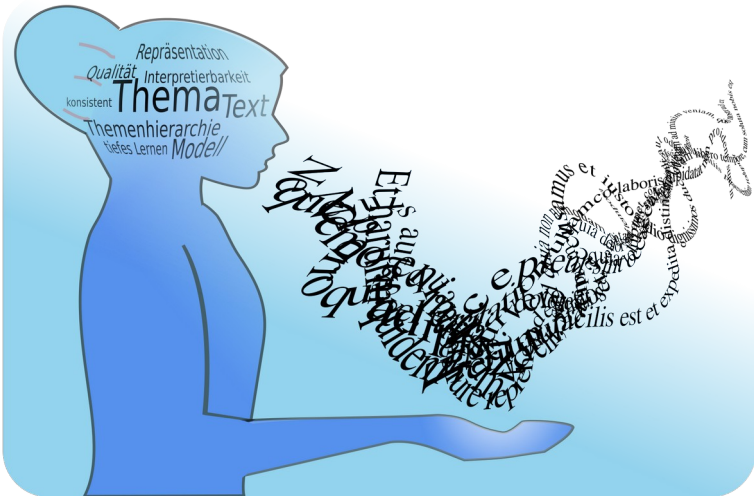
RPTU

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Veröffentlichung der Ergebnisse von Forschungsvorhaben im BMBF-Programm
“SeDis”
Semantic Disentanglement:
Unterscheidung von Stil und Thema in Textdaten



Förderkennzeichen:
01IS20048

Zuwendungsempfänger:
Rheinlandpfälzische Technische Universität
Kaiserslautern-Landau, Campus Kaiserslautern

Projektleiterin und Autorin des Berichts:
Jun.-Prof. Dr. Sophie Fellenz

Projektlaufzeit:
01.06.2020 – 31.05.2024

Das diesem Bericht zugrunde liegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IS20048 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei Sophie Fellenz.

Kurzbericht

Das Projekt SeDis hat zum Ziel, den von großen Sprachmodellen generierten Text besser steuerbar zu machen. Der Titel des Projektes, *Semantic Disentanglement*, beschreibt das Problem, wie unterschiedliche Faktoren, die in Daten enthalten sind, auseinandergehalten oder *disentangled* werden können. Im Projekt SeDis geht es um Disentanglement in Bezug auf Textstil und Thema in Texten. Die gelernten Repräsentationen für Stil und Thema sollen *disentangled* sein, sodass eine separate Kontrolle möglich wird, also eine Änderung des Stils, ohne, dass der Inhalt des Textes sich ändert und umgekehrt. Beispielsweise sollte es möglich werden, einen Harry Potter Roman in den Stil von Shakespeare zu konvertieren. Es sollte ein langer Artikel in Kurzform zur Verfügung gestellt oder für soziale Medien zusammengefasst werden können. Ein formal wissenschaftlicher Artikel sollte in vereinfachter Sprache wiedergegeben oder für eine andere Zielgruppe aufbereitet werden können.

Das Projekt wurde von 2020 bis 2024 von Sophie Fellenz geleitet, die vorher als Postdoktorandin beschäftigt war. Es ermöglichte ihr, erstmals eine eigene Arbeitsgruppe aufzubauen. Die Stellen dieser Arbeitsgruppe wurden erfolgreich überwiegend mit Frauen besetzt. Sophie Fellenz erhielt kurz nach Beginn des Projektes einen Ruf auf eine Juniorprofessur für Maschinelles Lernen ohne Tenure Track an der damaligen TU Kaiserslautern (heute RPTU Kaiserslautern-Landau), wo ihr eigenständige Forschungs- und Lehraufgaben übertragen wurden. Dadurch und durch die Förderung des BMBF konnte ihre Arbeitsgruppe erfolgreich eine Vorbildwirkung im Fachbereich Informatik entfalten und mehrere Nachwuchswissenschaftlerinnen und auch Nachwuchswissenschaftler im Maschinellen Lernen ausbilden. Sophie Fellenz konnte im Laufe des Projektes die Zwischenevaluation der Juniorprofessur bestehen und damit die generelle Berufbarkeit auf eine reguläre Professur nachweisen.

Im Laufe ihrer Promotion und Zeit als Postdoktorandin arbeitete Sophie Fellenz an Themenmodellen. Dies sind Modelle, mit denen die thematische Struktur in großen Datensätzen mit Textdokumenten untersucht werden kann. Zur Zeit der Antragstellung war das Sprachmodell GPT-2 soeben erschienen und konnte erstmals eine überzeugende Leistung in der Generierung von längeren Texten demonstrieren. Jedoch waren die generierten Text noch schwer steuerbar. Daraus ergaben sich die Startbedingungen für das Projekt. Vortrainierte große Sprachmodelle sollten mit Hilfe von Themenmodellen kontrollierbar gemacht werden. Die Expertise der Projektleiterin im Bereich der Themenmodelle und Textdaten in Verbindung mit neuesten Forschungsergebnissen im Bereich der Sprachmodelle, stellten eine einzigartige Möglichkeit dar, beide Themenbereiche gewinnbringend miteinander zu verknüpfen. Das Projekt sollte Forschung zu Sprachmodellen und Themenmodellen zusammenführen, um ein kombiniertes Modell zu entwickeln, das sowohl den Inhalt als auch den Textstil repräsentieren kann. Hierzu sollten tiefe neuronale Netze zum Einsatz kommen, wobei erforscht werden sollte, wie diese mit den Besonderheiten von Textdaten umgehen können. Während es erfolgreiche Methoden für Bilddaten bereits gab, war unklar, wie ähnliche Ergebnisse für Texte erzielt werden könnten.

Die konkreten Forschungsziele des Projektes waren, Modelle zu entwickeln die zwischen Schreibstil und Inhalt differenzieren und generativ Texte nach bestimmten Vorgaben produzieren können, die eine Hierarchie von Themen und Konzepten anhand großer Datensätze lernen und

die mit externem Wissen initialisiert und erweitert werden können. Umgekehrt sollte aus Texten auch Wissen extrahiert werden können und gelerntes Wissen in neuen Modellen und anderen Anwendungskontexten wiederverwendet werden können, um das Training zu beschleunigen und die Ergebnisse zu verbessern. Weiterhin sollten Trainingsverfahren untersucht werden, mit denen derartige Sprachmodelle und Modelle für Repräsentationen von Text effizient gelernt werden können. Die zentrale Fragestellung war herauszufinden, wie anhand spezifischer Vorgaben neuer Text generiert werden kann. Die Generierung von qualitativem und den Vorgaben entsprechendem Text ist für alle Branchen, die an und mit Texten arbeiten, von zentraler Wichtigkeit.

Das Vorhaben wurde in fünf Arbeitspakete aufgeteilt, die sich mit folgenden Fragestellungen beschäftigten. Die Hauptfragestellung war die Textgenerierung unter Beeinflussung von Textstil und Thema des Textes. Zur Beeinflussung des Stils sollten mit Hilfe von neuronalen Netzwerken Repräsentationen gelernt werden, die den jeweiligen Stil eines Textes repräsentieren. Zur Beeinflussung des Themas sollten neuartige Themenmodelle entwickelt werden. Die Integration und Extraktion von bestehendem Wissen in Texte sollte verbessert werden. Beispielsweise sollte die Übereinstimmung mit bekannten Fakten sichergestellt werden. Der Transfer zwischen verschiedenen Stilen und Themen sollte gezielt untersucht werden. Zuletzt sollten auch Trainingsverfahren für das Training und Finetuning von Sprach- und Themenmodellen sowie der benötigten Repräsentation untersucht werden.

Durch die schnelle Weiterentwicklung im Bereich der großen Sprachmodelle und der Ende 2022 erschienenen ersten Version von ChatGPT wurden einige Projektziele bereits frühzeitig erreicht. Beispielsweise ist das eingangs erwähnte Szenario, einen Text zusammenzufassen oder in einfache Sprache übertragen zu lassen, heute kein Problem mehr. Auch die Einbindung von externem Wissen oder Tools ist mit Hilfe von Techniken wie Retrieval Augmented Generation (RAG) bereits ein gut untersuchtes Problem. Die zu Beginn des Projekts breite thematische Aufstellung wurde daher im Laufe des Projekt stärker fokussiert, um Fortschritte in einzelnen Themenbereichen zu erzielen, die mit den übergeordneten Projektzielen in Zusammenhang stehen. Insbesondere wurde ein Fokus auf das Reinforcement Learning gelegt, ein wichtiges Trainingsverfahren, für das wir neue Methoden für text-basierte Entscheidungsfindung entwickelt haben. Weiterhin haben wir dynamische Themenmodelle entwickelt und evaluiert, sowie neue Methoden für die Evaluation von Textstil-Transfer vorgeschlagen.

Die in diesem Projekt entwickelten Methoden zur Evaluation von Stiltransfer und Dynamischen Themenmodellen können in der Zukunft verwendet werden, um neu entwickelte Modelle für den Textstiltransfer besser zu testen und zu vergleichen. Die entwickelten Reinforcement Learning Methoden haben zum Ziel, die allgemeine textbasierte Entscheidungsfindung mit Hilfe von neuronalen Netzen zu verbessern. Erkenntnisse im Bereich des Disentanglement verschiedener Aspekte von Textdaten haben zu verbesserten Themenmodellen geführt und werden in Zukunft auch für die Weiterentwicklung von Modellen zum Textstiltransfer und der kontrollierten Textgenerierung hilfreich sein.

Außerhalb des Anwendungsbereichs von Texten werden die gewonnenen Erkenntnisse und entwickelten Methoden auch in anderen Bereichen einen Einfluss haben. Beispielsweise spielt das Disentanglement von verschiedenen Aspekten im Bereich physikalischer Modelle eine große Rolle und die Generierung von synthetischen Daten nach speziellen Kriterien ist auch in anderen technischen Bereichen von großer Wichtigkeit.

