten Lernens, die bei der maschinellen Klassifikation verwendet werden können, wie Naïve Bayes oder Support Vector Machines (SVM) (Liu, 2015, S. 49). Auf diese Methoden wird nicht näher eingegangen, da sich nicht Teil der Arbeit sind.
Zudem folgte in der Forschung ein Paradigmenwechsel von feature-basierten Ansätzen zu einem Deep-Learning Ansatz, was auch die Emotionsklassifikation prägte, die nun oft auf neuronalen Netzen basiert (Bostan & Klinger, 2018, S. 2107). Beispiele hierfür sind Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory (BiLSTM) und Long Short-Term Memory (LSTM) (Bostan & Klinger, 2018, S. 2107).

Bei dem Ansatz über Deep Learning bieten annotierte Korpora die Grundlage und mittels Machine Learning werden die relevanten Merkmale des Textes abgeleitet (Gonzáles-Carvajal & Garrido-Merchán, 2020, S. 2). Deep Learning bedeutet auf Deutsch «mehrschichtiges Lernen», was eine Methode des maschinellen Lernens (Machine Learning) ist (Deng & Yu, 2014, S. 199-200). Deng und Yu definieren Deep Learning unter anderem folgendermassen: «A class of machine learning techniques that exploit many layers of non-linear information processing for supervised or unsupervised feature extraction and transformation, and for pattern analysis and classification.» (Deng & Yu, 2014, S. 199-200).

Nun wird genauer auf die Unterschiede zwischen dem feature-basiertem, linguistischen Verfahren und dem Deep Learning Ansatz eingegangen.
Ein Vorteil für den oben erklärten, linguistischen Ansatz ist die hohe Precision, wobei jedoch der Recall eher tiefe Werte aufweist (Gonzáles-Carvajal & Garrido-Merchán, 2020, S. 1-2). Ansätze mit maschinellem Lernen sind vor allem dann beliebt, wenn man mit grossen Mengen an Daten zu tun hat (Gonzáles-Carvajal & Garrido-Merchán, 2020, S. 2). Das grosse Problem an linguistischen Ansätzen ist der Faktor, dass sie auf Regeln basieren. Das heisst, wenn man die Sprache ändert, muss man aufgrund der damit verbundenen Änderungen in Satzstruktur und Alphabet auch die Regeln anpassen. Ausserdem müssen die Regeln für jede Textsorte angepasst werden, da beispielsweise in formalen Texten nicht im gleichen Stil geschrieben wird, wie in Beiträgen auf Twitter (Gonzáles-Carvajal & Garrido-Merchán, 2020, S. 2).
Eine Lösung für diese Probleme bietet das Sprachrepräsentationsmodell BERT (Bidirectional Encoder Representations from Transformers) (Gonzáles-Carvajal & Garrido-Merchán, 2020, S. 2). BERT ist ein mächtiges Tool und aktuell ein Lösungsansatz für die Verarbeitung von natürlicher Sprache und beruht auf dem neusten Stand der Technik (Devlin, Chang, Lee, & Toutanova, 2019, S. 1). BERT besteht aus zwei Teilen: dem Pre-training und dem Fine-tuning. Im Schritt des Pre-trainings wurde das Modell auf blossen Textdaten trainiert (unlabeled), wobei im Fine-Tuning dann annotierte Daten dazu kommen (labeled text) (Devlin, Chang, Lee, & Toutanova, 2019, S. 1). In dieser Arbeit wird auf weitere Erläuterungen zur genauen Funktionsweise von BERT verzichtet, da dies über den Rahmen des Kapitels hinaus geht. Weitere technische Details zu BERT und genauere Beschreibungen finden sich in der Arbeit von Devlin, Chang, Lee und Toutanova von 2019 «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding».
Wie BERT in dieser Bachelorarbeit verwendet wird, wird im Kapitel 3.3 genauer erläutert.

3.2 Korpora und Lexika

3.2.1 Annotierter Goldstandard

Als Basis für die Emotionsklassifikation mit einem maschinellen Ansatz wird ein annotierter Goldstandard benötigt. In dieser Arbeit spielen vor allem literarische Texte eine zentrale Rolle, deshalb sollte der Goldstandard aus Sätzen und Phrasen bestehen, die ebenfalls der Prosa zugeordnet werden können. Aus diesem Grund wurde als das REMAN (Relational Emotion ANnotation) Korpus der Universität Stuttgart verwendet (Kim & Klinger, Who Feels What and Why? Annotation of a Literature Corpus with Semantic Roles of Emotions, 2018). Das Korpus besteht aus englischen Texten aus dem Projekt Gutenberg, wobei die Phrasen nach Emotionen, dem

Empfindenden, dem Ziel und der Ursache der Emotion annotiert sind. Die annotierten Emotionen beschränken sich auf die Kategorien nach dem Modell von Plutchik: Wut, Angst, Vertrauen, Ekel, Freude, Trauer, Überraschung, Erwartung und dem Zusatz «Andere», für alle Emotionen, die keiner Kategorie zugeordnet werden konnten. (Kim & Klinger, Who Feels What and Why? Annotation of a Literature Corpus with Semantic Roles of Emotions, 2018). Die verwendete Emotionstheorie wurde im Abschnitt 2.2.2 genauer beschrieben.
Die Sätze wurden grundsätzlich mit dem Ziel annotiert, dass eine Emotion an einem konkreten Wort festgemacht wurde, ausser die Emotion wurde mit einer ganzen Phrase oder indirekt ausgedrückt. Zudem wurde die Entität festgelegt, welche definiert, ob eine Person, ein Event oder etwas anderes im Zentrum steht und schliesslich wurden die semantischen Verbindungen zwischen den Emotionen und den Phrasen festgelegt (Kim & Klinger, Who Feels What and Why? Annotation of a Literature Corpus with Semantic Roles of Emotions, 2018).

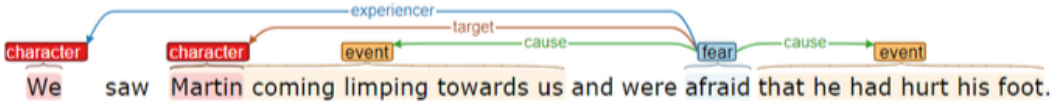

Abbildung 3 Beispiel der Annotation aus dem Annotationshandbuch des REMAN Korpus (Screenshot aus Guidlines.pdf, Downloadordner REMAN Korpus) (Kim & Klinger, Who Feels What and Why? Annotation of a Literature Corpus with Semantic Roles of Emotions, 2018)

Das Korpus besteht aus 1720 Textstellen, die aus 200 Büchern des Genres Fiktion aus dem Projekt Gutenberg extrahiert wurden. Diese 1720 Textstellen bestehen jeweils aus drei Sätzen, von welchen der mittlere Satz ein Wort aus dem NRC Lexikon von Mohammad und Turney enthält (Mohammad & Turney, 2013) (Kim & Klinger, Who Feels What and Why? Annotation of a Literature Corpus with Semantic Roles of Emotions, 2018).

Für die Arbeit sind nur die Annotationen der Emotionen relevant. Aus diesem Grund mussten die Daten vorgängig verarbeitet werden, sodass lediglich die Sätze und annotierten Phrasen sowie deren zugehörige Emotion extrahiert wurden. Zudem wurden alle Sätze, die mit «other-emotion» (de: «andere Emotion») annotiert worden waren, für diese Arbeit nicht weiter berücksichtigt. Die Art und Weise der Aufbereitung der Daten wird im Abschnitt 3.3 genauer erklärt. Die extrahierten Phrasen und Sätze wurden schliesslich in einer CSV-Datei abgespeichert, sodass sie zu einem späteren Zeitpunkt für das Trainieren und Testen des Klassifikationssystems verwendet werden konnten. Als CSV-Datei sind die Daten besser zugänglich und gut von den verwendeten Skripten einlesbar.

Alle im Rahmen der Arbeit erstellen CSV-Dateien sind nach dem gleichen Prinzip aufgebaut, damit das Einlesen in Skripte und Klassifikationssysteme möglichst problemlos funktionieren kann. Ausserdem sind die Daten dadurch besser vergleichbar, ohne dass sie erneut verarbeitet werden müssen.

Untenstehend findet sich ein Auszug aus der CSV-Datei, welche auf dem REMAN Korpus basiert. Hier wurden einerseits die annotierten Phrasen erfasst, wobei im Beispiel in Abbildung 4 die Phrase «of the best sort» die Emotion «joy» impliziert. Andererseits sind die ganzen Sätze aufgeführt, welche die annotierte Phrase enthalten und in welchen folglich die Emotion schliesslich über den ganzen Satz hinweg geltend ist.

text	**label**
…	…
She would brighten up greatly at this, taking it for a compliment of the best sort.	joy
of the best sort	joy
…	…

Abbildung 4 Auszug aus der CSV-Datei basierend auf dem REMAN Korpus

Aus der Aufbereitung der Daten resultierten 3177 annotierte Phrasen, Wörter und Sätze, basierend auf dem REMAN Korpus. Im Datenset sind einige Sätze mehrfach enthalten, da teilweise mehrere Emotionen in einer Textstelle enthalten sein können und folglich ein Satz mit mehreren Emotionen annotiert worden ist.

Da die Emotionen in dieser Arbeit eine zentrale Rolle spielen, wurde zu Beginn analysiert, mit welchem Anteil diese Emotionen in den annotierten Daten vorhanden sind. Hier fällt auf, dass «joy» (de: Freude) mit Abstand die grösste Häufigkeit aufweist und die anderen Emotionen mit einer ähnlichen Verteilung vorkommen.

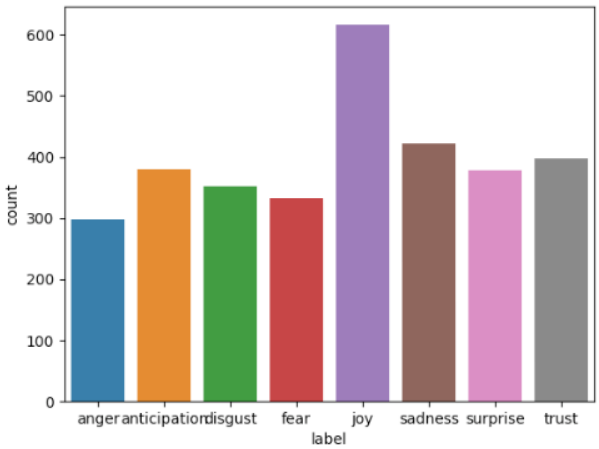

Abbildung 5 Emotionsverteilung der verarbeiteten Daten (REMAN) als Diagramm

joy	616
sadness	422
anticipation	398
trust	380
disgust	378
fear	352
surprise	332
anger	298

Abbildung 6 Absolute Zahlen der Emotionsverteilung (REMAN)

3.2.2 NRC Lexikon

Da das NRC Lexikon von Mohammad und Turney einerseits eine Grundlage bietet für das REMAN Korpus und andererseits für diese Bachelorarbeit von Relevanz ist, wird nun genauer darauf eingegangen. Mohammad und Turney (Mohammad & Turney, 2013) haben festgestellt, dass in vergangener Zeit in der Forschung bezüglich der Polarität von Texten viel Wert auf die Erstellung diverser Polaritäts-Lexika gelegt wurde. Für Emotionsanalysen gibt es hingegen nur sehr wenige und kleine Lexika. Mit Crowdsourcing erstellten sie deshalb eine Kollektion diverser Emotionslexika, wobei ein Teil davon ein Intensitäts-Emotionslexikon ist, welches in verschiedenen Sprachen existiert. Das englische Intensitätslexikon ist hierbei für diese Bachelorarbeit von Bedeutung.

word	**emotion**	**emotion-intensity-score**
happiest	joy	0.986
happiness	joy	0.984
bliss	joy	0.971
celebrating	joy	0.970
jubilant	joy	0.969
ecstatic	joy	0.954
elation	joy	0.944
beaming	joy	0.938
…	…	…
achieved	joy	0.500
soulful	joy	0.500
welcoming	joy	0.500
satisfied	joy	0.500
family	joy	0.500
meritorious	joy	0.500
…	…	…

Abbildung 7 Ausschnitt aus dem Intensitätslexikon für die Emotion «joy» (de: Freude)

Dieses Lexikon besteht aus einer Liste mit knapp 10'000 englischen Wörtern, die einerseits nach den acht Basisemotionen (nach Plutchik) annotiert wurden und andererseits jeweils mit einem Intensitätswert versehen sind. Dieser Wert gibt eine Auskunft darüber, wie stark ein gewisses Wort zur annotierten Emotion gehört. Ein Beispiel für die Emotion «Freude» beziehungsweise in Englisch «joy», wären die Wörter «celebrating» (de: feiern) und «satisfied» (de: zufrieden), welche zwar beide der Emotion «Freude» zugeordnet werden, jedoch eine unterschiedliche Stärke an Freude ausdrücken. So hat «celebrating» einen Intensitätswert von 0.97 und «satisfied» einen Wert von 0.50, wie der Abbildung 7 zu entnehmen ist.