Die Forschungstätigkeit von SeDis wurde über Konferenz- und Workshopbeiträge zum Thema Maschinelles Lernen verbreitet. Auch in der Lehre flossen die Arbeiten ein. Zum Beispiel wurde eine neue Vorlesung zum Thema generativer Sprachmodelle entwickelt und durchgeführt, mehrere Abschlussarbeiten zum Thema wurden betreut und studentische Projekte im Zusammenhang mit dem Projekt wurden durchgeführt.

Inhaltsverzeichnis des Schlussberichts

I Kurze Darstellung **2**

I.1 Ausgangsbasis . . . 2
I.2 Voraussetzungen . . . 2
I.3 Planung und Ablauf des Vorhabens . . . 3
I.3.1 AP 1: Disentanglement von Stil und Thema . . . 3
I.3.2 AP 2: Hierarchisches Topic Model mit Disentanglement . . . 3
I.3.3 AP 3: Hintergrundwissen . . . 3
I.3.4 AP 4: Transfer Learning/Domain Adaptation . . . 4
I.3.5 AP 5: Trainingsverfahren . . . 4
I.4 Wissenschaftlicher und Technischer Stand . . . 4
I.4.1 Themenmodelle . . . 4
I.4.2 Hierarchische Generative Modelle . . . 5
I.4.3 Disentanglement von Stil und Thema . . . 5
I.4.4 Transfer Learning und Domain Adaptation . . . 5
I.4.5 Wissensextraktion . . . 6

II Eingehende Darstellung **6**

II.1 Durchgeführte Projektarbeiten . . . 6
II.1.1 AP 1: Disentanglement von Stil und Thema . . . 6
II.1.2 AP 2: Hierarchisches Topic Model mit Disentanglement . . . 7
II.1.3 AP 3: Hintergrundwissen . . . 11
II.1.4 AP 4: Transfer Learning/Domain Adaptation . . . 13
II.1.5 AP 5: Trainingsverfahren . . . 14
II.2 Wichtige Positionen des zahlenmäßigen Nachweises . . . 17
II.3 Notwendigkeit und Angemessenheit der geleisteten Projektarbeiten . . . 17
II.4 Verwertbarkeit der Ergebnisse . . . 18
II.5 Fortschritt anderer Stellen im Bereich des Projekts . . . 18
II.6 Erfolgte oder geplante Veröffentlichungen . . . 18

I. Kurze Darstellung

I.1. Ausgangsbasis

Der Titel des Projektes, *Semantic Disentanglement*, beschreibt das Problem, wie unterschiedliche Faktoren, die in Daten enthalten sind, auseinandergehalten oder *disentangled* werden können. Immer dann, wenn abstrakte Repräsentationen gelernt werden, ist es wichtig, dass diese die tatsächlich in den Daten vorhandenen Faktoren gut abbilden und interpretierbare, eventuell auch für Menschen verständliche Dimensionen darstellen. Separate Kategorien sollten keine vermischten Repräsentationen haben und es sollte nicht mehrere redundante Repräsentationen für die gleiche Kategorie geben. Im Projekt SeDis geht es um Disentanglement in Bezug auf Textstil und Thema in Texten. Die gelernten Repräsentationen für Stil und Thema sollen *disentangled* sein, sodass eine separate Kontrolle möglich wird, also eine Änderung des Stils, ohne, dass der Inhalt des Textes sich ändert und umgekehrt.

Die konkreten Forschungsziele des Projektes waren wie folgt. Es sollten Modelle entwickelt werden,

- die zwischen Schreibstil und Inhalt differenzieren und generativ Texte nach bestimmten Vorgaben produzieren können,
- die eine Hierarchie von Themen und Konzepten anhand großer Datensätze lernen,
- die mit externem Wissen initialisiert und erweitert werden können und aus denen Wissen umgekehrt auch wieder extrahiert werden kann,
- deren gelerntes Wissen in neuen Modellen und anderen Anwendungskontexten wiederverwendet werden kann, um das Training zu beschleunigen und die Ergebnisse zu verbessern
- und die mit Hilfe von neuartigen Trainingsverfahren effizient trainiert werden können.

Die zentrale Ausgangsfrage war also: Wie kann strukturiertes Wissen aus unstrukturierten Textdaten gewonnen werden? Diese Frage kann weitergeführt werden zu: Wie kann dieses Wissen genutzt werden, um anhand spezifischer Vorgaben neuen Text zu generieren? Die Generierung von qualitativem Text ist für alle Branchen, die an und mit Texten arbeiten, von zentraler Wichtigkeit.

I.2. Voraussetzungen

Die Projektleiterin Sophie Fellenz war bis zum Förderbeginn als Postdoktorandin an der Johannes Gutenberg Universität in Mainz beschäftigt. Das Projekt ermöglichte ihr, erstmals eine eigene Arbeitsgruppe aufzubauen. Im Laufe ihrer Promotion und Zeit als Postdoktorandin arbeitete sie an Themenmodellen. Dies sind Modelle, mit denen die thematische Struktur in großen Datensätzen mit Textdokumenten untersucht werden kann. Zur Zeit der Antragstellung war das Sprachmodell GPT-2 [Rad+19] soeben erschienen und konnte erstmals eine überzeugende Leistung in der Generierung von längeren Texten demonstrieren. Jedoch waren die generierten Text noch schwer steuerbar. Daraus ergaben sich die Startbedingungen für das Projekt. Vortrainierte

große Sprachmodelle sollten mit Hilfe von Themenmodellen kontrollierbar gemacht werden. Die Expertise der Projektleiterin im Bereich der Themenmodelle und Textdaten in Verbindung mit neuesten Forschungsergebnissen im Bereich der Sprachmodelle, stellten eine einzigartige Möglichkeit dar, beide Themenbereiche gewinnbringend miteinander zu verknüpfen. Das Projekt sollte Forschung zu Sprachmodellen und Themenmodellen zusammenführen, um ein kombiniertes Modell zu entwickeln, das sowohl den Inhalt als auch den Textstil repräsentieren kann. Hierzu sollten moderne tiefe neuronale Netze zum Einsatz kommen, wobei erforscht werden sollte, wie diese mit diskreten Daten wie Texten umgehen könnten. Während es erfolgreiche Methoden für Bilddaten bereits gab, war unklar, wie ähnliche Ergebnisse für Texte erzielt werden könnten.

I.3. Planung und Ablauf des Vorhabens

Kurz nach Projektbeginn erhielt die Projektleiterin einen Ruf auf eine Juniorprofessur für Maschinelles Lernen ohne Tenure Track an der TU Kaiserslautern (heute RPTU Kaiserslautern-Landau). Daher wurde die Arbeitsgruppe an der TU Kaiserslautern aufgebaut. Die inhaltliche Ausrichtung blieb jedoch unverändert. Das Vorhaben wurde durch fünf Arbeitspakete grob gegliedert, die im folgenden kurz beschrieben werden sollen:

I.3.1. AP 1: Disentanglement von Stil und Thema

Das Ziel dieses Arbeitspaketes war die Entwicklung eines Sprachmodells, das Texte mit spezifischem Stil und Thema generieren kann. Dazu sollen spezielle Textstil-Embeddings gelernt werden, die den Stil eines Textes in einem Vektor repräsentieren. Diese Embeddings können dann als Input des Sprachmodells dienen, um dieses so zu steuern, dass der Text im gewünschten Stil ausgegeben wird. Hierzu sollten zunächst Testszenarien entworfen werden, mit denen die Funktionsweise des fertigen Modells validiert werden kann. Dann sollte die Erstellung eines Prototypen im Vordergrund stehen, der schließlich mit dem in AP 2 entwickelten hierarchischen Themenmodell integriert werden sollte.

I.3.2. AP 2: Hierarchisches Topic Model mit Disentanglement

In diesem Arbeitspaket sollte ein hierarchisches Themenmodell entwickelt werden, das Themen auf verschiedenen Abstraktionsebenen extrahieren kann. Hierbei sollte auch ein Sprachmodell integriert werden, welches die Textsegmentierung unterstützen sollte, sodass längere Texte automatisiert in kurze Sinnabschnitte und Wörter unterteilt werden können.

I.3.3. AP 3: Hintergrundwissen

Zur Extraktion von Wissen kann ein Stiltransfer ebenfalls hilfreich sein. Ein ausschweifend formulierter Text kann in einfache Sprache mit klar umrissenem Vokabular transferiert werden. Die Extraktion von Wissen wird dann standardisiert durchgeführt

In diesem Projekt sollte auch der umgekehrten Weg gegangen werden und Wissen vorgegeben werden, das in Sprachmodellen enthalten sein soll. Wir erforschen, wie relationales Wissen und logische Aussagen aus generiertem Text effizient extrahiert werden können, um diese mit externen Wissensdatenbanken abzugleichen. Wenn Diskrepanzen erkannt werden, soll durch eine zu entwickelnde Fehlerfunktion das Modell solange angepasst werden, bis der generierte Text frei von Fehlern ist. Die Extraktion der Relationen kann mit Hilfe von Methoden erfolgen, die z.B. auf

Reinforcement Learning beruhen und sich existierende Wissensdatenbanken mit gespeicherten Relationen zunutze machen.

I.3.4. AP 4: Transfer Learning/Domain Adaptation

Wir unterscheiden für unser Projekt zwischen zwei Szenarien, dem Transfer des Topics und dem Transfer des Stils. In dem Fall, dass das Topic transferiert werden soll und der Stil fixiert wird, kann zunächst durch einen Encoder eine latente Repräsentation des Textstils erstellt werden, die dann fixiert wird. Durch einen Austausch der latenten Repräsentation des Topics können dann Texte erzeugt werden, die den gleichen Stil und ein anderes Topic haben. Ebenso wird es möglich sein, den Stil beizubehalten und nur den Topic-Teil des Modells neu zu trainieren, so dass die extrahierten Topics vom Textstil abstrahieren. Das Verfahren für den Transfer des Stils ist analog zu dem für den Transfer des Topics.

Die Validierung des erfolgreichen Transfers kann durch Klassifikatoren erfolgen. Ein Klassifikator, der für die Topics eines Datensatzes trainiert ist, sollte vor dem Transfer ein Dokument zu seinem Topic zu ordnen und nach dem Transfer zu dem gewählten anderen Topic. Gleichzeitig muss sicher gestellt werden, dass der Stil gleich bleibt, was durch einen separaten Klassifikator für den Textstil erreicht werden kann. Dieser sollte keinen Unterschied zwischen den Texten vor und nach dem Transfer finden. Als Testdatensätze können hier Textcorpora mit bekannten Topics und bekanntem Textstil dienen.

Domain Adaptation verlangt die Adaption von sowohl den Topics als auch dem Stil. Z.B. kann ein Modell, das auf Sportnachrichten trainiert ist, für Nachrichten aus dem Kulturbereich adaptiert werden. Hierbei würde sich hauptsächlich das den Inhalt definierende Topic verändern, während der Schreibstil der Stil eines Zeitungsartikels bleibt. Aber auch einzelne Topics sind ähnlich zwischen beiden Nachrichtenbereichen. Z.B. haben Sportveranstaltungen genauso wie kulturelle Veranstaltungen eine zeitliche Abfolge mit Pausen und Highlights. Beide haben Geldgeber, Veranstaltungsorte und Zuschauer. Demnach könnte eine Adaption des bestehenden Sportmodells in Bezug auf das Hauptthema Sport zur Generierung von Nachrichten für den Kulturbereich dienen, ohne dass solche als Trainingsdaten in großer Zahl vorliegen. Die Validierung für die Domain Adaptation soll durch einen Klassifikator erfolgen, der generierte Texte von Originaltexten unterscheidet.

I.3.5. AP 5: Trainingsverfahren

Dieses Arbeitspaket soll ausgehend von einer intensiven Literaturrecherche, die Vor- und Nachteile existierender Trainingsverfahren analysieren und sie auf ihre Anwendbarkeit in dem Projekt hin prüfen. Dazu muss insbesondere die Skalierbarkeit im Hinblick auf die Größe des Datensatzes, des Vokabulars, der Anzahl der zu findenden Themen sowie der Komplexität der Komponente für den Textstil beurteilt werden. Hier sollten insbesondere Verfahren wie “Contrastive Learning”, “Reinforcement Learning”, aber auch “Normalizing Flows” oder weitere neu entwickelte Verfahren in Betracht gezogen werden.

I.4. Wissenschaftlicher und Technischer Stand

I.4.1. Themenmodelle

Variationale Autoencoder (VAEs) wurden erstmals 2013 von Kingma und Welling vorgestellt [KW13]. Seitdem gab es eine Vielzahl von Erweiterungen und Anwendungen dieses Modells. Um

Topics aus Texten zu extrahieren, wurden VAEs initial von Miao *et al.* [MYB16] angewandt. Dieses Modell verwendete eine Gauss'sche Aprioriverteilung, welche für Topic Models nicht geeignet ist. Auch Modelle, die das Training mit einer Dirichletverteilung ermöglichen, wurden seitdem entwickelt [SS17; Zha+18; Joo+19; FMM18; BK19a].

I.4.2. Hierarchische Generative Modelle

Die existierenden Arbeiten zu generativen neuronalen Netzwerken mit Hierarchien von Variablen sind relativ eingeschränkt, da Hierarchien schwierig zu trainieren sind und die Vorteile durch die detailliertere Repräsentation nicht immer die Nachteile kompensieren, die durch die gesteigerte Komplexität und größere Anzahl an Parametern hervorgerufen werden. Es gibt zwei Hauptansätze zum Trainieren generativer hierarchischer Modelle. Erstens können mehrere Schichten generativer Modelle übereinander geschichtet werden (siehe z.B. Burkhardt und Kramer [BK19b]). Zweitens kann die hierarchische Repräsentation auch in einer flachen Parameterstruktur gelernt werden [ZSE17]. Zhao *et al.* [ZSE17] schlagen hierzu einen Variational Ladder Autoencoder vor, um Hierarchien von Konzepten zu lernen, die mehr *disentangled* sind als die von korrespondierenden geschichteten Modellen. Dazu wird davon ausgegangen, dass komplexere Konzepte von tieferen Netzwerken modelliert werden müssen. Der Ansatz wird lediglich auf Bildern angewandt und zeigt eine gelernte Hierarchie, die so *disentangled* ist, dass Hintergrundfarbe, Gesichtsfarbe, Haarfarbe und andere Faktoren über einzelne Variablen gesteuert werden können.

I.4.3. Disentanglement von Stil und Thema

Die Frage, wie man Modelle lernen kann, die zwischen Stil und Thema in Texten unterscheiden, war ungelöst. Für Bilddaten gab es allerdings schon erfolgreiche Ansätze, die z.B. auf CNNs [GEB16], VAEs (conditional VAEs [SLY15; Yan+15]) oder adversarialen Netzen [Mak+16] beruhen. Diese Ansätze sind jedoch nicht ohne Weiteres auf Texte zu übertragen [Zha+17], da sie im Gegensatz zu Bildern keine kontinuierlichen Daten sind. Während es ohne Weiteres möglich war, ein Foto in den Stil eines Bildes von Picasso zu verwandeln, war es zu Projektbeginn nicht möglich, einen beliebigen Text in den Stil von z.B. Kafka umzuschreiben.

Bestehende Arbeiten beschäftigten sich erstens mit dem Stiltransfer (ohne explizites Disentanglement zwischen Stil und Topic), zweitens mit dem Disentanglement verschiedener Aspekte (jedoch nicht Stil und Topic) oder drittens mit der Generierung von Texten allgemein (ebenfalls ohne Disentanglement). Viertens gibt es Versuche, Syntax und Inhalt von Texten zu trennen [Gri+05; HG06]. Diese Arbeiten beziehen sich allerdings nicht auf den Textstil, welcher nicht allein durch die Syntax bestimmt wird, sondern z.B. auch durch die Art des verwendeten Vokabulars.

I.4.4. Transfer Learning und Domain Adaptation

Indem verschiedene Aspekte in Texten *disentangled* werden, wird die Interpretierbarkeit der gelernten Repräsentationen erhöht, was es leichter macht, ein Modell für andere Aufgaben anzupassen. Beispielsweise entwickelten Zhang *et al.* ein adversariales Netzwerk zur Domain Adaptation, das die Klassifikation von medizinischen Texte über Karzinome in verschiedene Aspekte separieren kann, sodass ein Klassifizierer auch für Lymphinvasion ohne annotierte Trainingsdokumente trainiert werden kann. Ansätze zur Domain Adaptation beruhen oft auf VAEs oder GANs [GL15; Gan+16], machen sich jedoch in Bezug auf Texte nicht ein Disentanglement von Stil und Thema zunutze.

I.4.5. Wissensextraktion

Zur Extraktion von Wissen trainierten Petroni *et al.* [Pet+19] Sprachmodelle und ließen einzelne Wörter in Sätzen ergänzen. Für die Berücksichtigung von externem Wissen beim Training, beschäftigen sich bestehende Arbeiten z.B. mit dem Lernen von Embeddings aus Wissensdatenbanken [RSR15]. Diese logischen Embeddings lernen jeweils ein Embedding für Paare von Entitäten und ein Embedding für Relationen. Diese können dann mit Verfahren zur Matrix Faktorisierung trainiert werden. Als Trainingsdaten dienen hier umfangreiche Wissensdatenbanken wie z.B. WordNet. Wie derartige Embeddings jedoch gewinnbringend in Language Models oder Topic Models integriert werden können, blieb offen.

II. Eingehende Darstellung

II.1. Durchgeführte Projektarbeiten

II.1.1. AP 1: Disentanglement von Stil und Thema

Unser Ziel war es, getrennte Inhalts- und Stilembeddings zu lernen. Dazu haben wir zunächst bestehende Embeddingmethoden getestet [Joh+19]. Um zu sehen wie gut diese Stilembeddings auf neue Datensätze generalisieren, haben wir in Abb. II.1 eine bestehende Methode [Joh+19], die auf einem Datensatz trainiert wurde, auf einem anderen getestet. Wie man sieht, sind die Embeddings nur begrenzt übertragbar. Daher haben wir einen neuen Ansatz zum Lernen von Stilembeddings entwickelt. Dazu gehen wir von folgenden Annahmen aus:

1. Trainingsdokumente $x_1, \ldots, x_n$ haben unbekannten Stil s_i und Inhalt (Bedeutung) m_i.
2. Es gibt in jedem x_i mindestens ein Wort (oder N-gram), das den Stil repräsentiert und eins, das den Inhalt repräsentiert.

Idee: Forciere beide Embeddings, Stilembeddings θ_s und Inhaltsembeddings θ_m, eine gewichtete Summe aus Wortembeddings zu sein.

$$\theta_s(x_i) = \sum_j x_{ij} \text{softmax}(f_s(x_i))_j, \tag{II.1}$$

wobei x_{ij} das Wortembedding für das Wort mit Index j im i-ten Dokument ist und f_s eine nichtlineare Transformation eines Dokumentes. Außerdem sollen θ_s, θ_m

1. nah an einer typischen Repräsentation für den jeweiligen Stil sein, die Centroid c_s genannt wird.
2. unterschiedliche Wörter erfassen, sodass das selbe Wort nur entweder zur Stilrepräsentation oder zur Inhaltsrepräsentation beiträgt und nicht zu beidem.
3. konsistent sein, in dem Sinne, dass ein Satz nicht mehreren Stilen zugeordnet wird.

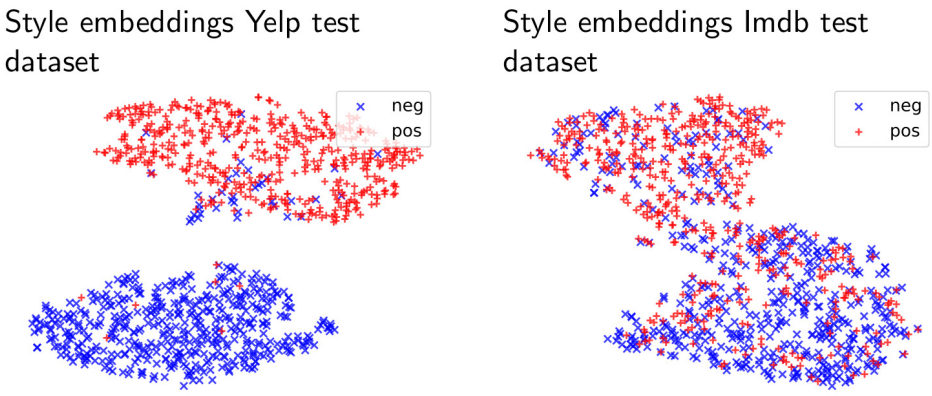


Abbildung II.1.: Diese Abbildung zeigt, dass die Stilembeddings, die auf dem Yelp-Datensatz gelernt wurden, für den Imdb-Datensatz nur begrenzt übertragen werden können. Diese Abbildung wurde von Phil Ostheimer im Rahmen des Projektes erstellt.