Damit zu einem späteren Zeitpunkt mit dem Emotionslexikon gearbeitet werden kann, wurde im Rahmen der Bachelorarbeit vorgängig die darin enthaltene Emotionsverteilung analysiert. Von den 9921 Emotionen im Lexikon gibt es am meisten Wörter, die mit der Emotion «fear» (de: Angst) annotiert wurden. Für die Emotion «surprise» (de: Überraschung) existieren am wenigsten Annotationen, was man den

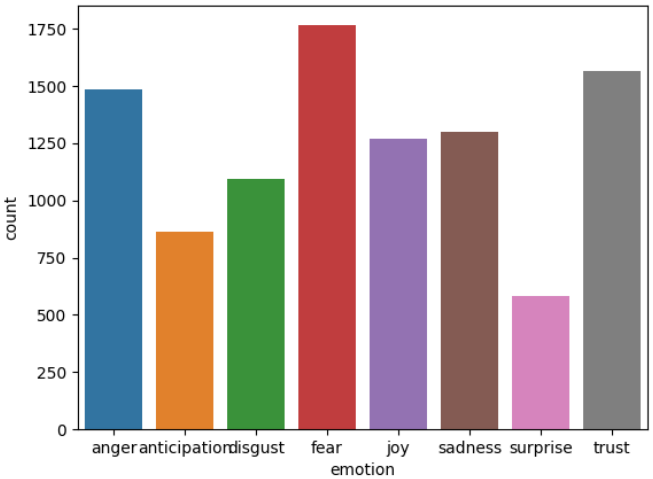

Abbildung 8 Emotionsverteilung im Intensitätslexikon

Abbildungen 8 und 9 entnehmen kann. Dies kann bei der Generierung eines halbautomatisch annotierten Silver Standard zu einer Herausforderung werden. Die Vermutung ist, dass wenn wenige Wörter mit der Emotion «surprise» annotiert sind, diese Emotion folglich in Texten vergleichbar weniger vorhanden sein wird.

fear	1765
trust	1564
anger	1483
sadness	1298
joy	1268
disgust	1094
anticipation	864
surprise	585

Abbildung 9 Absolute Zahlen der Emotionsverteilung im Intensitätslexikon

3.2.3 Halbautomatisch annotierter Silver Standard

Das REMAN Korpus besteht wie bereits erwähnt aus 1720 Textstellen. Bei der versuchten Emotionsklassifzierung mit BERT hat sich herausgestellt, dass die Grösse der Daten nicht ausreicht, um präzise Emotionsvorhersagen zu treffen, da BERT mit dieser Datengrundlage lediglich eine Genauigkeit von maximal 50 Prozent erzielt hat. Maximal bedeutet in diesem Kontext, dass bei mehreren Durchgängen und verschiedenen Anpassungen der Parameter, beim besten Durchgang höchstens 50 Prozent erreicht wurden. Aus diesem Grund wurde das Datenset mit einem Silver Standard erweitert. Der erwähnte Silver Standard besteht aus halbautomatisch annotierten Sätzen aus 80 Büchern des Projekt Gutenberg. Halbautomatisch bedeutet in diesem Fall, dass die Sätze mit einem Python-Skript aus den Büchern gefiltert wurden und mittels eines lexikonbasierten Ansatzes annotiert wurden. Auf den technischen Teil dieser Annotation wird später in Abschnitt 3.3 genauer eingegangen. In dieser Bachelorarbeit impliziert die Bezeichnung «Silver Standard», dass die Daten – im Vergleich zu einem Gold Standard – nicht vollumfänglich verifiziert sind, da diese maschinell annotiert wurden. Deshalb werden im Rahmen dieser Arbeit von nun die beiden Begriffe Gold und Silver Standard verwendet.
Um jedoch einen Überblick über die Qualität des Silver Standards zu erhalten, wurden die Annotationen mit einer Stichprobe überprüft. Die Auswertung der Stichprobe findet sich unter Kapitel 4.1.

Ein Ziel bei der Generierung eines Silver Standards war es, dass über die generierten Sätze eine gleichmässige Verteilung der Emotionen besteht. Dazu braucht es jedoch als Grundlage eine grosse Menge an Büchern, in welchen nach möglichen annotierbaren Sätzen gefiltert wird. Aus dem Filter resultierte ein Datenset mit rund 10’000 Sätzen, die mit den acht Basisemotionen annotiert wurden, wobei dort die Emotionen nicht gleich verteilt sind. Genauere Erläuterungen dazu befinden sich im Abschnitt 3.3.2.

3.3 Klassifikationssystem

3.3.1 Preprocessing des REMAN Korpus

Die Daten des REMAN Korpus sind in einer XML-Datei (Extensible Markup Language) festgehalten. Um die Daten besser zugänglich zu machen, sodass sie später für Trainings- und Testzwecke benutzt werden konnten, wurden die Daten mit einem Python Programm aus der XML-Datei extrahiert. Dem «Readme» des REMAN Korpus ist zu entnehmen, dass nur die Knoten mit der Bezeichnung «adjudicated» zu verwenden sind, da die «other»-Knoten zurückgewiesene Annotationen beinhalten. Das «Readme» findet sich im Anhang, beziehungsweise im Zusatzmaterial dieser Arbeit wieder.

Schliesslich wurde durch die Knoten iteriert, wobei alle annotierten Phrasen gesucht wurden, die mit einer Emotion der acht Basisemotionen annotiert waren («trust», «sadness», «joy», «anger», «anticipation», «surprise», «disgust», «fear»).

Ein Span hat einige Attribute, wobei prioritär die Attribute «type» und «modifier» beachtet wurden. Im Attribut «type» sind jeweils die Emotionen gespeichert. Das Attribut «modifier» kann die Werte «negated», «weak» und «strong» haben, wo jeweils für diese Arbeit keine Sätze und Phrasen extrahiert wurden, bei welchen der Modifikator negiert war. Denn in einem solchen Fall tritt die im Typ erfasste Emotion nicht auf. Da es schwierig ist, diese Information entsprechend für weitere Schritte für die Klassifizierung aufrecht zu erhalten und dem Klassifizierungssystem mitzugeben, wurden diese Phrasen nicht extrahiert. Ebenso wurden wie in vorhergehenden Abschnitten erwähnt, die Relationen zwischen den Emotionen und somit die Knoten mit dem Tag «relations» nicht berücksichtigt.

Folgend ein Auszug aus der XML-Datei, welcher die erwähnten Tags und Attribute aufzeigt. Die komplette Datei ist im Zusatzmaterial zu finden.

```
. . .
<document author="Ballantyne, R. M. (Robert Michael)" author_death_year="1894"
        book_title="The Wild Man of the West: A Tale of the Rocky Mountains"
        doc_id="24617|93859" genre="Indians of North America -- Juvenile fiction, Adventure
        and adventurers -- Juvenile fiction, Animals -- Juvenile fiction, Mothers and sons
        -- Fiction, Outdoor life -- Fiction, Natural history -- Juvenile fiction, Rocky
        Mountains -- Juvenile fiction, Conduct of life -- Juvenile fiction"
        url="http://www.gutenberg.org/ebooks/24617">
   <text>I see no prospect of a better state of things, for savage nature cannot be
        changed. It seems a hopeless case." There was a touch of pathos in the tone in which
        this was said that was very different from McLeod's usual bold and reckless manner.
   </text>
        <adjudicated>
           <spans>
              <span annotation_id="48266" annotatorId="B" cbegin="51" cend="82"
               type="other">savage nature cannot be changed</span>
              <span annotation_id="46208" annotatorId="C" cbegin="93" cend="109"
               type="sadness">a hopeless case.</span>
              <span annotation_id="46215" annotatorId="A" cbegin="95" cend="103"
               modifier="weak" type="fear">hopeless</span>
              <span annotation_id="46222" annotatorId="B" cbegin="95" cend="103"
               modifier="negated" type="anticipation">hopeless</span>
              <span annotation_id="47426" annotatorId="A" cbegin="104" cend="108"
               type="other">case</span>
              <span annotation_id="48280" annotatorId="B" cbegin="203" cend="211"
               type="character">McLeod's</span>
           </spans>
           <relations>
              <relation annotatorId="B" left="51" relation_id="48273" right="103"
               source_annotation_id="46215" target_annotation_id="48266" type="cause">
               savage nature cannot be changed[OTHER]...hopeless[FEAR]</relation>
              <relation annotatorId="B" left="95" relation_id="48287" right="211"
               source_annotation_id="46215" target_annotation_id="48280"
               type="experiencer">hopeless[FEAR]...McLeod's[CHARACTER]</relation>
              <relation annotatorId="A" left="104" relation_id="47433" right="109"
               source_annotation_id="46208" target_annotation_id="47426" type="target">
               a hopeless case.[SADNESS]...case[OTHER]</relation>
           </relations>
        </adjudicated>
        <other>
           <spans />
           <relations />
        </other>
</document>
. . .
```


Abbildung 10 Auszug aus der XML-Datei des REMAN Korpus

Pro gefundene Emotion wurden zwei Dictionaries zu einer Liste hinzugefügt. Ein erster Dictionary mit der Phrase und der dazugehörigen Emotion und ein zweiter Dictionary mit dem ganzen Satz und der Emotion. Dieser Satz ist der mittlere der drei Sätze, die im Knoten «text» hinterlegt sind.

Diese Liste von Dictionaries wurde dann in eine CSV-Datei geschrieben, mit den zwei Spalten «text» und «label».

3.3.2 Generierung des Silver Standards

Um zusätzliche Daten zu sammeln, mit dem Ziel die Performance von BERT zu verbessern, wurde, wie in Abschnitt 3.2.3 erwähnt, mit einem Python Programm ein Silver Standard generiert. Die verwendeten Bücher werden als Textdateien einzeln eingelesen und mithilfe der NLTK (Natural Language Processing Toolkit) Bibliothek in Sätze tokenisiert. Die einzelnen

Sätze werden dann dem Filter übergeben. Dieser Filter gibt den Wert «True» (de: Wahr) zurück, wenn ein Satz den Filter passiert und «False» (de: Falsch), wenn eine Eigenschaft auftritt, welche rausgefiltert werden soll. In diesem Fall werden betroffene Sätze nicht weiterverwendet. Zu den erwähnten Eigenschaften gehören Negationen wie beispielsweise «no», «not», «neither» und weitere. Diese könnten die Art und Weise, wie eine Emotion erlebt wird, verändern. Deshalb werden solche Sätze nicht berücksichtigt. Dasselbe gilt für Fragesätze, die auf ein Fragezeichen enden. Ebenfalls entfernt wurden Textstellen, die Links enthalten, Kapitelüberschriften oder Wörter, die nur in Grossbuchstaben geschrieben sind. Alle diese Textstellen indizieren, dass es sich nicht um den eigentlichen Text des Buches handelt, sondern um Titel und Informationen über das Buch und das Projekt Gutenberg. Die Berücksichtigung solcher Sätze würde ein verrauschtes Datenset zur Folge haben.
Die Sätze, die den Filter passiert haben, werden anschliessend mit dem NRC-Lexikon annotiert. Die Annotation basiert auf den Intensitätswerten des Lexikons, wobei Schwellenwerte erreicht werden müssen, damit ein Satz mit einer Emotion annotiert wird.

Als erstes Beispiel, um den Prozess zu illustrieren, wird folgender Satz verwendet:
« *That same night she made a discovery which increased her apprehension almost to terror.* »
Das Programm annotiert hier «apprehension» mit «fear» und einem Wert für die Intensität von 0.453. Ebenso ist «terror» im Lexikon enthalten mit einem Wert von 0.953.

That	*same*	*night*	*. . .*	*her*	***apprehension*** fear 0.453	*almost*	*to*	***terror*** fear 0.953

Abbildung 11 Darstellung der Annotation durch den lexikonbasierten Ansatz des Filters zur Generierung des Silver Standards

Das Programm addiert nun die Werte, die es pro Emotion im Satz findet und speichert diese folgendermassen in einem Dictionary ab: { 'fear': 1.406 }.
Da es hier nur eine Emotion im Satz hat, greift hier die Bedingung, dass der Wert grösser als 1 sein muss, da dann mindestens zwei Wörter im Satz mit derselben Emotion annotiert sein müssen und die Emotion mit genügender Intensität auftritt.

Ein zweites Beispiel deckt die Fälle ab, bei welchen in einem Satz mehrere Emotionen enthalten sind:
« *Yet nearer, and there began to come to our ears a great sound of mourning, the people on board and those on the shore crying and lamenting one to another so as to pierce the heart.* »

Hier wird mit der tabellarischen Darstellung gezeigt, wie das Programm den Satz annotiert. Zur besseren Übersicht werden Satzzeichen bei der Visualisierung weggelassen und unannotierte Teilsätze mit drei Punkten abgekürzt.

Yet	*nearer*	. . .	***mourning***	*the*	*people*	*on*	***board***	*and*	. . .	***crying***	*and*	***lamenting***	. . .	***heart***
			sadness				anticipation			sadness		sadness		joy
			0.969				0.234			0.781		0.656		0.547

Abbildung 12 Darstellung einer weiteren Annotation durch den lexikonbasierten Ansatz des Filters zur Generierung des Silver Standards

Daraus resultiert folgendes Dictionary: {'sadness': 2.406, 'anticipation': 0.234, 'joy': 0.547}, mit den summierten Intensitätswerten für die entsprechenden Annotationen.
Da in diesem Beispiel mehrere Emotionen existieren, berechnet das System nun die Verhältnisse zwischen den Emotionen. Über alle Emotionen gesehen ergibt dies hier einen Wert von 3.137. Dies bedeutet wiederum, dass «sadness» mit einem Verhältnis von 75.5 Prozent am stärksten vertreten ist und über dem im Rahmen dieser Arbeit gesetzten Mindestwert von 75 Prozent liegt. Dieser Mindestwert wurde gesetzt, damit ein Satz nur dann mit einer Emotion annotiert wird, wenn diese genug oft oder stark in einem Satz vorhanden ist. So fallen einzelne Wörter, die nichts zur Emotion beitragen, nicht zu stark ins Gewicht, wie im oberen Beispiel das Wort «heart» zeigt.

Die Sätze, welche den Filter passiert haben und eine Emotion zu einem genug hohen Wert enthalten, werden in einer CSV-Datei abgespeichert.
Das Verhältnis der Emotionen zueinander war zunächst sehr unausgeglichen und trotz der Erweiterung der Daten durch das Durchsuchen von weiteren Büchern, wurde die Anzahl der untervertretenen Labels wie «disgust» und «surprise» kaum erhöht.

In der Evaluation unter Abschnitt 4.3. wird überprüft, ob diese ungleiche Verteilung einen negativen Einfluss auf das Trainieren von BERT hat. Für diese Überprüfung wurde eine Auswahl aus dem Silver Standard verwendet, bei welcher von jeder Emotion maximal 600 Sätze enthalten waren. Um dies zu erreichen, wurden bei all den Emotionen, bei welchen mehr als 400 Sätze im Datenset enthalten waren, zufällig 600 Sätze ausgewählt, bei den restlichen beiden Emotionen wurden alle vorhandenen, annotierten Sätze verwendet. Daraus folgt eine

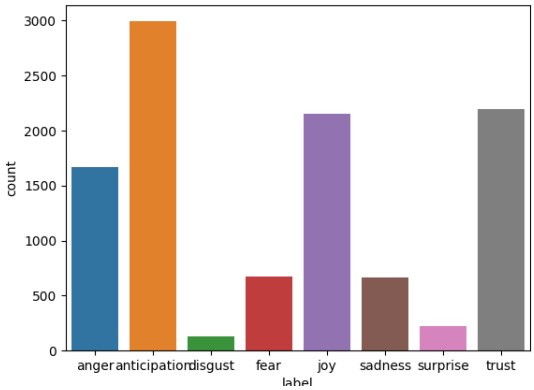

Abbildung 13 Verteilung der Emotion der vollständigen Version des Silver Standards (10702 Sätze)

Selektion des Datensets mit 3950 annotierten Sätzen. Die vollständige Version des Silver Standards enthält 10702 Sätze. Die Sätze wurden alle mit den acht Basisemotionen annotiert.
Aus der Liste von halbautomatisch annotierten Sätzen des Silver Standards wurden schliesslich mit einem Skript eine zufällige Stichprobe von 150 Datenzeilen extrahiert. Diese Stichprobe wurde manuell durchgearbeitet und die Annotationen wurden überprüft. Die Qualität der Stichprobe wird in der Evaluation unter 4.1 untersucht.