Diese drei Aspekte werden durch eine Lossfunktion abgebildet werden, die aus folgenden Teilen besteht. Ein Teil stellt sicher, dass die Daten auf Punkte abgebildet werden, die nahe am Clusterzentrum liegen:

$$L_{occ}(x_i) = ||\phi_s(\theta_s(x_i)) - c_s||^2 + ||\phi_m(\theta_m(x_i)) - c_m||^2. \tag{II.2}$$

Der andere Teil stellt sicher, dass Inhalts- und Stilrepräsentation möglichst unterschiedlich sind:

$$L_d(x_i) = -||\theta_s(x_i) - \theta_m(x_i)||^2. \tag{II.3}$$

Die so gelernten Embeddings werden in einem zweiten Schritt für die Anomalieerkennung in Texten genutzt, wie in Abb. II.2 illustriert. Dadurch können wir unsere Erkenntnisse zum Disentanglement von Stil und Inhalt auf das Feld der Anomalieerkennung übertragen. So können wir unterscheiden, ob eine erkannte Anomalie darauf beruht, dass der Inhalt des Textes anders ist, oder darauf, dass der Stil des Textes anders ist. Beispielsweise können wir damit unterscheiden, ob in einer Menge von Texten, einer viel negativer ist als der Rest oder ob der Gegenstand des Textes sich von dem anderer Texte stark unterscheidet.

Unsere empirischen Bewertungen bestätigen die Effizienz unseres Modells, was wir durch Visualisierungen und eine Reihe umfassender Experimente demonstriert haben. Außerdem haben wir gezeigt, dass das Modell in bestimmten Szenarien und Datensätzen erfolgreich sein kann, während die Leistung in anderen eingeschränkt ist. Die wichtigste Entdeckung ist, dass unser Modell hauptsächlich Inhaltsanomalien in Datensätzen mit Inhaltsvielfalt ins Visier nimmt, während es in einem Datensatz mit eingeschränktem Inhalt Inhalt und Stil separat fokussiert.

In einer weiteren Arbeit haben wir Disentanglement in einer Anwendung des Ingenieurswesens untersucht und Lift und Drag für das Airfoil-Design nutzbar gemacht. Dazu haben wir eine der ersten Arbeiten vorgestellt, die datengetrieben neue Airfoil-Designs ermöglicht und doch die Voreinstellung von Parametern erlaubt [Bha+23].

II.1.2. AP 2: Hierarchisches Topic Model mit Disentanglement

Hier haben wir zunächst verschiedene hierarchische Themenmodelle verglichen. Auf der einen Seite war dies das statistische Topic model [Ble+03], auf der anderen Seite ein hierarchisches Topic Model, das auf neuronalen Netzen beruht [Dua+21], und ein hybrides Modell [Iso+20]. Im Ergebnis haben wir festgestellt, dass das Modell, welches auf neuronalen Netzen beruht, keine

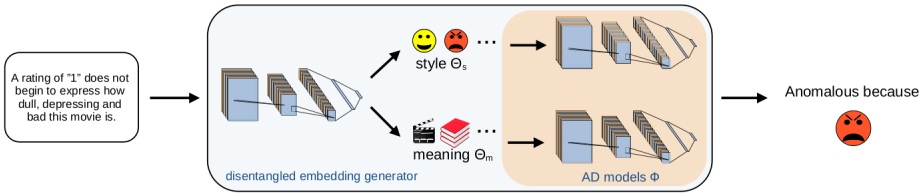


Abbildung II.2.: Diese Abbildung zeigt wie man separate Stil- und Inhaltsembeddings verwenden kann, um anomale Dokumente herauszufiltern, die entweder in Bezug auf Inhalt oder in Bezug auf Stil anomal sind. Diese Abbildung wurde von Phil Ostheimer im Rahmen des Projektes erstellt.

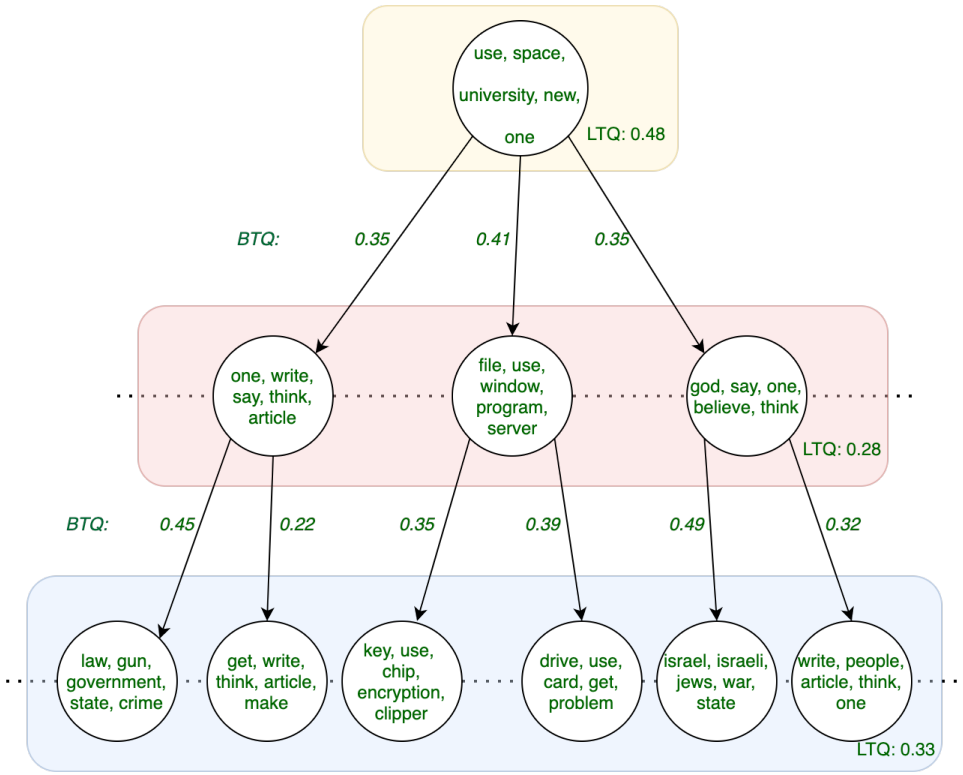


Abbildung II.3.: Ein Themenbaum mit zugehörigen BTQs und LTQs. Diese Abbildung wurde von Mayank Nagda im Rahmen des Projektes erstellt.

sehr gute Hierarchie lernt. Die Themen waren nur in der untersten Ebene der Hierarchie nachvollziehbar. Dies war allerdings an der jeweiligen Veröffentlichung nicht ersichtlich, da es bisher keine Evaluationsmethode gab, mit der man Themenhierarchien beurteilen konnte. Sämtliche Maßstäbe bezogen sich nur auf die einzelnen Themen und beachteten die Hierarchie nicht. Um das zu ändern und die Schwachpunkte bestehender Modelle besser beurteilen zu können, haben wir eine neue Evaluationsmethode entwickelt.

Die vorgeschlagene Methode (siehe Abb. II.3) baut auf der Intuition auf, dass eine gute Themenhierarchie zum einen in jeder Ebene kohärente Themen hat und zum anderen in jedem Zweig des Baumes kohärent ist. Wir berechnen daher die Kohärenz in Abzweigungen und in Ebenen eines Baumes und nutzen die resultierende Kohärenz um die Zweig-Themenqualität (BTQ) und die Ebenen-Themenqualität (LTQ) zu berechnen. Die Aggregation dieser beiden ergibt die hierarchische Themenqualität (HTQ).

Wir berechnen die Kohärenz in der Baumstruktur, indem wir einem Framework von Röder et al. folgen [RBH15]. Das Framework repräsentiert ein Kohärenzmaß als eine Kombination von vier Mengen: Segmentierung S, Konfirmationsmaß M, Wahrscheinlichkeitsschätzung P, und Aggregierung Σ. Die Autoren gruppieren diese Mengen in Dimensionen, die den Konfigurationsraum C aufspannen. Er wird beschrieben als $C = S \times M \times P \times \Sigma$. Zusammengefasst repräsentiert die Menge S verschiedene Möglichkeiten, eine Wortmenge in kleinere Untermengen aufzusplitten. Die Menge M beschreibt verschiedene Arten von Konfirmationsmaßen, die die Übereinstimmung zwischen Paaren von Worten bewerten, z.B. das NPMI (normalized pointwise mutual information) von zwei Wörtern. Die Menge P besteht aus verschiedenen Methoden, Wortwahrscheinlichkeiten in der Menge M abzuschätzen. Zuletzt besteht die Menge Σ aus Methoden, die verwendet werden, um die Werte zu aggregieren die von einem Konfirmationsmaß berechnet werden. Wir fügen eine weitere Dimension in den Konfigurationsraum hinzu, die die hierarchischen Wortmengen H_W repräsentiert. Wir definieren den modifizierten Konfigurationsraum wie folgt: $C^{(h)} = H_W \times S \times M \times P \times \Sigma$.

Die hierarchische Wortmenge H_W modifiziert den Input der Segmentierungsmenge, um die hierarchische Struktur miteinzubeziehen. H_W repräsentiert alle Wortmengen, die von Themen in Zweigen und Ebenen eines Baumes geformt werden. Für eine bestimmte Baumstruktur beinhaltet die Wortmenge W_B^b einer bestimmten Abzweigung b Wörter des Elternknoten und der Kindknoten. Genauso beinhaltet die Wortmenge W_L^l die Wörter von allen Knoten in einer bestimmten Ebene des Baumes. Wir verwenden die Wortmengen W_B und W_L, um BTQ und LTQ zu berechnen.