3.3.3 BERT

Wie in den theoretischen Grundlagen erklärt wurde, kann das Sprachmodell BERT (Bidirectional Encoder Representations from Transformers) unter anderem für Aufgaben im Bereich der Textklassifikation verwendet werden (Devlin, Chang, Lee, & Toutanova, 2019, S. 1). Dieses Modell gehört aktuell in diesem Anwendungsbereich zum State-of-the-Art und ausserdem hat die Recherche gezeigt, dass sich erst wenige Arbeiten mit der Emotionsanalyse mithilfe von BERT befasst haben. Aus diesem Grund wurde für diese Bachelorarbeit BERT trainiert, um die Emotionen zu klassifizieren.
Die Grundlage für das Trainieren liefert «ktrain», eine Aggregation der Bibliothek TensorFlow Keras (Maiya, 2020). «ktrain» erweist sich als gutes Package, da es viele Funktionen besitzt, die für diese Arbeit relevant sind und den Aufbau eines Klassifikationssystems erleichtern.
Eine weitere Quelle für den Code liefert zudem ein Python Notebook von der Webseite «Kaggle», welches ebenfalls Sentiment Analyse mithilfe von BERT durchführt (Chakraborty, 2020). Aufgrund der Tatsache, dass für diese Arbeit ein gesamtes Datenset verwendet wurde, musste dieses für Testen und Training entsprechend aufgeteilt werden. Hierfür wurden die Grundlagen aus dem Notebook des Assignment 5 «News Classification using BERT» aus dem Kurs «Introduction to Machine Learning» vom Herbstsemester 2020 bei Dr. Manfred Klenner an der Universität Zürich verwendet.
Parameter, wie die maximale Länge der Input Sequenz, das Modell und die Batch-Grösse wurden im Rahmen der Arbeit mehrfach angepasst und verändert, um herauszufinden, welche Einstellungen die besten Resultate erzielen. Grundsätzlich erwies sich die maximale Länge von 256 Tokens bei der Input Sequenz als sinnvoll, ein Batch Size von 10 sowie das Modell «Bert-base-case». Die Lernrate und die Anzahl der Epochen beim Training wurden für die verschiedenen Datensets jeweils angepasst, sodass bestmögliche Resultate erzielt wurden. Im Falle der Daten des REMAN Korpus ergaben 20 Epochen mit einer Lernrate von 1e-5 die optimale Einstellung. Bei den aggregierten Daten aus dem REMAN Korpus und dem Silver Standard erzielten 10 Epochen bei einer Lernrate von 1e-4 eine höhere Genauigkeit.

3.3.4 Emotion Predictor – Erweiterung von BERT

Das Ziel der Arbeit war ein System, welches einzelne Sätze oder längere Textstellen als Input verwendet und die entsprechende Emotion ausgibt. Deshalb musste das trainierte Modell noch weiter ausgebaut werden. Dieses wurde nach den erwähnten 10 Epochen abgespeichert und kann nun im Google Colab «Emotion Predictor» hochgeladen werden. Aufgrund des Modells gibt dieses System einerseits die Emotion aus, welche den Text zu dominieren scheint, aber auch eine Übersicht darüber, für welche Emotion sich das Modell zu welchen Prozentzahlen entschieden hat.

Der «Emotion Predictor» fragt zuerst die Benutzenden, ob sie eine ganze Textdatei hochladen, oder manuell einen Text eingeben möchten. Werden eine oder mehrere Textdateien hochgeladen, wird über diese iteriert und die Emotionen der einzelnen Dokumente werden separat aufgeführt und aufgeschlüsselt ausgegeben. Schreiben die Benutzenden manuell einen Text in das Eingabefeld, sagt das Modell ebenfalls die Emotion des eingegebenen Textes voraus und gibt die aufgeschlüsselte Verteilung der Emotionen aus. Ist dieser Vorgang abgeschlossen, fragt das System, ob noch mehr Text klassifiziert werden soll, worauf die Benutzenden mit ja oder nein antworten können. Würde das Modell erneut trainiert werden, für den Fall, dass bessere oder andere Daten zur Verfügung stehen, könnte das neue Modell problemlos im oberen Teil des Codes eingelesen werden und das System würde anhand dessen über die Emotion entscheiden. Untenstehend ein Ausschnitt aus dem Output des Codes, für den Upload des Buches «A dreamers palace» aus dem Gutenberg Projekt.

```
Hello! I am the Emotion Predictor.
If you input me a book or a part of a text, I will tell you the emotions occurring in this text! Maybe it helps you analyzing the text.
Please note: I am not perfect, I am just a machine, I cannot show any emotions, so it was hard for me to learn to detect emotions in texts. But I will try my best!


Do you want to upload a file or input text directly?
Write «upload» or «write» into the box please.
 -->upload
Dateien auswählen a_dream..._tales.txt
• a_dreamers_tales.txt(text/plain) - 217177 bytes, last modified: 26.4.2021 - 100% done
Saving a_dreamers_tales.txt to a_dreamers_tales.txt

##############################################################################

UPLOADED TEXT: a_dreamers_tales
Downloading: 100% 213k/213k [00:00<00:00, 417kB/s]

Downloading: 100% 436k/436k [00:00<00:00, 1.19MB/s]

__________________________
Input Text:  Title: A Dreamers Palace  Author: Lord Dunsany [Edward J. M. D. Plunkett]    PREFACE   I hope for this book that it may come into the hands of those th
__________________________
The Emotion Predictor would classify this text as:       ANGER
__________________________
The exact distribution of how the predictor decided for:

joy             6.15
trust           0.67
sadness         0.44
anticipation    0.33
disgust         0.53
surprise        0.14
fear            0.23
anger           91.51
__________________________
Do you want to input another text?
Write «yes» or «no» into the box please.
 -->no

BYEBYE!
```

Abbildung 14 Emotion Predictor Output nach einem Dateiupload

Ein weiteres Beispiel von einem manuell eingegebenen Text, der korrekterweise als «fear» klassifiziert wurde.

```
_____________________________
Do you want to input another text?
Write «yes» or «no» into the box please.
 -->yes
Do you want to upload a file or input text directly?
Write «upload» or «write» into the box please.
 -->write
Input your text: This house is haunted. I am sure that something strange is going on here. Let's move on quickly.
_____________________________
Input Text: This house is haunted. I am sure that something strange is going on here. Let's move on quickly.
_____________________________
The Emotion Predictor would classify this text as:       FEAR
_____________________________
The exact distribution of how the predictor decided for:

joy             0.07
trust           0.06
sadness         0.33
anticipation    0.07
disgust         0.36
surprise        13.36
fear            85.44
anger           0.32
_____________________________
Do you want to input another text?
Write «yes» or «no» into the box please.
 -->no

BYEBYE!
```

Abbildung 15 Emotion Predictor Output nach eingegebenem Text

4 Evaluation

Im Kapitel der Evaluation werden diverse Datensets und Systeme, die im Rahmen dieser Arbeit geschrieben, produziert und verändert wurden, analysiert. Einige Evaluationen führten zum Verwerfen einiger Ideen und Hypothesen, wobei jeweils in den entsprechenden Abschnitten erklärt wird, wenn eine Strategie aus bestimmten Gründen nicht weiterverfolgt oder abgeändert wurde.

Einleitend sollte bemerkt werden, dass bei den Evaluationen, den Tabellen und Beispielen zu den Top Losses, die Resultate der besten Trainingsdurchgänge verwendet wurden. Bei erneutem Training können leicht andere Werte resultieren.

4.1 Überprüfung der Stichprobe

Der Silver Standard wurde, wie in Abschnitt 3.3.2 genauer erläutert wird, mithilfe eines Skripts generiert. Die maschinelle Annotation von Sätzen anhand eines lexikonbasierten Ansatzes hat zur Folge, dass es grundsätzlich keine Kontrollinstanz gibt, die die Klassifikation der Sätze überprüft, wie dies bei einer manuellen Annotation mit beispielsweise drei Annotierenden der Fall wäre. Aus diesem Grund wurde aus dem Set des Silver Standards eine zufällige Stichprobe von 150 Sätzen extrahiert, wobei die Emotionen in einer zufälligen Anzahl vorkommen.

Die Stichprobe wurde manuell von zwei Annotierenden durchgearbeitet und mit den Emotionen klassifiziert. Anschliessend wurden die Annotationen mit dem Silver Standard verglichen.

Die Bedingungen für die überprüfenden Annotationen waren folgende: Den Annotierenden lagen nur die Sätze vor, ohne Emotionslabel, welche sie klassifizieren mussten. Sätze, die gemäss den Annotierenden keine Emotion enthalten, wurden mit «nothing» annotiert. Es handelte sich hierbei oft um einleitende Sätze rund um das Gutenberg Projekt. Grundsätzlich galt die Regel, dass sich die beiden Annotierenden klar für eine Emotion entscheiden müssen, auch wenn mehrere enthalten sind oder die Emotion im Satz aus anderen Gründen unklar ist. Dies war die Vorgabe, da sich der Silver Filter ebenfalls für eine Emotion entscheiden musste. Ist es für die Annotierenden nicht möglich sich zu entscheiden, musste die gewählte Emotion an erster Stelle gesetzt werden und die restlichen, möglichen Emotionen folgen danach. Diese Informationen über nicht eindeutige Emotionen sind für die Evaluation und die Diskussion dieser Arbeit von Relevanz, da dort unter anderem die Schwierigkeit der Emotionsklassifizierung festgestellt werden kann. Um die Übereinstimmungen der Annotierenden zu messen, wurde jeweils nur die zuerst notierte Emotion verwendet.

Um einerseits den Silver Standard zu überprüfen, aber auch um zu messen, wie schwierig eine menschliche, manuelle Annotation sein kann, wurden die nachfolgenden Tabellen erstellt. Diese sollten die Genauigkeiten und Übereinstimmungen der Annotierenden aufzeigen.

	Annotator 1 (Silver Filter)	**Annotator 2**	**Annotator 3**	**Total**
anger	19	15	15	**49**
anticipation	25	27	28	**80**
disgust	9	8	9	**26**
fear	18	23	15	**56**
joy	17	19	22	**58**
trust	17	16	14	**47**
sadness	24	19	23	**66**
surprise	21	20	19	**60**
Rest	-	3	5	**8**
Total	**150**	**150**	**150**	**450**

Abbildung 16 Anzahl der annotierten Emotionen pro Annotierenden (Annotator 1 ist der Silver Filter beziehungsweise der Silver Standard, Annotator 2 und 3 sind menschliche Annotierende)

	A1 vs. A2	**A1 vs. A3**	**A2 vs. A3**	**Total**
anger	13	11	10	34 = 69.4%
anticipation	21	18	16	55 = 68.75%
disgust	8	8	7	23 = 88.5%
fear	15	11	12	38 = 67.86%
joy	16	15	14	45 = 77.59%
trust	15	12	12	39 = 82.98%
sadness	17	17	15	55 = 91.67%
surprise	19	19	17	49 = 74.24%
Total	**124 = 82.67%, κ = 0.80**	**111 = 74%, κ = 0.70**	**103 = 68.67%, κ = 0.64**	**75% (Mittelwert)**

Abbildung 17 Übereinstimmung zwischen den Annotierenden. A1 steht für Annotator 1 usw. Für die Prozentzahlen in der rechten Spalte wurden jeweils die Summe durch das Total der oberen Tabelle dividiert, um die Genauigkeit bei den Emotionen zu berechnen. Die unterste Spalte zeigt jeweils das Verhältnis der totalen Übereinstimmungen der Annotierenden zur Gesamtzahl von 150. 75% ist der Durchschnitt über alle Prozentzahlen der Übereinstimmungen.

In Abbildung 16 findet sich die Anzahl annotierten Label pro Emotion. Dass insgesamt vom Label «disgust» am wenigsten Annotationen vorhanden sind und von «anticipation» am meisten, erscheint sinnvoll. Da auch im Datenset des Silver Standards eine ähnliche Verteilung vorliegt, wie in Abschnitt 3.2.3 aufgezeigt wurde.

Abbildung 17 zeigt die Übereinstimmungen der Annotierenden, wobei die Annotierenden mit «A» abgekürzt wurden und jeweils zwei miteinander verglichen wurden. Es wurde festgestellt, dass der Silver Standard (A1) mit der ersten annotierenden Person am meisten Übereinstimmungen hat. Im Schnitt haben somit die Annotierenden eine Übereinstimmung von 75 Prozent. Im Rahmen dieser Arbeit sollten diese Werte ausreichend sein, um den Silver Standard weiter zu verwenden, da die Übereinstimmungen zwischen menschlichen Annotierenden und der

maschinellen Annotation mit 82.7 und 74 Prozent sogar sind ist als die Übereinstimmung zwischen den beiden manuellen Annotationen (A2 und A3) mit 68.7 Prozent.

Schliesslich wurde die Interrater Reliabilität über alle Werte hinweg berechnet, jeweils für jedes Annotierenden-Paar. Für den Silver Filter und den Annotierenden 1 ergab dies ein Kappa von 0.80, für den Silver Filter und den zweiten Annotierenden ergab es ein Kappa von 0.7 und für den Vergleich zwischen beiden menschlichen Annotierenden ergab es ein Kappa von 0.64.
Diese Werte sind zufriedenstellen, zumal nach Landis und Koch (1977) Werte von 0.61 bis 0.80 als beachtliche Übereinstimmung gelten (Landis & Koch, 1977).
Auch die Universität Stuttgart hat bei der Erstellung des REMAN Korpus die Übereinstimmungen der Annotierenden mittels dem Kappa Koeffizienten gemessen und erhielt dort jeweils paarweise tiefere Werte, was jedoch an der Eindeutigkeit der Klassifikation der Sätze liegen kann. Ausserdem liegen nicht dieselben Berechnungen zugrunde und die Annotationsweise weicht leicht ab. Die genauen Werte und Grundlagen der Berechnungen der Universität Stuttgart bei der Annotation vom REMAN Korpus findet sich in der Arbeit von Evgeny Kim und Roman Klinger (Kim & Klinger, Who Feels What and Why? Annotation of a Literature Corpus with Semantic Roles of Emotions, 2018, S. 1350).

Diese Analyse zeigt grundsätzlich auf, dass das Klassifizieren von Emotionen für Menschen, sowohl als auch für Maschinen eine Herausforderung ist, da oftmals keine eindeutige Entscheidung getroffen werden kann. Im Abschnitt 5.1 wird näher auf diese Schwierigkeiten bei der Klassifizierung eingegangen.

4.2 BERT und der Gold Standard aus dem REMAN Korpus

Vorerst musste evaluiert werden, ob das REMAN Korpus ausreicht, um BERT so zu trainieren, dass das System neue Texte mit einer grossen Genauigkeit und Qualität klassifizieren kann.
Hierzu wurden die nach Abschnitt 3.3. aufbereiteten Phrasen des REMAN Korpus verwendet und in ein Trainings- und ein Testset aufgeteilt. Nach mehrfachem Verändern der einzelnen Komponenten beim Finetuning, wurde schliesslich über alle Emotionen hinweg betrachtet, eine durchschnittliche Genauigkeit von 50 Prozent erzielt. Dies betrifft den Durchgang mit der bestmöglichen Performance aufgrund der eingestellten Parameter des Finetunings. Zu diesen Parametern gehören unter anderem die Anzahl der Epochen, was die Durchgänge des Trainings meint, aber auch die maximale Länge der Tokens und die Lernrate.

Wie in Abschnitt 3.3.1 erwähnt, wurden die Sätze und Phrasen, die mit «other-emotion» (de: andere Emotion, kurz: «other») annotiert wurden, aus dem Set der Trainingsdaten entfernt. Der Grund dafür ist die schlechtere Performance von BERT im Falle der Berücksichtigung der «other»-Labels. Es wird vermutet, dass dies daran liegen kann, dass die «other»-Labels von grösserer Häufigkeit sind und somit BERT grosszügig Sätze, bei welchen aus Sicht des Systems keine Zuordnung möglich ist, mit «other» klassifiziert.
Aus diesem Grund wurden die entsprechenden Sätze entfernt, sodass BERT eine Entscheidung treffen muss und Sätze und Phrasen in jedem Fall einer der acht Basisemotionen zuordnen muss. Folgend ein Vergleich der Konfusionsmatrizen der Datensets mit und ohne dem «other»-Label. Hierbei wird klar, weshalb die Entscheidung für ein Datenset ohne die «other»-Labels fiel.

	precision	recall	f1-score	support
joy	0.56	0.57	0.57	437
trust	0.44	0.51	0.47	281
sadness	0.49	0.40	0.44	306
anticipation	0.38	0.37	0.37	257
disgust	0.40	0.44	0.42	241
surprise	0.56	0.56	0.56	250
fear	0.47	0.42	0.44	245
anger	0.34	0.38	0.36	204
other-emotion	0.32	0.30	0.31	287
accuracy			0.45	2508
macro avg	0.44	0.44	0.44	2508
weighted avg	0.45	0.45	0.45	2508

Abbildung 18 Konfusionmatrix nach Training und Testing von BERT. Daten inkl. «other»-Label

	precision	recall	f1-score	support
joy	0.67	0.55	0.60	442
trust	0.49	0.54	0.51	276
sadness	0.45	0.51	0.48	293
anticipation	0.37	0.45	0.40	256
disgust	0.48	0.48	0.48	242
surprise	0.67	0.53	0.59	281
fear	0.44	0.47	0.46	231
anger	0.38	0.39	0.39	203
accuracy			0.50	2224
macro avg	0.49	0.49	0.49	2224
weighted avg	0.51	0.50	0.50	2224

Abbildung 19 Konfusionmatrix nach Training und Testing von BERT. Daten ohne «other»-Label

Diese Resultate wurden jeweils mit 20 Trainings-Epochen und einer Lernrate von 1.5e-5 erreicht. Weiter lieferte bei beiden das Modell «bert-base-case» ein besseres Resultat als «bert-base-uncased». Zudem erwies sich eine maximale Länge der Token von 256 als optimal.

Um in einem nächsten Schritt zu diskutieren, welche Sätze dem System auf Basis der Trainingsdaten am meisten Schwierigkeiten bereiten, werden die Daten untersucht, die beim Testing die schlechteste Performance (= Top Losses) aufweisen.
Der nachfolgenden Tabelle können die Top 10 der Verluste abgelesen werden. Im folgenden Abschnitt wird nun genauer auf die einzelnen Problemstellen eingegangen und analysiert, worin jeweils die Schwierigkeit liegt.

ID	Loss	True	Predicted	Sätze / Phrasen
666	8.01	fear	joy	Back to Kennington Road by omnibus, riding outside, her eyes and hair doing execution upon a young man in a very high collar, who was, she saw, terribly tempted to address her, but, happily for himself, could not pluck up courage.
1823	7.82	fear	joy	When I saw Niles pass through the street and the crowd following, I was afraid that a matter that's very serious to some of us would be turned into horseplay, and so I came along, too.
2	7.71	joy	fear	awe
726	7.69	fear	joy	Through all its tremor, there was a look of constancy that greatly pleased me.
1172	7.14	surprise	sadness	disgraceful
123	7.09	anticipation	sadness	It seemed a tragedy to want nothing--and yet he wanted something, something.
1971	7.07	surprise	sadness	Gwynne turned the cold surprise of his eyes upon her.
1301	7.03	anger	sadness	envy
2176	7.03	anger	sadness	envy
2013	6.93	surprise	sadness	It gnashes and whispers, completing the gloom with mystery.

Abbildung 20 Top 10 Losses von BERT mit Trainingsdaten aus dem REMAN Korpus

Einerseits haben einzelne Wörter wie «disgraceful» (de: schändlich, erbärmlich) und «envy» (de: Neid, Eifersucht) schlecht performt. Es fällt auf, dass bei der Annotation von «disgraceful» beim Gold Standard mit «surprise» ebenfalls keine zutreffende Emotion klassifiziert wurde. Um herauszufinden, weshalb der Gold Standard folgendermassen annotiert wurde, werden die ursprünglichen Daten des REMAN Korpus nochmals untersucht.

Die folgende Textstelle ist die Grundlage der Annotation, wobei für die Klassifizierung hauptsächlich der mittlere Satz berücksichtigt wurde:

« Forty thousand roubles--paid down on the nail!" ***The scene was growing more and more disgraceful; but Nastasia Philipovna continued to laugh and did not go away****. Nina Alexandrovna and Varia had both risen from their places and were waiting, in silent horror, to see what would happen. »*

Das Adjektiv «disgraceful» kann in diesem Kontext mit grosser Wahrscheinlichkeit mit «erbärmlich» übersetzt werden. Folglich bedeutet der Satz übersetzt: «Die Szene wird immer erbärmlicher, aber Natasia Philipovna lacht weiterhin und geht nicht weg.»

Bei genauerem Analysieren der Annotationsdaten fällt auf, dass am Wort «disgraceful» mehrere Emotionen festgemacht wurden, und zwar: «disgust», «surprise» und «sadness», wobei alle drei Annotierenden eine Übereinstimmung bei «disgust» zeigen.

Da für den Satz und die Phrase mehrere Annotationen existieren, waren diese auch dreifach im Datenset, mit welchem BERT trainiert und getestet wurde, enthalten. Es lässt sich daraus schliessen, dass BERT im Training bereits auf die Annotation von «disgraceful» mit dem Label «sadness» gestossen ist und schliesslich beim Testing das Wort entsprechend annotiert hat.

Für eine weitere Überprüfung wurde noch das Intensitäts-Lexikon zu Rate gezogen, um zu untersuchen, was gemäss diesem eine passende Emotion für das Wort «disgraceful» sein kann. Hier ergibt sich, dass das Lexikon «disgraceful» mit einer Intensität von 0.625 als «disgust» oder mit einem Wert von 0.547 als «anger» einstuft.

Aus dieser Analyse wird abschliessend geschlossen, dass diese Fehlannotation von BERT kaum berücksichtigt werden kann, da das Wort mehrfach im Datenset vorkommt, wobei es jeweils mit einer anderen Emotion annotiert wurde. Zudem kann «disgraceful» je nach Zusammenhang anders annotiert werden. Da die Annotierenden eine Übereinstimmung beim Label «disgust» haben, was das Lexikon ebenfalls bestätigt, kann man abschliessend sagen, dass diese Emotion in diesem Fall am zutreffendsten wäre.

Da für die weiteren einzeln annotierten Wörter im Datenset vergleichbare Beobachtungen gemacht werden konnten, wird daraus geschlossen, dass es schwierig ist, einzelne Wörter zu klassifizieren, da jeweils der Kontext, in welchem das Wort steht, fehlt. Der Kontext ist jedoch für eine Klassifizierung essenziell, wie das obere Beispiel mit «disgraceful» gezeigt hat. Dies erklärt das Problem bei Wörtern wie «envy» und «awe» in der Tabelle 20.

Nachfolgend werden noch weitere Beispiele aus Tabelle 20 beleuchtet, die eine schlechte Performance aufweisen. Ein falsch klassifizierter Satz ist folgender, welcher gemäss Gold Standard das Label «fear» trägt, von BERT aber mit «joy» gelabelt wurde:

« *Back to Kennington Road by omnibus, riding outside, her eyes and hair doing execution upon a young man in a very high collar, who was, she saw, terribly tempted to address her, but, happily for himself, could not pluck up courage.* »

Um genauer zu analysieren, woran BERT in diesem Beispiel gescheitert ist, wurden mit der Funktion *predictor.explain()* von *ktrain* mehr Informationen zur Annotation ausgegeben.

y=joy (probability **0.978**, score **4.410**) top features

Contribution?	Feature
+5.193	Highlighted in text (sum)
-0.783	<BIAS>

back to kennington road by omnibus, riding outside, her eyes and hair doing execution upon a young man in a very high collar, who was, she saw, terribly tempted to address her, but, happily for himself, could not pluck up courage.

Abbildung 21 Output von predictor.explain() zur Klassifizierung eines Satzes durch BERT

Diese Funktion markiert diejenigen Wörter in grün, welche für die falsche Klassifikation ausschlaggebend waren. Die rot markierten Wörter sprechen für eine Klassifikation mit dem eigentlichen, richtigen Label, in diesem Fall «fear».

Das Wort «happily» scheint die Klassifikation stark beeinflusst zu haben, sodass BERT den Satz abschliessend mit der Emotion «joy» klassifiziert hat.

Zum Vergleich folgt ein Blick in die ursprünglichen Daten des REMAN Korpus. «happily» wird als «joy» annotiert, «could not pluck up courage» wird mit «fear» annotiert, wobei der junge Mann («young man») derjenige ist, der die Emotion erlebt und die Frau («she») ist das Ziel.

Zusammenfassend kann man sagen, dass es auf die Sichtweise ankommt. Schaut man die Textstelle aus der Sicht der Frau an, könnte «joy» eine passende Emotion sein, da sie grundsätzlich froh scheint, dass er sie nicht angesprochen hat. Aus der Sichtweise des Mannes ist jedoch «fear» die passende Emotion, da er nicht mutig genug ist, die Frau anzusprechen, scheint er ängstlich zu sein. Daraus wird klar, dass sich hierbei die Klassifizierung durch ein maschinelles System schwierig gestaltet.

Ein weiteres Beispiel, an welchem man die Entscheidungsweise von BERT gut erkennt, ist das Folgende.

y=sadness (probability **0.989**, score **4.849**) top features

Contribution?	Feature
+5.409	Highlighted in text (sum)
-0.560	<BIAS>

it seemed a tragedy to want nothing--and yet he wanted something, something.

Abbildung 22 Output von predictor.explain() zur Klassifizierung eines weiteren Beispielsatzes durch BERT

Hier wäre gemäss dem Gold Standard die Emotion «anticipation» passend, was nach dem REMAN Korpus am Wort «wanted» festgemacht wurde. «tragedy» wurde zusätzlich mit «sadness» annotiert.

In Abbildung 22 sieht man, dass BERT bei der Klassifizierung ähnliche Punkte berücksichtigt hat. Somit scheint «tragedy» den grössten Einfluss auf die Entscheidung gehabt zu haben. Trotzdem hat BERT «wanted» als ein Wort wahrgenommen, welches für das Label «anticipation» spricht, was man der roten Markierung entnehmen kann.

Die zuvor beschriebenen Resultate mit einer Genauigkeit bei der Klassifikation von 50 Prozent sind die bestmöglichen Ergebnisse, die im Rahmen dieser Arbeit mit BERT und der entsprechenden Aufbereitung des REMAN Korpus erreicht werden konnten. Für die bestmöglichen Werte wurde das Set mit Phrasen und Sätzen benötigt. Beachtet man nur die Sätze aus dem REMAN Korpus, resultieren schlechtere Werte beim Testen von BERT.

Hierbei stellt sich jedoch das Problem, dass der Silver Standard nur aus Sätzen und nicht wie der Gold Standard des REMAN Korpus aus Sätzen und Phrasen besteht. Für eine bessere Vergleichbarkeit wird BERT deshalb noch mit einem Datenset trainiert, welches nur aus den Sätzen des REMAN Korpus besteht, obwohl dieses Modell schlechtere Werte liefert, wie die folgende Analyse genauer aufzeigt.

	precision	recall	f1-score	support
joy	0.45	0.44	0.44	218
trust	0.28	0.24	0.26	134
sadness	0.31	0.37	0.34	147
anticipation	0.18	0.27	0.22	129
disgust	0.31	0.27	0.29	119
surprise	0.48	0.43	0.46	136
fear	0.40	0.32	0.35	123
anger	0.27	0.25	0.26	106
accuracy			0.34	1112
macro avg	0.34	0.32	0.33	1112
weighted avg	0.35	0.34	0.34	1112

Abbildung 23 Konfusionsmatrix von BERT mit ganzen Sätzen aus dem REMAN Korpus

Mit einer für diesen Fall optimalen Lernrate von 2e-5 und 20 Epochen, erreichte BERT 34 Prozent.