Für die Aggregierung führen wir einen Diversitäts-Maßstab ein [BK19c; DRB20]. Alle Konfirmations-Maßstäbe $\phi_N = \{\phi_1, \phi_2 ... \phi_{|s|}\}$ von allen Untermengen-Paaren S_i werden aggregiert, in dem der Mittelwert gebildet wird und werden mit dem Diversitäts-Maßstab (d) multipliziert, was die Themenqualität ergibt. Wir berechnen BTQ und LTQ für alle Abzweigungen und Ebenen im Baum. Sie werden wie folgt formuliert: $BTQ = \frac{\sum_{b=1}^{B} \phi^{W_B^b} \cdot d_b}{B}$ und $LTQ = \frac{\sum_{l=1}^{L} \phi^{W_L^l} \cdot d_l}{L}$. Der Mittelwert von BTQ und LTQ ist unsere finale Themenqualität eines kompletten Themenmodells.

Als ersten Schritt, um Themenmodelle mit Sprachmodellen zu verbinden, haben wir ein hybrides Modell entwickelt, das in der Lage ist, auf ungefiltertem Text zu lernen und Themen mit hoher Qualität zu lernen [NF24]. Dieses Modell macht sich zunutze, dass Wörter in Texten Abhängigkeiten mit unterschiedlichen Auflösungen haben. Es gibt im lokalen Kontext Abhängigkeiten, die oftmals syntaktischer Natur sind und es gibt Abhängigkeiten über weitere globale Kontexte, die tendenziell eher inhaltlicher Natur sind. Um das Wort „die“ vorherzusagen, ist es nicht nötig über den aktuellen Satz hinaus zu gehen, da es lediglich ein Artikel mit syntaktischer Funktion ist. Sagt man allerdings ein Wort wie „Netzwerk“ vorher, so ist es durchaus hilfreich zu wissen, ob das Wort oder andere ähnliche Wörter schon zwei Seiten vorher erwähnt wurden.

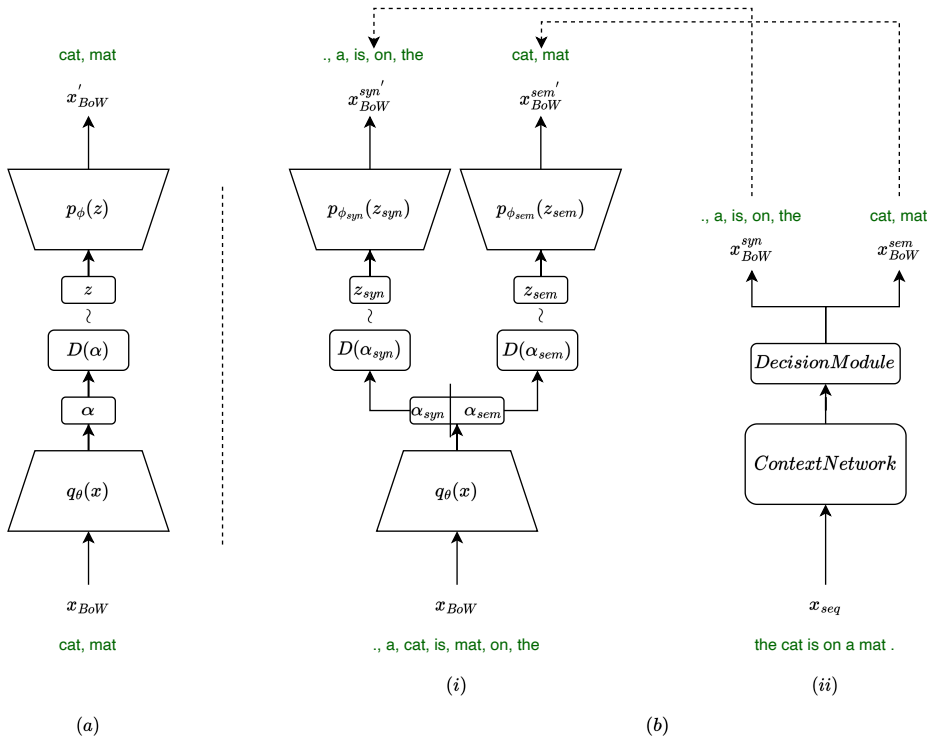


Abbildung II.4.: Diese Abbildung zeigt (a) die Standardarchitektur eines neuronalen Themenmodells und (b) unser vorgeschlagenes Modell. Es besteht aus einem Themenmodell mit VAE (links) und einem Kontextnetzwerk (rechts). Das Themenmodell hat zwei Decoder, einen für syntaktische Wörter und einen für semantische Wörter. Das Entscheidungsmodul entscheidet mit Hilfe des Kontextnetzwerk, welche Wörter syntaktisch sind und welche semantisch. Diese Abbildung wurde von Mayank Nagda im Rahmen des Projektes erstellt.

Diese Tatsache haben wir uns zunutze gemacht, um ein in Abbildung II.4 dargestelltes Modell zu entwickeln, das aus einem Themenmodell besteht, das semantische und syntaktische Themen modellieren kann, und durch ein Sprachmodell ergänzt wird, welches die Wörter in syntaktische und semantische Wörter differenziert. Das Ergebnis ist in Tabelle II.1 dargestellt, wo man an einem einfachen Datensatz mit Reviews von Yelp sehen kann, wie sowohl syntaktische als auch semantische Themen erzeugt werden.

In vielen Anwendungen sind Texte bereits in Kategorien oder Labels eingeteilt. Oftmals sind Texte zumindest teilweise gelabelt oder sollen für eine bestimmte Anwendung gelabelt werden. Für diesen Fall sind überwachte Themenmodelle geeignet. Solche überwachten Anwendungsfälle können auch verwendet werden, um bestehende Themenmodelle zu evaluieren und festzustellen, inwiefern die gelernten Themen mit den manuell verteilten Labels übereinstimmen. Hierzu haben wir neuronale Themenmodelle mit Labels trainiert. Beispielhafte Ergebnisse für gefundene Themen sind in Tabelle II.2 zu sehen. Hier kann man erkennen, dass das überwachte Themenmodell, die vorgegebenen Themen besser abdeckt. Beispielweise hat es separate Themen für *Windows* und *Mac*, sowie für *Autos* und *Motorräder*, während in dem unüberwachten Modell diese Themenpaare vermischt werden.

In diesem Arbeitspaket haben wir uns weiterhin mit dynamischen Themenmodellen beschäftigt [Jam+24] und ein neues kontinuierliches Themenmodell entwickelt, welches über die Zeit hinweg mit neuen Daten upgedated werden kann. Dynamische Themenmodelle wie Dynamic LDA erfassen die Entwicklung von Themen im Laufe der Zeit, erfordern jedoch, dass der gesamte Korpus von Anfang an verfügbar ist, was in der Praxis oft nicht realistisch ist. Um dieses Problem zu lösen, wurden Online-Themenmodelle entwickelt, die sich an Daten anpassen, die

Tabelle II.1.: Das vorgeschlagene Modell (oben) erzeugt syntaktische und semantische Themen, während das baseline-basierte neuronale Themenmodell (unten) dies nicht tut. Gezeigt sind fünf von zehn ausgewählten Themen des Yelp-Datensatzes. Die aktuellen Baselines ohne Vorverarbeitung führen zu unverständlichen und ineffektiven Themen, während die Vorverarbeitung den Verlust der reichhaltigen syntaktischen Repräsentation des Korpus zur Folge hat.

		Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
Ours	syntactic	i	good	la	ich	.	years	future	is	he	would
		me	loved	de	der	,	month	moments	has	she	could
		she	great	et	hier	!	days	chaser	always	they	did
		my	amazing	le	da	_	minutes	aspiring	than	you	when
		in	very	pas	nach	"	hour	acting	want	this	am
	semantic	theater	tacos	hotel	company	pancakes	salad	meat	game	dessert	cars
		band	bbq	rooms	office	waffle	sandwich	bbq	movie	chocolate	office
		seats	mexican	casino	wash	butter	dressing	red	games	cake	tire
		stage	burrito	bathroom	salon	breakfast	grilled	side	sports	gelato	repair
		store	wings	pool	car	cafe	bread	cut	watch	cupcakes	tires
Baseline	preprocessed	store	burger	drink	hotel	et	tacos	restaurant	love	und	buffet
		car	good	bar	rooms	un	mexican	table	store	ist	buffets
		work	fries	steak	hotels	le	salsa	food	amazing	der	rib
		shop	burgers	great	floor	de	chips	server	favorite	die	prime
		business	hot	table	vegas	pour	taco	sushi	home	das	steak
	raw	came	if	to	it	les	-	&	!	-	,
		were	get	the	the	pour	)	!	?	.	bread
		table	people	for	of	la	(	got	card	ribs	*
		food	into	if	,	des	fish	nails	said	orange	chicken
		very	room	she	for	par	happy	salon	charge	cabbage	sandwich

sequenziell eintreffen. Diese Modelle haben jedoch das Problem, dass sie langfristig erworbene Informationen vergessen. Hier setzt das Continual Learning an, das es ermöglicht, neue Probleme zu erlernen, ohne bereits gelerntes Wissen zu verlieren. Unser Continual Neural Topic Model (CoNTM) speichert Informationen über Zeit mit einem globalen Prior und erfasst lokale Muster in aktuellen Zeitabschnitten. Dadurch kann CoNTM globale thematische Muster verfolgen und deren zeitliche Entwicklung beobachten, ohne zuvor Erlerntes zu vergessen. Das graphische Modell für unseren Ansatz ist in Abb. II.5 zu sehen und der Lern-Algorithmus in Alg. 1. Unsere Experimente zeigen, dass das Modell zeitliche Entwicklungen gut abbildet und im Vergleich zu bisherigen Modellen effizienter ist, sowie eine bessere Themenqualität erreicht.