Die bestmögliche Genauigkeit von 50% ist nicht zufriedenstellend, jedoch kann dieser Wert mit denjenigen der Universität Stuttgart verglichen werden.

Kim und Klinger (2018) haben in ihrem Projekt mit dem REMAN Korpus ein Baseline Model erstellt, um die manuellen Annotationen zu rekonstruieren. Sie haben verschiedene Einstellungen ausprobiert, um nebst den Emotionen auch die Entitäten, die von der Emotion betroffenen und die Ursachen der Emotion maschinell zu annotieren (Kim & Klinger, Who Feels What and Why? Annotation of a Literature Corpus with Semantic Roles of Emotions, 2018, S. 1352). Da in dieser Bachelorarbeit jedoch bei den Daten nur die Emotionen verwendet wurden, werden beim Vergleich der Baseline Resultate nur die Experimente für die Emotionserkennung berücksichtigt. Kim und Klinger haben verschiedene Modelle untersucht. Einerseits verwenden sie einen lexikonbasierten Ansatz, welcher auf Regeln basiert und einen sogenannten Bag-of-Words-Ansatz mit einem mehrschichtigen Perzeptron. Weiter versuchen sie die Annotationen mittels eins Modells zu rekonstruieren, welches auf Conditional Random Fields (CRF) basiert und mit einem Bidirectional Long Short-Term Memory mit CRF-Layers (biLSTM-CRF) (Kim & Klinger, Who Feels What and Why? Annotation of a Literature Corpus with Semantic Roles of Emotions, 2018, S. 1352-1353). Auf die Funktionsweise dieser Modelle wird an dieser Stelle nicht weiter eingegangen. Details können der Arbeit von Kim und Klinger (2018) entnommen werden (Kim & Klinger, Who Feels What and Why? Annotation of a Literature Corpus with Semantic Roles of Emotions, 2018).

Bei einer strikten Auswertung haben Kim und Klinger über alle Modelle hinweg betrachtet, höchstens einen F1-Wert von 43% erhalten (Kim & Klinger, Who Feels What and Why? Annotation of a Literature Corpus with Semantic Roles of Emotions, 2018, S. 1353). Eine

strikte Auswertung heisst in diesem Fall, dass die maschinellen Annotationen der Modelle exakt den manuellen Annotationen des Gold Standards entsprechen müssen. Für eine bessere Vergleichbarkeit wurden für diese Arbeit deshalb nur die Resultate der strikten Auswertung der Universität Stuttgart betrachtet. Beim Stuttgarter Modell, welches auf biLSTM-CRF basiert, wurden mit 43% die besten Werte erzielt, die anderen Modelle erreichten jeweils zwischen 11-31%, wobei die ausführlichen Resultate der Arbeit von Kim und Klinger entnommen werden können (Kim & Klinger, Who Feels What and Why? Annotation of a Literature Corpus with Semantic Roles of Emotions, 2018, S. 1353).
Da BERT in dieser Bachelorarbeit mit dem Gold Standard mindestens einen F1-Score von 34% und maximal 50% erreicht, kann gesagt werden, dass BERT hierbei ein Modell ist, auf welchem weiter aufgebaut werden kann, da es grundsätzlich bessere Resultate liefert als die Modelle, welche Kim und Klinger (2018) geprüft haben.

4.3 BERT und der Silver Standard

Da BERT mit den Daten des REMAN Korpus bei jeweils 50 beziehungsweise 34 Prozent richtig entscheidet, wurde versucht mit halb-automatisch annotierten Daten die Klassifizierung durch BERT zu verbessern. Dazu wurde BERT mit dem Silver Standard trainiert. Ähnlich wie bei den Daten des REMAN Korpus wurden die Sätze des Silver Standards von BERT eingelesen und mit einer Lernrate von 1.5e-5 und 20 Epochen trainiert.
Dieser Vorgang lieferte eine Genauigkeit von 81 Prozent, was 31 Prozentpunkte höher ist als nach dem Training mit den Daten aus dem REMAN Korpus.

Da die Emotionen im Silver Standard sehr ungleich verteilt sind, wurde zuerst überprüft, welche Auswirkung die Verteilung der Emotionen auf die Trainingsresultate hat. Dazu wurde BERT zuerst mit der Selektion von knapp 4000 Sätzen trainiert, wobei 66 Prozent erreicht wurden. In diesem komprimierten Datenset sind von jeder Emotion maximal 600 Sätze enthalten.
Anschliessend wurde BERT mit denselben Einstellungen erneut trainiert, jedoch mit dem gesamten Datenset, welches aus 10702 Sätzen besteht.

	precision	recall	f1-score	support
joy	0.90	0.84	0.87	1498
trust	0.86	0.77	0.81	1557
sadness	0.60	0.74	0.66	461
anticipation	0.83	0.85	0.84	2085
disgust	0.40	0.31	0.35	95
surprise	0.80	0.72	0.76	159
fear	0.56	0.71	0.62	463
anger	0.86	0.85	0.86	1174
accuracy			0.81	7492
macro avg	0.73	0.72	0.72	7492
weighted avg	0.82	0.81	0.81	7492

Abbildung 24 BERT mit komplettem Silver-Datenset

	precision	recall	f1-score	support
joy	0.73	0.62	0.67	427
trust	0.63	0.58	0.60	418
sadness	0.71	0.70	0.71	428
anticipation	0.54	0.58	0.56	420
disgust	0.37	0.36	0.36	87
surprise	0.76	0.71	0.73	160
fear	0.60	0.71	0.65	406
anger	0.77	0.79	0.78	419
accuracy			0.66	2765
macro avg	0.64	0.63	0.63	2765
weighted avg	0.66	0.66	0.66	2765

Abbildung 25 BERT mit Selektion des Silver Standards

Aus den Abbildungen 24 und 25 wird klar, dass der vollständige Silver Standard die besseren Resultate liefert, da dieser eine Genauigkeit von 81 Prozent erreicht. Deshalb wird für die weiteren Schritte das Set aus 10702 annotierten Sätzen verwendet.
Um zu analysieren, wo bei diesem Modell Fehler entstehen, werden erneut die Sätze ausgegeben, die bei diesem Durchgang die schlechtesten Resultate geliefert haben. Dies bedeutet, sie weisen bei «Loss» die grössten Werte auf. Dieser Wert des «Loss» beschreibt die Grösse des Fehlers.

ID	Loss	True	Predicted	Sätze / Phrasen
244	9.43	fear	joy	So we are forced to the second—that Madame Renauld lied for the sake of the man she loved—or in other words, for the sake of Georges Conneau.
2570	9.36	disgust	joy	They seem for the most part shabby in attire, dingy of linen, lovers of billiards and brandy, and cigars and greasy ordinaries.
1372	9.26	anticipa-tion	anger	This, of itself, was enough to set us in good humour, but the richest part of it was this, that the two thirstiest men that ever came out of Scotland (Mr. Shuan being dead) were now shut in the fore-part of the ship and condemned to what they hated most--cold water.
4547	9.23	anticipa-tion	joy	Then the Sage went up to Michael and looked him hard in the face awhile, and then said: "Yea, honest he is unless the kindred of Michael of the Hatch of Swevenham have turned thieves in the third generation."
1780	9.21	joy	anticipation	To add to their embarrassment, this same son was now appearing over the Lustich Hill--an opportune moment for a pleasing digression.
4281	9.17	trust	joy	Wonderful lawns all round, and shrubberies and a lake that you can just see where the ground dips beyond the fields.
4413	9.0	anticipa-tion	trust	He always acted on the principle of making himself acquainted with every road and track and by-path, every stream, pond, river, and spring in the land.
5119	8.99	joy	anticipation	So monsieur turned on, meeting new treasures at every rise of a leaf; and at the end of the book he came upon two sheets of paper, of much more recent date than anything he had yet seen, which puzzled him considerably.
2347	8.99	anger	sadness	Mr. Carmyle bought it from some lord or other who had been losing money on the Stock Exchange.
2332	8.96	anger	joy	None of the lamps that hung on the walls had been turned up; the silence was oppressive, cheerless.

Abbildung 26 Top 10 Losses des Trainings mit dem Silver Standard

Aus Abbildung 26 wird klar, dass die Vorhersage von BERT besser sein kann als der annotierte Silver Standard. Als Beispiel folgender Satz:

« *So monsieur turned on, meeting new treasures at every rise of a leaf; and at the end of the book he came upon two sheets of paper, of much more recent date than any-thing he had yet seen, which puzzled him considerably.* »

Dieser Satz wurde im Silver Standard mit der Emotion «joy» annotiert. BERT sagte hier jedoch «anticipation» voraus, was in diesem Satz die passendere Emotion wäre. Ähnliches fällt auf bei ID 4281, wo die Vorhersage von BERT ebenfalls passender beurteilt werden kann. Daraus folgt, dass BERT gewisse Sätze treffender annotiert, als sie mit dem lexikonbasierten Ansatz im Silver Standard klassifiziert wurden.

4.4 BERT mit dem Gold und Silver Standard

Das optimale Klassifikationssystem folgt aus einer Konsolidierung des Gold und des Silver Standards. Beim Mischen der Silver und Gold Annotationen erreicht BERT eine Genauigkeit von 72 Prozent. Dies ist noch tiefer als bei der Verwendung von ausschliesslich dem Silver Standard, ohne die Daten des Gold Standards.

	precision	recall	f1-score	support
joy	0.77	0.77	0.77	1725
trust	0.75	0.69	0.72	1688
sadness	0.54	0.66	0.60	619
anticipation	0.76	0.78	0.77	2235
disgust	0.31	0.33	0.32	212
surprise	0.64	0.56	0.60	294
fear	0.55	0.57	0.56	578
anger	0.78	0.76	0.77	1252
accuracy			0.72	8603
macro avg	0.64	0.64	0.64	8603
weighted avg	0.72	0.72	0.72	8603

Abbildung 27 Gold und Silver Daten als Input für BERT

Um die Performance zu verbessern wurden diverse Variationen des Datensets als Input verwendet. Die Sätze des Gold Standards wurden nochmals gefiltert, da einige Sätze mehrfach annotiert sind und bei der Auswertung aufgefallen ist, dass diese Sätze von BERT falsch annotiert werden. Dieses Problem kann auftreten, weil die Daten zufällig in Training- und Test Set aufgeteilt werden und die gleichen Sätze mit unterschiedlichen Emotionen mehrfach im Datenset vorkommen, das heisst, dass ein bestimmter Satz mit der ersten Annotation in das Trainingsset fallen kann und die anders klassifizierte Variante sich im Testset befindet.

Ein Satz ist teilweise mehrfach im Datenset beziehungsweise mit unterschiedlichen Emotionen annotiert, da sich einerseits die Annotierenden nicht einig waren, oder weil mehrere Emotionen im Satz vorkommen.

Um das Problem der Verwirrung von BERT zu lösen, wurden diese Sätze mit einer Python-Funktion nochmals angeschaut und es wurden nur die Sätze verwendet, bei welchen sich das Skript für genau eine Emotion entscheiden konnte. Die Entscheidung erfolgte, wenn ein Satz mit zwei gleichen und einer weiteren Emotion annotiert worden war. Dann wurde die Emotion ausgewählt, welche die höhere Häufigkeit in diesem Satz hat.

Ausserdem wurde, wie oben im Text festgestellt, beim Training und der Validierung mit dem Silver Standard eine Genauigkeit von 81 Prozent erreicht, beim konsolidierten Datenset, bestehend aus Gold und Silver Standard wurden jedoch nur 72 Prozent erzielt. Deshalb wurde überprüft, ob es allenfalls zu besseren Resultaten kommt, wenn BERT mit dem Silver Standard

trainiert und mit dem Gold Standard getestet wird. Doch dies erzielte eine noch schlechtere Performance, wodurch dies wieder verworfen und nicht weiterverfolgt wurde.

Ein weiterer Versuch bestand darin, die Sätze des Gold Standards mit dem Silver Filter zu annotieren und nur die übereinstimmenden Emotionen ins Datenset aufzunehmen. Da der lexikonbasierte Ansatz aber nicht immer gleich annotierte wie die Annotierenden des REMAN Korpus und einige Sätze keine Wörter im Lexikon hatten beziehungsweise wenige Wörter mit geringer Intensität, stellte sich dieser Weg nicht als eine geeignete Lösung heraus.

Schlussendlich wurde das Set des Silver Standards nur mit den Sätzen aus dem Gold Standard gemischt, bei welchen sich die oben erwähnte Python-Funktion für eine Emotion entscheiden konnte, beziehungsweise die Sätze, die bereits nur mit einer Emotion annotiert waren.
So befindet sich jeder Satz mit nur einer Emotion annotiert in den Daten. Zudem musste die Lernrate auf 1e-4 angepasst werden und Zahl der Epochen auf 10 reduziert werden. So wurde für den F1-Score einen Wert von 84 Prozent erzielt.

	precision	recall	f1-score	support
joy	0.87	0.86	0.87	1600
trust	0.81	0.88	0.84	1603
sadness	0.74	0.76	0.75	517
anticipation	0.89	0.89	0.89	2177
disgust	0.50	0.40	0.44	146
surprise	0.86	0.75	0.80	226
fear	0.79	0.70	0.74	508
anger	0.87	0.86	0.87	1181
accuracy			0.84	7958
macro avg	0.79	0.76	0.78	7958
weighted avg	0.84	0.84	0.84	7958

Abbildung 28 Konfusionsmatrix von BERT mit Clean Silver und Gold Daten

Bei einem dieser besten Durchgänge des Trainings wurden erneut die Top Losses ausgegeben, um zu analysieren, wo die Schwierigkeiten hierbei liegen. Die Sätze mit den grössten Fehlerwerten sind der Abbildung 29 zu entnehmen. Es ist auffällig, dass bei diesem Durchgang vor allem Sätze mit den Labels «sadness» und «anger» falsch klassifiziert wurden, obwohl diese gemäss der Konfusionsmatrix nicht diejenigen Emotionen, mit der schlechtesten Genauigkeit sind.