II.1.3. AP 3: Hintergrundwissen

Zu diesem Arbeitspaket hatten wir uns zunächst auf die Einbindung von Wissensgraphen (knowledge graphs) in neuronale Netzwerke fokussiert, was in einer Publikation resultierte, bei der es um die Erkennung von Argumentstrukturen ging [Li+21]. Da Sprachmodelle mittlerweile sehr viel implizites Weltwissen beinhalten, ist in einigen Bereichen der Einsatz von Wissensgraphen inzwischen nicht mehr zielführend. In Bereichen, in denen Spezialwissen oder eine Repräsentation einer bestimmten beschriebenen Umgebung erforderlich sind, sind Sprachmodelle allein jedoch nicht ausreichend. Als Beispiel haben wir uns mit textbasierten Spielen beschäftigt, die eine wohldefinierte, text-basierte Umgebung bereitstellen. Um in diesen Spielen erfolgreich zu sein, ist Wissen über Sprache sowie Wissen über das Spiel und seine Mechanik erforderlich.

		Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
	DVAE	windows	bible	players	space	car
		mac	god	team	launch	bike
		ati	truth	pitching	lunar	mph
		card	belief	player	mission	engine
		vga	christians	season	earth	riding
	Labeled DVAE	mac	windows	controller	car	riding
		quadra	microsoft	ide	cars	ride
		simms	fonts	bios	engine	bike
		slot	setup	isa	wheel	bikes
		monitors	z	hd	ford	rear
Label		pc-hardware	ms-windows	mac-hardware	autos	motorcycles

Tabelle II.2.: Ergebnisse von einem von uns trainierten überwachten Themenmodell (unten) im Vergleich mit einem unüberwachten Modell (oben). Man sieht, dass im unüberwachten Modell mehrere Themen (z.B. Windows und Mac) in einem Thema vermischt werden.

Algorithm 1 CoNTM algorithm

Input: Number of time slices T, stream of documents $\{\mathbf{w}^{i,1}\}_{i=1}^{n_1}, \cdots, \{\mathbf{w}^{i,T}\}_{i=1}^{n_T}$, number of Topics K, and number of training steps J
Output: Topics: $\{\hat{\phi}_t^{\text{local}}\}_{t=1}^{T}$, $\hat{\phi}^{\text{global}}$

for $t = 1 \dots T$ **do**
  Initialize θ
  Initialize $\Delta\hat{\phi}_t^{\text{local}}$
  $\{\mathbf{w}^{i,t}\}_{i=1}^{n_t} \leftarrow$ documents arrived at time t
  for $j = 1 \dots J$ **do**
    Update θ by the gradient

$$\nabla_\theta \sum_{i=1}^{n_t} \mathcal{L}(\theta, \mathbf{w}^{i,t} \mid \hat{\phi}^{\text{global}}, \Delta\hat{\phi}_t^{\text{local}})$$

    Update $\Delta\hat{\phi}_t^{\text{local}}$ by the gradient

$$\nabla_{\Delta\hat{\phi}_t^{\text{local}}} \sum_{i=1}^{n_t} \mathcal{L}(\theta, \mathbf{w}^{i,t} \mid \hat{\phi}^{\text{global}}, \Delta\hat{\phi}_t^{\text{local}})$$

  end for
  Set $\hat{\phi}_t^{\text{local}} \leftarrow \hat{\phi}^{\text{global}} + \Delta\hat{\phi}_t^{\text{local}}$
  Set $\hat{\phi}^{\text{global}} \leftarrow (1 - \rho_t)\hat{\phi}^{\text{global}} + \rho_t\hat{\phi}_t^{\text{local}}$
end for

Wie der Erfolg von ChatGPT und ähnlichen Modellen gezeigt hat, ist RL eine sehr vielversprechende Trainingsmethode für den NLP-Bereich, der bisher hauptsächlich Methoden aus dem sogenannten Q-Learning verwendet hat, oder im Fall von ChatGPT einfache "policy gradient"-Methoden. In unserer Arbeit haben wir eine modernere RL-Methode eingesetzt, die in der Robotik großen Erfolg hat und die Vorteile von Q-Learning und policy gradient miteinander verbindet: Soft Actor Critic (SAC). SAC ist eine Maximum Entropy RL Methode, bei der Zustände eher angesteuert werden, über die bislang wenig Informationen im System sind. Das heißt, dass automatisch die Exploration des Agenten gefördert wird. Im Bereich text-basierter Spiele haben wir Verbesserungen damit erreicht [LDF23; Li+24].

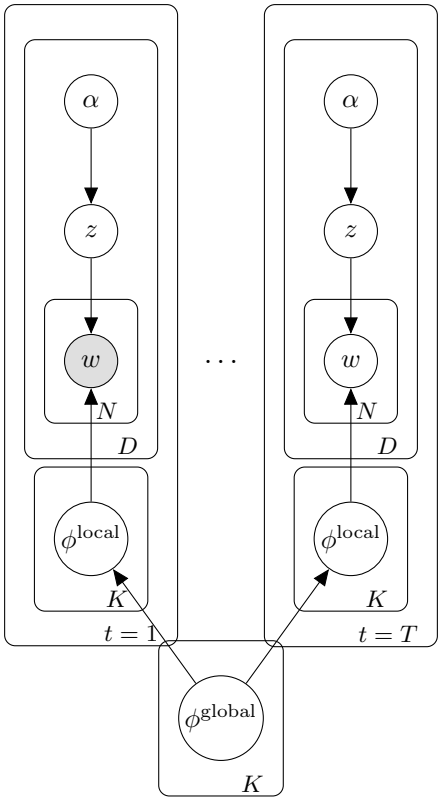


Abbildung II.5.: Das graphische Modell für unsere Methode. Die lokalen Modelle der einzelnen Zeitabschnitte t sind durch die globalen Parameter ϕ^{global} miteinander verbunden. Die Abbildung für dieses graphische Modell wurde von Charu James im Rahmen des Projektes erstellt.

In einer Kollaboration mit der Arbeitsgruppe Thermodynamik wurde eine Ontologie für das Feld der Thermodynamik erstellt, um in einem ersten Schritt Klausuraufgaben der Thermodynamik-Vorlesung automatisiert zu beantworten [Vol+24]. In einem nächsten Schritt werden wir uns näher damit befassen, wie die Antwort in Textform so gegeben werden kann, dass garantiert werden kann, dass die physikalischen Gesetze und Zusammenhänge, die in dem Wissensgraphen der Ontologie gespeichert sind, nicht verletzt werden.

Eine weitere Forschungsrichtung der Arbeitsgruppe beschäftigt sich mit hybriden ML-Methoden, die physikalisches Wissen und datengetriebenes Lernen miteinander verbinden [Jir+23; Spe+24; Nag+24]. Hierfür verwenden wir zwar keine Textdaten als Ein- und Ausgabe, jedoch ist das Prinzip ähnlich, da auch hier Vorhersagen des Modells auf Konsistenz mit gewissen Bedingungen geprüft werden.

II.1.4. AP 4: Transfer Learning/Domain Adaptation

Wir haben uns in diesem Arbeitspaket auf den Bereich Stiltransfer konzentriert. Es wurde eine Vielzahl von bestehenden Modellen erprobt und die vielversprechendsten ausgewählt. Das aus unserer Sicht beste, bestehende Modell zum Stiltransfer ohne Disentanglement ist der denoising adversarial Autoencoder [She+20]. Wir haben hier die Dimensionalität der Repräsentation im latenten Raum untersucht, um festzustellen, welchen Einfluss sie auf die Qualität der Ergebnisse hat. Zusammengefasst, sind unsere Erkenntnisse wie folgt:

- Wenn der Inhalt eines transferierten Satzes von Bedeutung ist, sollte der latente Raum größer sein.
- Wenn der Stil gut transferiert werden soll, kann der latente Raum kleiner sein.

- Wenn beides wichtig ist, sollte eine mittlere Größe des latenten Raumes gewählt werden

Die Qualität des Stiltransfers wird mit Stilklassifizierern gemessen, die einen transferierten Text klassifizieren, um festzustellen, ob sich sein Stil geändert hat. Wenn der latente Raum sehr klein gewählt wird, dann ist die Diversität der generierten Texte sehr gering. D.h. es werden nur sehr kurze, einfache Texte generiert, die nicht starkt variieren. Diese sind vom automatischen Klassifizierer leichter zu klassifizieren. D.h. viele existierende Methoden zum Stiltransfer erzielen gute Ergebnisse, indem sie sehr wenig komplexe Sätze generieren. Dadurch ist die Einordnung dieser Transfer-Methoden problematisch, da die Qualität der generierten Texte oft gering ist.

Die Notwendigkeit einer Standardisierung von Evaluierungsmetriken und experimentellen Setups für Stiltransfer für Text (TST) hat stark an Wichtigkeit gewonnen, da zunehmend neue Methoden entwickelt werden, deren relative Effektivität unklar ist. Daher haben wir eine Überblicksarbeit zu TST-Methoden und Evaluierungspraktiken erstellt, die den Bedarf an Standardisierung verdeutlicht und einen ungedeckten Bedarf an Validierung mit menschlicher Evaluation aufdeckt. Insgesamt ist die Standardisierung von Evaluierungsmetriken und experimentellen Setups notwendig, um eine Vergleichbarkeit von Methoden und einen Fortschritt in der TST-Forschung zu gewährleisten. Hierzu haben wir insgesamt 89 Publikationen in Bezug auf ihre Evaluationsmethoden ausgewertet [Ost+23].

Aspect	**Count**
No human evaluation	21/89
>1 non-validated metric	33/89
Statistical significance	3/89
Metrics	
Fluency	34
Content preservation	35
Style transfer strength	23
Other	24
Papers	
Fluency	45/89
Content preservation	66/89
Style transfer strength	77/89
Other	34/89

Tabelle II.3.: Gezeigt wird die Vielfalt der automatisierten TST-Evaluierungsmetriken. Alle untersuchten TST-Papiere setzen mindestens eine ein. Die Verwendung nicht validierter Metriken ist jedoch ein häufiges Phänomen und die statistische Signifikanz wird nur in einem Bruchteil der untersuchten Papiere berichtet.