ID	Loss	True	Predicted	Sätze / Phrasen
2313	10.01	anger	trust	The disclosure revealed four faint livid marks running parallel across the arm, just above the elbow.
7372	10.0	anger	trust	Our French acquaintance being in evening dress had perforce confined himself in his sartorial eccentricities to a flowing silk knot in place of the more conventional, neat bow.
7517	9.74	sadness	anticipation	Young Pitt, whom he addressed, turned towards him a face from which the ruddy tan of the sea had faded almost completely during those months of captivity.
2997	9.68	anger	trust	If you dare to say so you'll be called on to have some very serious conversation with my father's daughter!
4626	9.66	anger	trust	then father sed have you did ennything rong to ennyone Albert whitch mite want to get even with you and old J. Albert he sed he hadent done rong to a living sole as far as he gnew and he sed i gess George they must have got in the rong side of the house and they ment it for you insted of me and father sed that may be so Albert but it is almity quear that they shood call me J. Albert Clark and hang the sine on your side of the house and J. Albert dident know what to say to this and so he sed i gess that is quear but peeple do quear things sumtimes.
4622	9.44	sadness	anticipation	Colles told me you weren't wanting in intelligence.
2916	9.43	anger	trust	At the barriers he was met by the attorney Willingwood, the chief lawyer who had conducted the Falworth case before the High Court of Chivalry, and who was to attend him during the administration of the oaths before the King.
7501	9.39	sadness	trust	The latter shifted uncomfortably under the confident flash of his blue eyes.
1501	9.33	anger	anticipation	It is the favourite stratagem of our passions to sham a retreat, and to turn sharp round upon us at the moment we have made up our minds that the day is our own.
444	9.32	sadnesss	anticipation	Miss Crawley was now in the habit of writing to Mr. Waxy her solicitor almost every day in the week, for her arrangements respecting her property were all revoked, and her perplexity was great as to the future disposition of her money.

Abbildung 29 Top 10 Losses des besten Durchgangs mit BERT

Ein Beispiel für ein falsch klassifizierter Satz ist folgender: *« The latter shifted uncomfortably under the confident flash of his blue eyes. »,* was auf Deutsch übersetzt heisst: *« Letztere bewegte sich unbehaglich unter dem selbstbewussten Aufblitzen seiner blauen Augen. »*

Diesen Satz klassifizierte das System als «trust», statt «sadness». Die Verwirrung besteht wahrscheinlich darin, dass einerseits «er» selbstbewusst auftritt und «Letztere» fühlt sich unbehaglich. Da stellt sich erneut die Frage, welche Person man in das Zentrum der Emotionsanalyse stellt. Für BERT überwiegt hier das Wort «confident», wie Abbildung 30 aufzeigt.

y=trust (probability **0.989**, score **5.150**) top features

Contribution?	Feature
+5.753	Highlighted in text (sum)
-0.603	<BIAS>

the latter shifted uncomfortably under the confident flash of his blue eyes.

Abbildung 30 predictor.explain des Satzes der ID 7501

4.5 Evaluation des Emotion Predictors

BERT wurde schliesslich ausgebaut, sodass unbekannte Texte als Input gegeben werden können und das Klassifikationssystem die Emotionen erkennt. Hierzu gibt es einerseits eine Übersicht über alle gefundenen Emotionen aus, andererseits nennt das System die Emotion, welche über den ganzen Text gesehen, überwiegt. Wie dieser Output aussieht, wurde im Abschnitt 3.3.4 genauer erklärt.

Um die Performance des Systems genauer anzuschauen, wurden diverse Bücher aus verschiedenen Genres aus dem Gutenberg Projekt als Input verwendet. In Abbildung 31 folgt eine Übersicht über die Bücher und die Genres, welche in das System eingegeben wurden. In der letzten Spalte der Tabelle sind die Emotionen eingetragen, welche der Emotion Predictor ausgegeben hat. Bei Klassifikationen, wo die Entscheidung für eine Emotion bei über 90% liegt, wurde nur diese Emotion in die Tabelle eingefügt. Wenn mehrere Emotionen Prozentzahlen über 5% erhalten haben, wurden diese jeweils zusätzlich aufgeführt.

Genre	Titel	Autor	Emotion: Prozent
Crime	Brood Of The Witch Queen	Sax Rohmer	**joy**: 98.65%
	Dead Man Tell No Tales	E. W. Hornung	**joy**: 99.31%
	Murder In The Gunroom	H. Beam Piper	**anger**: 79.88% **fear**: 13.22% **anticipation**: 5.11%
Fantasy	A Dreamers Tales	Lord Dunsay	**anger**: 91.51%
	Gulliver Of Mars	Edwin L. Arnold	**anger**: 94.86%
	The Gods Of Pegana	Lord Dunsay	**joy**: 31.94%, **trust**: 28.6%, **anticipation**: 23.12%
Horror	Draculas Guest	Bram Stoker	**surprise**: 84.52%, **joy**: 12.09%
	Nightmare Abbey	Thomas Love Peacock	**joy**: 77.71%, **sadness**: 18.38%
	The Turn Of The Screw	Henry James	**anger**: 36.29%, **fear**: 21.08%, **disgust**: 33.69%
	Wuthering Heights	Emily Brontë	**joy:** 99%
Humor	Book Of Nonsense	Edward Lear	**joy**: 99.81 %
	Comic History of USA	Bill Nye	**trust**: 69.52 %, **joy**: 24.94%
	Man On The Bummel	Jerome K. Jerome	**anger**: 96.44%
	The Man With Two Left Feet	P. G. Wodehouse	**anticipation**: 99.82%
Romance	Kate Vernon	Mrs. Alexander	**anger**: 95.07%
	Wastralls	C. A. Dawson Scott	**anger**: 99.68%
Science Fiction	The Hanging Stranger	Philip K. Dick	**anticipation**: 99.73%
	The Hour Of The Dragon	Robert E. Howard	**sadness**: 79.78%, **anticipation**: 16.53%
	The Lost World	Arthur Conan Doyle	**joy**: 95.37%

Abbildung 31 Genres der getesteten Bücher und ihre Klassifizierung

Vergleicht man die Emotionen, die vom System klassifiziert wurden mit dem Genre, scheint es zumindest teilweise Sinn zu ergeben. Als gutes Beispiel dient das Buch «The Turn Of The Screw» aus dem Genre «Horror», bei welchem die Emotionen «anger», «fear» und «disgust» erkannt wurden. Ähnliches ist bei dem Buch «Murder In The Gunroom» festzumachen. Auch dort scheinen die Emotionen in etwa dem Genre zu entsprechen.

Es hat jedoch Bücher, bei welchen die erkannten Emotionen willkürlich scheinen. Beispielsweise bei «Wuthering Heights», einem Buch aus dem Genre Horror, welches jedoch zu 99% «joy» enthält. Es wäre möglich, dass einige Bücher mit der Emotion Freude annotiert wurden, wenn diese am Schluss ein Happy End haben. Dies ist jedoch lediglich eine Hypothese und müsste bei einer allfälligen Weiterführung dieser Arbeit genauer untersucht werden.
Weiter fällt auf, dass die erkannten Emotionen in den Büchern den Verteilungen der Emotionen im Datenset entsprechen. Die Verteilung im Datenset ist der Abbildung 32 zu entnehmen, wodurch klar wird, dass die Emotionen «anticipation», «joy» und «trust» und «anger» am häufigsten vertreten sind. In Abbildung 33 ist die Verteilung der Emotionen, die vom System in den Büchern gefunden wurde, abgebildet. Um diese Verteilung zu erstellen, wurden die Prozentzahlen pro Emotion über alle Bücher zusammengerechnet. Bei diesem Vergleich kommt es lediglich auf das Verhältnis an, weshalb diese Berechnung ausreicht. Es kann den beiden Diagrammen entnommen werden, dass bei den Büchern vor allem «joy», «anger» und «anticipation» überwiegend gefunden worden sind, was ungefähr den häufig vertretenen Emotionen im Trainingsset entspricht.

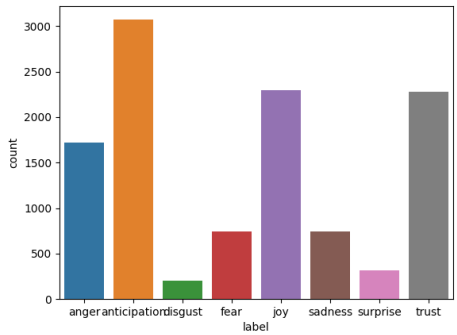

Abbildung 32 Verteilung der Emotionen im Datenset für Training und Validation.

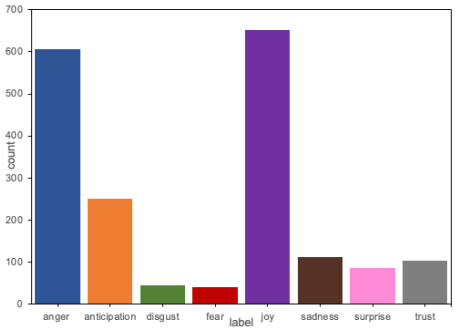

Abbildung 33 Aggregierte Prozentzahlen der Emotionen in den Büchern.

Der Versuch, die Emotionen und Genre zu vergleichen, stellt eine Grenze dieser Arbeit dar. Mit dem vorhandenen Datenset und dem trainierten Modell ist es folglich nicht möglich, aufgrund der Emotionen Bücher und Texte einem Genre zuzuordnen. Dafür wären mehrere Daten nötig und zudem müssten für eine Überprüfung die genauen Inhalte der Bücher bekannt sein,

so wie die Emotionen, die tatsächlich in den Büchern enthalten sind. Dann könnte man Vergleiche ziehen, wie gut das Modell die Emotionen im Buch erkennt und inwiefern diese einen Zusammenhang mit dem Genre haben.

5 Diskussion

Basierend auf der Evaluation werden nun die Herausforderungen bei der Emotionsanalyse durch ein Computersystem, welches auf Machine Learning basiert, genauer diskutiert. Einige der Schwierigkeiten existieren bereits bei der manuellen Annotation von Emotionen in einer schwächeren Form, wodurch die maschinelle Einordnung von Texten zusätzlich erschwert wird.

5.1 Herausforderungen der Emotionsanalyse

Bei der Emotionsanalyse stösst man auf verschiedene Hindernisse, die eine gute Performance eines Klassifikationssystems erschweren.
Nachfolgend werden nun die Herausforderungen beleuchtet, die im Rahmen dieser Arbeit aufgetreten sind. Einige der Schwierigkeiten sind vor allem bei der Kontrolle der Stichprobe des Silver Standards, die manuell durchgearbeitet wurde, aufgefallen. Aus dieser Stichprobe dienen einige Beispiele als Grundlage für die folgende Diskussion.

5.1.1 Fehlende Kategorien bei den Emotionen

Wie zu Beginn in Abschnitt 2.2. erläutert, gibt es verschiedene Emotionstheorien, die auf unterschiedlicher Auswahl der Emotionen basieren. Im Rahmen dieser Arbeit wurden ausschliesslich die acht Basisemotionen nach Plutchik verwendet.
Bei der Klassifikation ist aufgefallen, dass jedoch bei der Beschränkung auf die acht Basisemotionen «Angst», «Freude», «Erstaunen», «Wut», «Ekel», «Trauer», «Vertrauen», «Aufmerksamkeit» oder auf Englisch «fear», «joy», «surprise», «anger», «disgust», «sadness», «trust», «anticipation» die Schwierigkeit besteht, gewisse Sätze eindeutig einzuordnen.
Fehlende Emotionen sind beispielsweise solche wie Neid, Peinlichkeit oder Liebe. Es gibt Fälle, wo es beispielsweise um Liebe geht, der Satz aber trotzdem einer der acht Basisemotionen zugeordnet werden muss, da die Liebe beispielsweise in einer freudigen Situation auftritt oder die Liebe die handelnden Personen freudig stimmt. Darauf würde die Klassifikation mit der Emotion «Freude» naheliegen. Ähnliches gilt für Neid, da dies eine Emotion ist, die oft begleitet wird von anderen Emotionen, wie Wut oder Trauer. Ein Gefühl, welches aber eher für sich allein steht, ist die Peinlichkeit.
Ein Beispiel hierfür aus der Stichprobe des Silver Standards ist folgender Satz: « *Bo blushed rosy red, and the cowboy's face lost something of its somber hue.* »
Dies bedeutet nach Übersetzung der Verfasserin auf Deutsch: « *Bo errötete rosig, und das Gesicht des Cowboys verlor etwas von seinem düsteren Farbton.* »

Diesem Satz ist eindeutig zu entnehmen, dass die Person «Bo» peinlich berührt ist und deshalb rosig errötet. Es ist jedoch schwierig, eine passende Emotion in der Liste der acht Basisemotionen zu finden. Man könnte einerseits mit «Trauer» oder «Wut» argumentieren, da «Bo» vor der Errötung einen düsteren Farbton hatte, was darauf hindeuten könnte, dass die Person zuvor wütend oder bedrückt war. Das Erröten und Verlieren des düsteren Farbtones können aber auch bedeuten, dass der Cowboy nun glücklicher ist, als er es zuvor war. Demnach könnte man mit «Freude» beziehungsweise «joy» als passendes Label argumentieren.
Aus diesen Argumenten wird klar, dass es in einem solchen Fall, wo «Peinlichkeit» eher zutreffend ist, schwierig ist, den Satz mit den zur Verfügung stehenden Emotionen zu annotieren. Unter diesen Emotionen findet man, wie man im Beispiel sieht, kaum zu einer eindeutigen Entscheidung, da je nach Kontext und Blickwinkel Verschiedenes in Frage kommt.

5.1.2 Entscheidung für eine Emotion

Dies führt in der Diskussion zur nächsten Schwierigkeit: Die eindeutige Entscheidung für eine Emotion. Dies ist eine Herausforderung, da in Sätzen oftmals mehrere Emotionen vorkommen, beziehungsweise keine eindeutige Einordnung möglich ist, da zum einen der Kontext nicht bekannt ist. Zum anderen ist nicht immer klar, welche Person man im Text betrachtet beim Annotieren. Diese Schwierigkeit wird jedoch im Abschnitt 5.1.3 genauer diskutiert.

Ein Beispiel aus der Stichprobe zeigt das Problem des Annotierens mit fehlendem Kontext auf:
« *He talked of the men he'd shot, of the fights he'd had.* »
Gemäss dem lexikonbasierten Ansatz des Silver Standards ist dieser Satz mit «anger» (de: Wut) zu klassifizieren. Die Kontrolle der Stichprobe hat ergeben, dass dies nicht zwingend falsch ist, jedoch auch «sadness» (de: Trauer) in Frage kommen würde. Doch um eine genaue Aussage zu machen, müsste man mehr über die Person im Satz wissen. Ist er wütend auf die Personen, gegen die er gekämpft hat oder hasserfüllt, da es seine Gegner waren? Oder spricht er in Trauer darüber, dass er Männer erschiessen musste? Die Antworten auf solche Fragen fehlen in diesem Fall, da keine Informationen über die Person vorhanden sind.

Ein ähnliches Beispiel wäre folgender Satz: *« This moment made her forget all her troubles and all her anguish »*, was übersetzt bedeutet, dass sie all ihre Sorgen und Ängste vergessen konnte. Ohne Kontext ist es kaum möglich, diesen Satz einer Emotion zuzuordnen, da unbekannt ist, was dieser Moment (Im Satz: «this moment») genau für eine Situation darstellt. Ist sie voller Wut und im Begriff jemanden anzugreifen, dann würde wohl die Wut all ihre Ängste und

Sorgen vergessen lassen. Es ist aber auch möglich, dass sie fliehen muss und sie dabei all ihre Sorgen vergisst und sich Dinge traut, vor welchen sie unter normalen Umständen Angst hätte, sie diese aber durch das Adrenalin der Flucht völlig ausblendet. In diesem Fall wäre die passende Emotion «Angst» beziehungsweise «fear». Auch lexikonbasiert würde man den Satz tendenziell mit «Angst» oder «Wut» klassifizieren, da die Wörter «troubles» und «anguish» mit diesen Labels im NRC-Lexikon erfasst sind. «anguish» könnte zusätzlich im Zusammenhang von «Trauer» vorkommen.

Eine weitere Möglichkeit besteht darin, den Satz mit «Freude» (en: «joy») zu klassifizieren, da die Person im Text allenfalls einen freudigen, schönen Moment erlebt, den sie all ihre Sorgen und Ängste vergessen lassen. Beispielsweise ein Wiedersehen mit einer geliebten Person.
Würde man einen solchen Satz einer Gruppe Annotierenden vorlegen und verlangen, dass sie sich eindeutig für eine Emotion entscheiden müssten, würden wahrscheinlich diverse unterschiedliche Klassifikationen erfolgen. Dies sieht man auch in den Daten vom REMAN Korpus. Dort sind diverse Sätze mehrfach im Datenset erfasst, mit unterschiedlichen Emotionen. Nicht nur, weil tatsächlich mehrere Emotionen im Satz vorhanden sind, sondern auch, da sich die Meinungen und Entscheidungen der Annotierenden unterscheiden.

Die Einordnung gewisser Sätze scheint folglich abhängig von der Interpretation der annotierenden Person zu sein. Daraus wird klar, dass es für eine Maschine ebenso schwierig ist, klare Entscheidungen zu treffen, da durch Unsicherheiten die Daten, auf deren Grundlage das System trainiert wird, verrauscht sein können.

5.1.3 Abhängigkeit von Blickwinkel

Manchmal ist keine klare Entscheidung für eine Emotion möglich, da teilweise mehrere Akteure in einem Satz vorkommen und sich dann die Frage stellt, wessen Emotion annotiert werden sollte.
Ein Beispiel hierfür ist ein Satz, welcher schon in Abschnitt 4.2 analysiert wurde: « *Back to Kennington Road by omnibus, riding outside, her eyes and hair doing execution upon a young man in a very high collar, who was, she saw, terribly tempted to address her, but, happily for himself, could not pluck up courage*. »
Wie im erwähnten Abschnitt bereits erklärt wurde, kommt es hier auf die Sichtweise an. Aus Sicht der Frau könnte «joy» zutreffen, da sie froh ist, wenn der Mann sie nicht anspricht. Da der Mann sich jedoch fürchtet sie anzusprechen, wäre seine Emotion die Angst.

5.1.4 Verneinungen, Modalverben

Was mit den Akteuren zusammenhängt ist zusätzlich die Tatsache, ob eine Emotion wirklich erlebt wurde. Ein Gefühl kann auch gedacht werden oder angedroht werden. Zum Beispiel: *« Wenn du den Termin absagst, dann bin ich wütend. »* Zum jetzigen Zeitpunkt ist noch niemand wütend. Die Person droht lediglich an, wütend zu sein, falls das Gegenüber den Termin absagt. Ob diese überhaupt in Betracht zieht, den Termin abzusagen, sprich wie wahrscheinlich das Eintreten der Emotion ist, wird ebenfalls nicht klar aus diesem Satz. Folglich sollte man korrekterweise solche Sätze nicht mit dem Label «Wut» klassifizieren, da es noch nicht eingetreten ist.

Ähnliches gilt für Verneinungen oder Phrasen, aus welchen klar wird, dass das Gegenteil zur erwähnten Emotion gemeint ist. Ein treffender Teilsatz findet sich im REMAN Korpus: « *I would be the very last to complain of it* », was nach der Verfasserin übersetzt bedeutet: *« Ich wäre der Allerletzte, der sich darüber beschweren würde ».*

Das Wort «complain» würde man korrekterweise mit Trauer oder Wut annotieren, wobei im REMAN Korpus in diesem Beispiel beide Annotationen vorhanden sind.

Aber der Teil «I would be the very last» impliziert, dass die Person eben nicht Trauer oder Wut verspürt, da er oder sie sich in diesem Kontext nicht beschweren würde. Daraus würde als zusammenfassende Emotion eher «Freude» resultieren.

Direkte Hinweise auf gegenteilige oder nicht erlebte Emotionen sind eindeutige Verneinungen wie unter anderem «nicht», «noch nicht» oder «weder…noch».

Was beim obigen Beispiel noch angefügt werden muss, ist, dass man aus diesem Kontext nicht genau weiss, ob die Freude schon eingetreten ist, da ein Modalverb «would» darin vorkommt. Dieses deutet darauf hin, dass die Person sich freuen würde, es aktuell aber noch nicht tut.

Dies ist gleichzeitig ein Beispiel dafür, dass nicht nur bei Verneinungen, sondern auch bei Modalverben bei einer Annotation Vorsicht geboten ist. Modalverben wie «können», «würden», «sollten», Satzstrukturen im Konjunktiv oder Futur deuten darauf hin, dass die im Text erwähnten Gefühle noch nicht erlebt wurden. Aus diesem Grund wurden bei der Erstellung des Silver Standards Sätze mit Modalverben und Verneinungen entfernt.

5.1.5 Ironie und Sarkasmus

Was an die Schwierigkeit der nicht tatsächlich erlebten Gefühle anschliesst, sind Ironie und Sarkasmus als weitere Herausforderungen bei der Emotionsanalyse. Ironie und Sarkasmus sind Stilmittel, welche die Erwartungen von einem Lesenden nicht erfüllen, wie Klinger in seiner Arbeit «Strukturierte Modellierung von Affekt in Text» schreibt (Klinger, 2020, S. 14). Ironie,

Sarkasmus, aber auch Satire sind oftmals besonders emotionsgeladen, werden dann aber humoristisch wieder aufgelöst (Klinger, 2020, S. 14). Eine differenzierte, separate Behandlung von Ironie und Sarkasmus kann folglich die Sentiment Analyse verbessern (Klinger, 2020, S. 14).

5.1.6 Veränderung der Sprache über die Zeit

Wie Schmidt, Burghardt und Wolff in ihrer Arbeit zu den Herausforderungen für Sentiment Analysis bei literarischen Texten erklären, ist eine weitere Schwierigkeit die Veränderung der Sprache über die Zeit. Ältere Texte basieren auf einem Vokabular, welches sich über die Zeit hinweg verändert hat und nicht mehr der heutigen Standardsprache entspricht (Schmidt, Burghardt, & Wolff, Herausforderungen für Sentiment Analysis-Verfahren bei literarischen Texten, 2018, S. 5). Das veraltete Vokabular ist vor allem bei einem lexikonbasierten Ansatz problematisch, wenn Sentiment-tragende Wörter nicht mehr erkannt werden, da sie sich in ihrer Rechtschreibung über die Zeit hinweg verändert haben (Schmidt, Burghardt, & Wolff, Herausforderungen für Sentiment Analysis-Verfahren bei literarischen Texten, 2018, S. 5). Ein weiteres Problem sind Wörter, die in diesem Sinne heutzutage nicht mehr in dieser Form existieren, die jedoch damals beispielsweise als eine Beleidigung galten und deshalb von einem System mit einer negativen Emotion annotiert werden sollten, wie beispielsweise «Pöbel» oder «Schlingel» (Schmidt, Burghardt, & Wolff, Herausforderungen für Sentiment Analysis-Verfahren bei literarischen Texten, 2018, S. 5). In ihrer Arbeit haben Schmidt, Burghardt und Wolff erklärt, dass sich bei einigen Wörtern über die Zeit die sogenannte Sentiment-Bedeutung verändert, was bedeutet, dass ein Wort, wie beispielsweise «Freier», heute negativer wahrgenommen wird, als es damals der Fall war (Schmidt, Burghardt, & Wolff, Herausforderungen für Sentiment Analysis-Verfahren bei literarischen Texten, 2018, S. 5). Eine Lösung, die von ihnen vorgeschlagen wurde, ist die Lemmatisierung bei der Verarbeitung der Texte oder auf das Anwendungsgebiet abgestimmte Sentiment Lexika (Schmidt, Burghardt, & Wolff, Herausforderungen für Sentiment Analysis-Verfahren bei literarischen Texten, 2018, S. 5-6).

5.2 Fazit

Zusammenfassend kann gesagt werden, dass das Sprachmodell BERT für die Emotionsanalyse verwendet werden kann und wie im Abschnitt 4.2 anhand der Arbeit von Kim und Klinger (2018) erklärt wurde, bessere Resultate liefern kann als andere Modelle, wie beispielsweise biLSTMs oder lexikonbasierte Verfahren. Der Wert des F1-Scores von 84 Prozent, welcher beim besten Trainingsdurchgang von BERT erreicht wurde, ist zufriedenstellend, wenn man diesen mit den Baseline Modellen der Universität Stuttgart vergleicht. Grundsätzlich gibt es

viele Schwierigkeiten und Herausforderungen, denen man sich bei der Emotionsanalyse bewusst sein muss und welche bei der maschinellen, aber auch bei der manuellen, menschlichen Klassifizierung auftreten. Dass sich die Werte mit wachsendem Datenset verbessert haben, bestätigt die Annahme, dass für Emotionsanalyse mittels maschinellen Lernens grosse Mengen an annotierten Daten benötigt werden. Im Vergleich zu den vorhandenen Daten dank Social Media, den Beiträgen auf Twitter und den Kommentaren auf Facebook, gibt es im Gebiet der Prosa Literatur noch wenig annotierte Daten, wie die Recherche ergeben hat und im Kapitel der Diskussion zusammengefasst wurde.
Da beispielsweise die eindeutige Klassifizierung eines Satzes nach einer Emotion auch für Menschen eine grosse Herausforderung ist, stellt es eine grosse Schwierigkeit dar, das System besser zu trainieren, da oftmals die Daten verrauscht sind und je nach Meinung und Auffassung der annotierenden Person ein Satz mit mehreren Emotionen annotiert sein kann. Dadurch ist für die Maschine eine exakte Klassifizierung eine grosse Herausforderung.

Es hat sich gezeigt, dass die Klassifizierung für Menschen leichter wird, sobald mehr Kontext des zu klassifizierenden Satzes zur Verfügung steht. Am Einfachsten wird es dann, wenn ein Mensch ein Buch in voller Länge lesen kann und laufend die Emotionen beurteilt. Die Performance einer Maschine könnte mit einem ähnlichen Vorgehen verbessert werden, wenn die Maschine die Informationen der einzelnen Personen im Buch oder im Text kennt und sammelt und laufend die Emotionen mit den Charakterzügen und den Erlebnissen der Protagonisten abgleicht und aktualisiert. Daraus folgt, dass ein System nicht nur mit den Sätzen und den dazugehörigen Emotionen trainiert werden sollte, sondern auch mit den dazugehörigen Zusammenhängen, wie der Person, welche die Emotion erlebt oder den Ursachen. Wobei bei den Ursachen betrachtet werden müsste, wie der Zustand der betroffenen Person zuvor war und wie sich diese Person während der Geschichte verändert und entwickelt hat und welche Beziehungen diese zu anderen Charakteren der Geschichte hat. Gibt man dem System ein ganzes Buch, würde dies bedeuten, dass es dieses ähnlich durcharbeiten muss, wie ein Mensch und dieses in einem ähnlichen Sinne verstehen muss.

5.3 Ausblick

Im letzten Abschnitt der Evaluation wurde gezeigt, inwiefern ein weiterer Schritt möglich wäre. Hierfür wird ein Datenset benötigt, in welchem Bücher nach Genre und ihren enthaltenen Emotionen annotiert sind. Als Leserin oder Leser eines Buches ist es nach Beenden grundsätzlich möglich, einen Schluss zu ziehen und zu entscheiden, welche Emotion überwiegt. So würde beispielsweise bei einer tragischen Liebesgeschichte viel Traurigkeit enthalten sein. Dieses Datenset könnte man schliesslich mit dem Modell klassifizieren und dann vergleichen, an welchen Stellen der Gold Standard mit der Annotation des Systems übereinstimmt und anhand dessen evaluieren, wie gut das Modell tatsächlich funktioniert.
Schliesslich könnte ein optimiertes Modell dazu verwendet werden, Bücher automatisch einem Genre zuzuordnen.