Basierend auf diesen Erkenntnissen wurde von uns ein neues einheitliches Verfahren zur Evaluation mit großen Sprachmodellen vorgeschlagen [Ost+24]. Hierzu wird ein generierter Text wiederum einem Sprachmodell vorgelegt, welches den Text anhand verschiedener Dimensionen bewertet, die für den Stiltransfer relevant sind: Transferstärke, Inhaltskonsistenz und Text-Flüssigkeit. Bestehende Evaluierungsmethoden werden hierdurch durch eine einheitliche und stärkere Methode ersetzt, wie unser Vergleich mit durch Menschen erzeugten Annotationen ergeben hat.

II.1.5. AP 5: Trainingsverfahren

In diesem Arbeitspaket legten wir den Fokus auf das sogenannte „contrastive learning“. Die Kernidee ist in Abb. II.6 illustriert. Die Daten werden in einen Raum projiziert, in dem Texte mit dem selben Stil sich ähnlich sind, während Texte mit unterschiedlichem Stil leicht auseinander

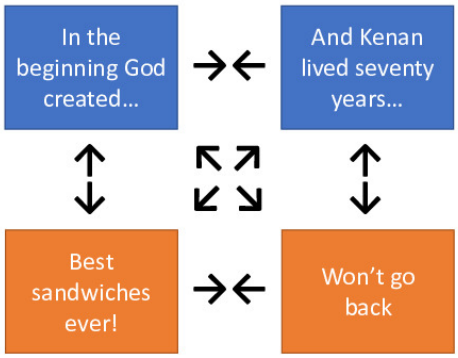


Abbildung II.6.: Unser Trainingsziel besteht darin, Satz-Embeddings desselben Stils nahe beieinander zu positionieren. In diesem Beispiel werden Rezensionen (in Orange) zueinander hingezogen, und Sätze einer Bibelversion (in Blau) werden ebenfalls nahe beieinander positioniert. Gleichzeitig werden die Embeddings verschiedener Stile, in diesem Fall Bibel vs. Rezensionen, voneinander entfernt. Diese Abbildung wurde von Phil Ostheimer im Rahmen des Projektes erstellt.

gehalten werden können. Formal lernen wir neuronale Encoder f und g, die jeweils zwei Inputsätze x und $\tilde{x}$ auf Vektorrepräsentationen mappen, sodass $f(x) \approx g(\tilde{x})$ genau dann, wenn x und $\tilde{x}$ den gleichen Stil haben: $\text{style}(x) = \text{style}(\tilde{x}) = s$. Um das zu erreichen, formen wir zunächst Paare von Sätzen, von denen einige den gleichen Stil haben und einige einen anderen Stil. Dann nutzen wir die „contrastive“ Zielfunktion, um die Vektorrepräsentationen der Sätze im selben Stil näher zueinander zu führen und die mit unterschiedlichem Stil auseinander zu drängen.

Für einen gegebenen Satz x mit $\text{style}(x) = s$, enthält unsere Menge an Kandidaten-Sätzen X_{cand} Sätze mit dem gleichen Stil und mit unterschiedlichen Stilen (siehe Abb. II.7): $X_{\text{cand}} = X_s \cup \overline{X_s}$. In unseren Experimenten setzen wir eine einfache Architektur ein, um die Ähnlichkeit zwischen den Outputs von f and g zu berechnen, nämlich das innere Produkt $f(x)^\top g(\tilde{x})$. Andere Funktionen, die in Betracht gezogen werden können, sind z.B. die „cosine similarity“ oder Kernelfunktionen. Wir verwenden eine einfache Architektur, da der Encoder unseres Modells bereits sehr mächtige Stil-Repräsentationen lernen soll, die leicht verglichen werden können.

Die Wahrscheinlichkeit dafür, dass ein Kandidatensatz $\tilde{x} \in X_{\text{cand}}$ den gleichen Stil hat wie x ist gegeben durch:

$$p(\tilde{x}|x, X_{\text{cand}}) = \frac{\exp[f(x)^\top g(\tilde{x})]}{\sum_{\tilde{x}' \in X_{\text{cand}}} \exp[f(x)^\top g(\tilde{x}')]}. \tag{II.4}$$

Unsere Zielfunktion ist es, die Wahrscheinlichkeit zu maximieren, alle Sätze $\tilde{x} \in X_{\text{cand}}$ zu identifizieren, für die $\tilde{x} \in X_{\text{cand}}$ wobei $\text{style}(x) = \text{style}(\tilde{x}) = s$ für alle Sätze x in den Trainingsdaten D:

$$\sum_{x \in D} \sum_{\tilde{x} \in X_s} \log p(\tilde{x}|x, X_{\text{cand}}) \tag{II.5}$$

Eine Visualisierung der gelernten Stil-Repräsentationen ist in Abb. II.8 zu sehen.

Wir haben unsere Stilembeddings aufbauend auf vortrainierten Sprachmodellen trainiert. Dazu haben wir BERT [Dev+18] und RoBERTa-Modelle verwendet und diese dann mit einer Verlustfunktion mit einem Contrastive-Loss getuned. Weiterhin haben wir diese Verlustfunktion auch mit einem Standard-Cross-Entropy-Loss kombiniert und die beiden Teile einzeln verglichen. Außerdem wurden one-shot und few-shot Szenarien verglichen. Ein zentrales Ergebnis war, dass die Schedule für die Lernrate von immenser Wichtigkeit ist. Wenn dieser Parameter nicht richtig gesetzt ist, schlägt das ganze Training fehl. Ein weiteres Ergebnis war, dass die Resultate mit mehr Trainingsdaten besser werden, aber auch schon mit wenigen Trainingsdaten gute Ergebnisse möglich sind (few shot). Wir haben verschiedene Architekturen auf einer Vielzahl von

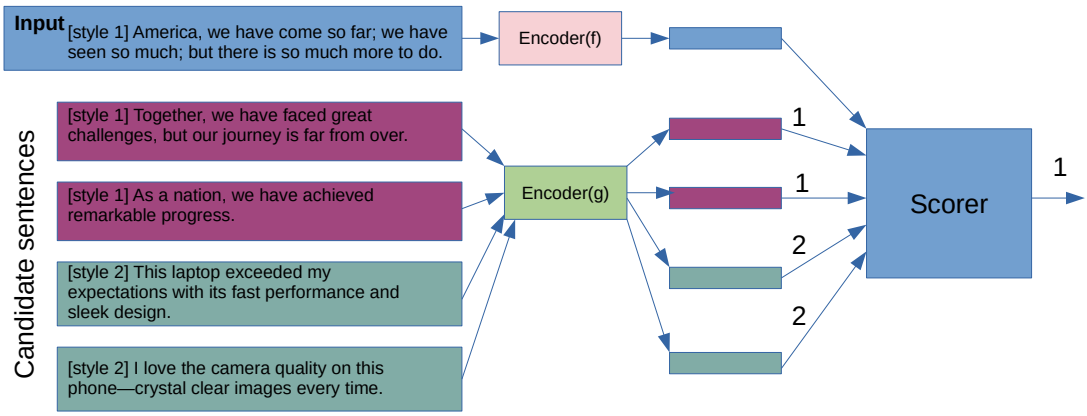


Abbildung II.7.: Lernen der Stil-Embeddings.

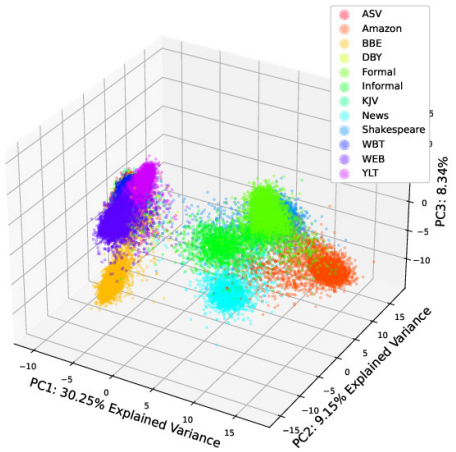


Abbildung II.8.: Visualisierung von Stil-Embeddings (mittels PCA), die von unserer Methode gelernt wurden. Die Abkürzungen mit drei Buchstaben bezeichnen verschiedene Versionen der Bibel. Diese Abbildung wurde von Phil Ostheimer im Rahmen des Projektes erstellt.

Datensätzen getestet. Unter anderem haben wir auch den StylePTB-Datensatz [Lyu+21] getestet, der eine Menge verschiedener Textstile enthält, die sehr detailliert voneinander abgegrenzt sind. In Tabelle II.4 zeigen wir die vorläufigen Ergebnisse, für die ein BERT-Modell, mit Finetuning auf dem StylePTB-Datensatz, für die Stilklassifikation anderer Stile eingesetzt wurde. Wie zu sehen ist, zeigen die mit Contrastive Learning gelernten Stilembeddings keine Verbesserung im Vergleich zu einem mit einem Standard-Cross-Entropy-Loss trainierten Modell.

Dataset	Not	CEL	CEL+CL	CL
Yelp	94.3	92.9	90.5	90.1
Amazon	74.9	72.1	74.5	55.6
GYAFC	88.4	88.0	82.1	82.1
Shakespeare	85.7	**86.4**	83.4	83.5
Bible	50.9	**52.8**	42.9	46.7

Tabelle II.4.: Vergleich der auf StylePTB gelernten Style Embeddings eines BERT-Modells, angewendet auf die genannten Stile/Datensätze zur Stilklassifikation, ohne Finetuning auf StylePTB-Datensatz (Not), mit Finetuning durch einen Cross-Entropy-Loss (CEL), einer Kombination aus Cross-Entropy-Loss und Contrastive-Loss (CEL+CL), und nur Constrastive-Loss (CL)

Unsere Ergebnisse zeigen daher, dass im Rahmen der bestehenden Datensätze, Stilembeddings am besten mit dem Cross-Entropy-Loss gelernt werden können. Die verwendeten Datensätze sind jedoch in Bezug auf die Vielfalt der enthaltenen Textstile eher eingeschränkt. In zukünftigen Arbeiten sollten zunächst besser geeignete Datensätze für das Training identifiziert werden.