Ein weiterer Anwendungsbereich wäre im Feld des Computer-assisted Language Learning möglich. Zum Spracherwerb und zur Festigung der Sprache werden in Schulen Bücher gelesen und analysiert und Textstellen interpretiert. Ein Computersystem, welches befähigt ist, Emotionen in Texten zu erkennen oder allenfalls sogar die Zusammenhänge und Beziehungen der Personen im Buch, sowie die Charakterzüge der Protagonisten, könnte eine unterstützende Funktion im Bereich des Spracherwerbs haben. So könnte das System den Lernenden in der Fremdsprache dem Niveau angepasste Textstellen oder sogar Bücher zum Lesen vorlegen. Mit Fragen zur Textanalyse kann daraufhin abgefragt werden, wie gut die Lernenden den Text verstanden haben, wobei die Antworten direkt vom System überprüft werden können. Somit könnten Textanalysen mithilfe eines Computersystems durchgeführt werden.

Literatur- und Quellenverzeichnis

Ahmad, K., & Devitt, A. (2013). *Is there a language of sentiment? An analysis of lexical resources for sentiment analysis. Language Resources and Evaluation.* (Bd. 47). 47:475–511: Springer.

Alm, C. O., Roth, D., & Sproat, R. (2005). Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing* (S. 579-586). Vancouver, British Columbia, Canada: Association for Computational Linguistics.

Bostan, L.-A.-M., & Klinger, R. (2018). An Analysis of Annotated Corpora for Emotion Classification in Text. In *Proceedings of the 27th International Conference on Computational Linguistics* (S. 2104-2119). Santa Fe, New Mexico, USA: Association for Computational Lingustics.

Chakraborty, A. (November 2020). *Kaggle*, 1. Abgerufen am 10. März 2021 von https://www.kaggle.com/anik424/sentiment-analysis-multiclassification-using-bert#Building-Model-using-Transformer.

Deng, L., & Yu, D. (Mai 2014). *Deep Learning: Methods and Applications.* Abgerufen am 05. Mai 2021 von https://www.microsoft.com/en-us/research/publication/deep-learning-methods-and-applications/.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* Abgerufen am März 2021 von https://arxiv.org/abs/1810.04805.

Gabler Wirtschaftslexikon. (2021). *Gabler Wirtschaftslexikon Online*. Abgerufen am 2. März 2021 von https://wirtschaftslexikon.gabler.de/definition/emotion-35195.

github.com/amaiya/ktrain. (2021). *github.com/amaiya/ktrain.* Abgerufen am 29. März 2021 von Github ktrain: https://github.com/amaiya/ktrain.

Gonzáles-Carvajal, S., & Garrido-Merchán, E. C. (2020). *Comparing BERT against traditional machine learning text classification.* Madrid, Spanien: CoRR: abs/2005.13012.

Kim, E., & Klinger, R. (2018). *A Survey on Sentiment and Emotion Analysis for Computational Literary Studies.* Stuttgart: University of Stuttgart.

Kim, E., & Klinger, R. (2018). Who Feels What and Why? Annotation of a Literature Corpus with Semantic Roles of Emotions. In *Proceedings of the 27th International Conference on Computational Linguistics* (S. 1245-1359). Santa Fe, New Mexico, USA: Association for Computational Lingustics.

Klinger, R. (2020). *Strukturierte Modellierung von Affekt in Text.* Universität Stuttgart.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. In *Biometrics 33(1)* (S. 159–174). doi:10.2307/2529310.

Liu, B. (2015). *In Sentiment Analysis: Mining Opinions, Sentiments, and Emotions.* Cambridge: Cambridge University Press.

Maiya, A. S. (2020). *ktrain: A Low-Code Library for Augmented Machine Learning.* Abgerufen am 22. März 2021 von https://arxiv.org/abs/2004.10703.

Mohammad, S. M., & Turney, P. D. (2013). *Crowdsourcing a Word-Emotion Association Lexicon.* Ottawa, Ontario, Canada: Institute for Information Technology, National Research Council Canada.

Project Gutenberg. (kein Datum). *Project Gutenberg*. Abgerufen am 02. Mai 2021 von https://www.gutenberg.org/.

Schmidt, T., Burghardt, M., & Dennerlein, K. (2018). "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, T., Burghardt, M., & Wolff, C. (2018). Herausforderungen für Sentiment Analysis-Verfahren bei literarischen Texten. In M. Burghardt, & C. Müller-Birn, *INF-DH-2018.* Bonn: Gesellschaft für Informatik e.V.

Traue, H. C., & Kessler, H. (2003). Psychologische Emotionskonzepte. In W. H. Stephan A, *Natur und Theorie der Emotion* (S. 20-33). Paderborn: Mentisverlag.

Wikipedia. (21. Oktober 2012). *Plutchik, Rad der Emotionen.* Abgerufen am 2. März 2021 von https://de.wikipedia.org/wiki/Robert_Plutchik#/media/Datei:Plutchik-wheel_de.svg.

Die Quellen der Abbildungen sind jeweils in der Bildunterschrift erwähnt. Wenn keine Quelle angegeben ist, wurden die Diagramme und Tabellen eigenhändig im Rahmen dieser Arbeit erstellt. Das Dokument, von welchem Abbildung 3 entnommen wurde (Guidelines REMAN Korpus) ist dem Anhang beigefügt.

Anhang

Im bei der Abgabe beigelegten File befinden sich folgende Ordner:

- /_PYHTON/
- /_EVALUATION
- /_PREDICTOR/
- Topic_Paper_Final_Thesis.pdf
- selbststaendigkeitserklaerung.pdf
- Bachelorarbeit_kthommen_16580011.pdf

Daten und Skripte für Preprocessing und Generierung des Silver Filter

/_PYTHON/

- get_information.py ..... Diagramme mit Emotionsverteilungen erstellen
- /PREPROCESSING/ ..... Vorverarbeitung des REMAN Korpus
  - /input/reman-version1.0.xml Original REMAN-XML Datei
  - /input/reman Gesamter Download Ordner des REMAN Korpus. Inkl. Readme und Guidelines (Abb. 3 stammt aus den Annotationsguidelines)
  - preprocess_phrase.py Python Skript für Vorverarbeitung des REMAN Korpus
- /SILVERFILTER/ ..... alle Dateien um den Silver Standard zu generieren
  - /input/
    - /gutenberg/ Bücher als Input zur Generierung des Silver Standards
    - /NRC-Emotion-Intensity-Lexicon-v1.txt
  - /output/
    - /silver_data.csv/ Vollständiger Silver Standard
    - /silver_gold_merge.csv/ Silver Standard + Gold Standard (komplett)
    - /silver_selection.csv/ Selektion aus dem Silver Standard (max. 600 Sätze pro Emotion)
    - /silver_stichprobe.csv/ Stichprobe
  - silver_filter.py Generierung des Silver Standards

Dateien zur Evaluation

/_EVALUATION/

- predictor_output.txt ..... Exakter Output des Emotion Predictors beim Input diverser Bücher (Kapitel 4.5)
- stichprobe_auswertung.xlsx ..... Excel Datei mit den Resultaten der Stichprobe inkl. Auswertung und Berechnungen

BERT Modell

/_PREDICTOR/

- /BOOKS/ ..... Input für den Emotion Predictor (Kapitel 4.5)
- /TRAINING_DATA/ ..... Trainingsdaten für BERT
  - BERT_gold_sentences.csv Nur Gold Standard Sätze
  - gold_training_data.csv Gold Standard Phrasen und Sätze
  - silver_clean_gold_merge.csv Finales Set für Training und Validation von BERT
- emotion_predictor.ipynb ..... Google Colab Emotion Predictor (Kapitel 4.5)
- saved_model ..... Ordner mit Modell (gleicher Inhalt wie zip)
- saved_model.zip ..... Gespeichertes, trainiertes Modell von BERT, welches in emotion predictor wieder geladen werden kann
- V3_BERT.ipynb ..... Finales BERT (84 % F1 Score)

Readme REMAN Korpus

(Gesamter Ordner mit XML-Datei und Readme – Download von https://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/reman /, Zugriffsdatum: 1. März 2021)

* This folder contains the following files:
 - reman-version1.0.xml: the corpus itself
 - reman-schema.xsd: XSD schema
 - reman.dtd: DTD definition
 - guidelines.pdf: official annotation guidelines
 - license.txt: full text of license information
* The REMAN (Relational EMotion ANnotation) corpus provides the annotation of emotions with the roles of experiencer, cause, and target in 1720 sentence triples subsampled from the Project Gutenberg (www.gutenberg.org).

XML file structure:
1. "adjudicated" node:

Includes annotations accepted by the curator. The node includes two layers: "spans" and "relations". "spans" include information about the annotations of phrases. "relations" include information about the annotation of relations.
2) "other" node is structured identically to "adjudicated" node but includes annotations that were rejected by the curator.

* Please cite the corpus as:
@inproceedings{Kim2018,
  author = {Kim, Evgeny and Klinger, Roman},
  title = {Who Feels What and Why? Annotation of a Literature Corpus with Semantic Roles of Emotions},
  booktitle = {Proceedings of COLING 2018, the 27th International
      Conference on Computational Linguistics},
  month = {August},
  year = {2018},
  address = {Santa Fe, USA},
}

* The corpus can be found at: http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/reman.html
* The REMAN corpus is licensed under Creative Commons 4.0 license (https://creativecommons.org/licenses/by/4.0/). See file license.txt for more information.
* NOTE: Our license does not include Project Gutenberg's license.
* The current version of the corpus is 1.0
* Contact information: Evgeny Kim evgeny.kim@ims.uni-stuttgart.de

Verwendete Bücher aus dem Gutenberg Projekt

Die gelisteten Bücher sind alle auf der Webseite des Gutenberg Projekt zu finden und die heruntergeladenen Versionen befinden sich im abgegebenen Zip-File.

Gutenberg Projekt: https://www.gutenberg.org/

Liste der Bücher für den Silver Standard

Titel	Autor
A Matter of Proportion	Anne Walker
A Strange Disappearance	Anna Katharine Green
A Study in Scarlet	Conan Doyle
Adam Bede	George Eliot
Adventure of Sally	P. G. Wodehouse
Alice’s Adventures in Wonderland	Lewis Carroll
Battles with the Sea	R. M. Ballantyne
Black Amazon of Mars	Leigh Brackett
Black Ivory	R. M. Ballantyne
Bones in London	Edgar Wallace
Book of Wonder	Edward J. M. D. Plunket, Lord Dunsay
Captain Blood	Rafael Sabatini
Carmilla	J. Sheridan LeFanu

Charlie to the Rescue	R. M. Ballantyne
Diary of a Nobody	George and Weedon Grossmith
Dickory Dock	L. T. Meade
Frankenstein	Mary Wollstonecraft (Godwin) Shelley
Ghost Stories of an Antiquary	M. R. James
Gread Expectations	Charles Dickens
Great Gatsby	F. Scott Fitzgerald
Greenmantle	John Buchan
Helpfully Yours	Evelyn E. Smith
Huckleberry Finn	Mark Twain (Samuel Clemens)
In the Fog	Richard Harding Davis
Irish Fairy Tales	James Stephens
Key Out of Time	Andre Alice Norton
Kidnapped	Robert Louis Stevenson
Love Among Chickens	P. G. Wodehouse
The Mabinogion	Lady Charlotte Guest
Magic of Oz	L. Frank Baum
Men Of Iron	Howard Pyle
My Fair Planet	Evelyn E. Smith
Night Land	William Hope Hodgson
No Hero	E. W. Hornung
Only a Girl's Love	Charles Garvice
Other People's Money	Emile Gaboriau
Plague Ship	Andre Norton
Prester John	John Buchan
Pride and Prejudice	Jane Austen
Robin Hood	Howard Pyle
Sherlock Holmes	Arthur Conan Doyle
Star Born	Andre Norton
Star of India	Alice Perrin
The Strange Case of Dr. Jekyll and Mr. Hyde	Robert Louis Stevenson
Tarzan of the Apes	Edgar Rice Burroughs
The Big Otter	R. M. Ballantyne
The Black Arrow	Robert Louis Stevenson
The Blue Tower	Evelyn E. Smith
The Cash Boy	Horatio Alger Jr.
The Castaways	Captain Mayne Reid
The Cave in the Mountain	Lieut. R. H. Jayne
The Cliff Climbers	Captain Mayne Reid
The Crystal Stopper	Maurice LeBlanc
The Devil Doctor	Sax Rohmer
The Doorway	Evelyn E. Smith
The Gloved Hand	Burton E. Stevenson
The Holladay Case	Burton E. Stevenson
The Hound of the Baskervilles	Arthur Conan Doyle
The Last Supper	T. D. Hamm
The Leavenworth Case	Anna Katherine
The Lone Star Ranger	Zane Grey
The Man in The Brown Suit	Agatha Christie
The Man of the Forest	Zane Grey

The Middle of Things	J. S. Fletcher
The Moon Rock	Arthur J. Rees
The Moonstone	Wilkie Collins
The Murder on the Links	Agatha Christie
The Paradise Mystery	J. S. Fletcher
The Red House Mystery	A. A. Milne
The Riddle of the Forzen Flame	Mary E. Hanshew, Thomas W. Hanshew
The Secret Adversary	Agatha Christie
The Sign of the Four	Arthur Conan Doyle
The Time Traders	Andre Norton
The Yellow Claw	Sax Rohmer
Thirty-Nine Steps	John Buchan
To The Last Man	Zane Grey
Vanity Fair	William Makepeace Thackeray
The Well At the World's End	William Morris
Wildfire	Zane Grey
Wit and Humor of America, Volume II.	Various

Liste der Bücher um den Emotion Predictor zu testen:

Titel	**Autor**
Brood Of The Witch Queen	Sax Rohmer
Dead Man Tell No Tales	E. W. Hornung
Murder In The Gunroom	H. Beam Piper
A Dreamers Tales	Lord Dunsay
Gulliver Of Mars	Edwin L. Arnold
The Gods Of Pegana	Lord Dunsay
Draculas Guest	Bram Stoker
Nightmare Abbey	Thomas Love Peacock
The Turn Of The Screw	Henry James
Wuthering Heights	Emily Brontë
Book Of Nonsense	Edward Lear
Comic History of USA	Bill Nye
Man On The Bummel	Jerome K. Jerome
The Man With Two Left Feet	P. G. Wodehouse
Kate Vernon	Mrs. Alexander
Wastralls	C. A. Dawson Scott
The Hanging Stranger	Philip K. Dick
The Hour Of The Dragon	Robert E. Howard
The Lost World	Arthur Conan Doyle