II.2. Wichtige Positionen des zahlenmäßigen Nachweises

Das Vorhaben SeDis stellt ein personalintensives Forschungsvorhaben dar, sodass ein hoher Anteil der finanziellen Ressourcen für wissenschaftliches Personal aufgewendet wurde. Weiterhin wurden für das Training und die Ausführung der vortrainierten rechenintensiven Sprachmodelle zwei DGX A100 Systeme beschafft.

II.3. Notwendigkeit und Angemessenheit der geleisteten Projektarbeiten

Die durchgeführten Arbeiten entsprachen den im Antragstext festgehaltenen Planungen und waren in ihrem Umfang weitgehend hinreichend, um die festgelegten Ziele zu erreichen. Eine kostenneutrale Verlängerung musste nicht beantragt werden, da der Projektfortschritt der Planung angemessen war. Die im Projekt erzeugten Ressourcen, insbesondere der Aufbau und der Transfer von Wissen zum Training und der Verwendung von vortrainierten Sprachmodellen, waren für den Aufbau der Arbeitsgruppe essentiell. Im Hinblick auf das Ziel der Förderung, eine Nachwuchsgruppe zu etablieren, die langfristig erfolgreich arbeitet, und Nachwuchswissenschaftlerinnen gezielt und nachhaltig zu fördern, bewerten wir die geleisteten Arbeiten als sehr erfolgreich. Unter anderem waren mehr als die Hälfte der über das Projekt angestellten Promovierenden Frauen.

II.4. Verwertbarkeit der Ergebnisse

Die in diesem Projekt entwickelten Methoden zur Evaluation von Stiltransfer und Dynamischen Themenmodellen können in der Zukunft verwendet werden, um neu entwickelte Modelle besser zu testen und zu vergleichen. Die entwickelten Reinforcement Learning Methoden haben zum Ziel, die allgemeine Entscheidungsfindung mit Hilfe von neuronalen Netzen zu verbessern. Erkenntnisse im Bereich des Disentanglement verschiedener Aspekte von Textdaten haben zu verbesserten Themenmodellen geführt und werden in Zukunft auch für die Weiterentwicklung von Modellen zum Textstiltransfer und der kontrollierten Textgenerierung hilfreich sein.

Außerhalb des Anwendungsbereichs von Texten werden die gewonnenen Erkenntnisse und entwickelten Methoden auch in anderen Bereichen einen Einfluss haben. Beispielsweise spielt das Disentanglement von verschiedenen Aspekten im Bereich physikalischer Modelle eine große Rolle und die Generierung von synthetischen Daten nach speziellen Kriterien ist auch in anderen technischen Bereichen von großer Wichtigkeit.

Die Forschungstätigkeit von SeDis wurde über Konferenz- und Workshopbeiträge zum Thema Maschinelles Lernen verbreitet. Auch in der Lehre flossen die Arbeiten ein. Zum Beispiel wurde eine neue Vorlesung zum Thema generativer Sprachmodelle entwickelt und durchgeführt, mehrere Abschlussarbeiten zum Thema wurden betreut und studentische Projekte im Zusammenhang mit dem Projekt wurden durchgeführt.

II.5. Fortschritt anderer Stellen im Bereich des Projekts

Der extreme Fortschritt im Bereich großer Sprachmodelle und die Verfügbarkeit dieser Modelle für die breite Öffentlichkeit in der Zwischenzeit haben zu enormen Veränderungen in der Forschungslandschaft geführt. ChatGPT wurde im November 2022 veröffentlicht und hat eine Welle an Publikationen und neuen Forschungsrichtungen ausgelöst. Insbesondere das sogenannte "Instruction-based Fine Tuning" [Ouy+22], das auf Reinforcement Learning basiert, hat Sprachmodelle besser nutzbar gemacht. Außerdem gab es Weiterentwicklungen im Bereich des Prompting, um bestehendes Wissen besser zu integrieren, was im "Retrieval Augmented Generation (RAG)" [Lew+20] resultierte.

II.6. Erfolgte oder geplante Veröffentlichungen

Im Rahmen des Projektes erfolgte Publikationen:

[Wil+24] Justus Will u. a. „Enhancing Realism in Batch Distillation Simulations: Data-Efficient Time Series Style Transfer with Transformers". In: *ECML/PKDD, Workshop on Machine Learning for Chemistry and Chemical Engineering*. 2024.

[Ost+24] Phil Ostheimer, Mayank Nagda, Marius Kloft und Sophie Fellenz. „Text Style Transfer Evaluation Using Large Language Models". In: *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-Coling)*. 2024.

[Man+24] Laura Manduchi u. a. „On the Challenges and Opportunities in Generative AI". In: *arXiv preprint arXiv:2403.00025* (2024).

[NF24] Mayank Nagda und Sophie Fellenz. „Putting Back the Stops: Integrating Syntax with Neural Topic Models". In: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. 2024.

[Jam+24] Charu James, Mayank Nagda, Nooshin Haji Ghassemi, Marius Kloft und Sophie Fellenz. „Evaluating Dynamic Topic Models“. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*. 2024.

[Li+24] Weichen Li, Rati Devidze, Waleed Mustafa und Sophie Fellenz. „Ethics in Action: Training Reinforcement Learning Agent for Moral Decision-making In Text-based Adventure Games“. In: *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2024.

[LDF23] Weichen Li, Rati Devidze und Sophie Fellenz. „Learning to Play Text-based Adventure Games with Maximum Entropy Reinforcement Learning“. In: *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*. 2023.

[Ost+23] Phil Ostheimer, Mayank Kumar Nagda, Marius Kloft und Sophie Fellenz. „A Call for Standardization and Validation of Text Style Transfer Evaluation“. In: *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, Juli 2023, S. 10791–10815. DOI: 10.18653/v1/2023.findings-acl.687.

[Li+21] Weichen Li, Patrick Abels, Zahra Ahmadi, Sophie Burkhardt, Benjamin Schiller, Iryna Gurevych und Stefan Kramer. „Topic-Guided Knowledge Graph Construction for Argument Mining“. In: *2021 IEEE International Conference on Big Knowledge, ICBK 2021, Auckland, New Zealand, December 7-8, 2021*. Hrsg. von Lei Chen und Baltasar Fernández-Manjón. IEEE, 2021, S. 315–322. DOI: 10.1109/ICKG52313.2021.00049.

[Nag+21] Mayank Kumar Nagda, Charu James, Marius Kloft und Sophie Burkhardt. „Hierarchical Topic Evaluation: Statistical vs. Neural Models“. In: *Bayesian Deep Learning Workshop at NeurIPS*. 2021.

Im Zusammenhang mit dem Projekt erstellte Publikationen:

[Mon+24] Marcio Monteiro, Charu Karakkaparambil James, Marius Kloft und Sophie Fellenz. „Characterizing Text Datasets with Psycholinguistic Features“. In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, 2024.

[Nag+24a] Mayank Nagda, Phil Ostheimer, Thomas Specht, Frank Rhein, Fabian Jirasek, Marius Kloft und Sophie Fellenz. „PITs: Physics-Informed Transformers for Predicting Chemical Phenomena“. In: *ECML/PKDD, Workshop on Machine Learning for Chemistry and Chemical Engineering*. 2024.

[Nag+24b] Mayank Nagda, Phil Ostheimer, Thomas Specht, Frank Rhein, Fabian Jirasek, Marius Kloft und Sophie Fellenz. *SetPINNs: Set-based Physics-informed Neural Networks*. 2024. arXiv: 2409.20206 [cs.LG].

[Spe+24] Thomas Specht, Mayank Nagda, Sophie Fellenz, Stephan Mandt, Hans Hasse und Fabian Jirasek. „HANNA: Hard-constraint Neural Network for Consistent Activity Coefficient Prediction“. In: *Chem. Sci.* (2024), S. -. DOI: 10.1039/D4SC05115G.

[Vol+24] Luisa Vollmer, Sophie Fellenz, Fabian Jirasek, Heike Leitte und Hans Hasse. „KnowTD - An Actionable Knowledge Representation System for Thermodynamics“. In: *Journal of Chemical Information and Modeling* 64.15 (2024). PMID: 39042488, S. 5878–5887. DOI: 10.1021/acs.jcim.4c00647. eprint: https://doi.org/10.1021/acs.jcim.4c00647.

[Liz+24] Philipp Liznerski, Saurabh Varshneya, Ece Calikus, Sophie Fellenz und Marius Kloft. „Reimagining Anomalies: What If Anomalies Were Normal?“ In: *arXiv preprint arXiv:2402.14469* (2024).

[Hau+23] Tony Hauptmann, Sophie Fellenz, Laksan Nathan, Oliver Tüscher und Stefan Kramer. „Discriminative machine learning for maximal representative subsampling“. In: *Scientific Reports* 13.20925 (2023). DOI: https://doi.org/10.1038/s41598-023-48177-3.

[Fra+23] Billy Joe Franks, Benjamin Dinkelmann, Marius Kloft und Sophie Fellenz. „Ordinal Regression for Difficulty Estimation of StepMania Levels“. In: *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*. 2023.

[Bha+23] Asmita Bhat, Nooshin Haji-Ghassemi, Deepak Nagaraj und Sophie Fellenz. „Constraint-Based Parameterization and Disentanglement of Aerodynamic Shapes using Deep Generative Models“. In: *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*. 2023.

[Har+23] Fabian Hartung u. a. „Deep Anomaly Detection on Tennessee Eastman Process Data“. In: *Chemie Ingenieur Technik* 95.7 (2023), S. 1077–1082. DOI: https://doi.org/10.1002/cite.202200238. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cite.202200238.

[Jir+23] Fabian Jirasek, Sophie Fellenz, Robert Bamler, Michael Bortz, Marius Kloft, Stephan Mandt und Hans Hasse. „Making thermodynamic models of mixtures predictive by machine learning: matrix completion of pair interactions“. In: *ECML/PKDD, Workshop on Neuro-Explicit AI and Expert-informed Machine Learning for Engineering and Physical Sciences*. 2023.

[Bur+21] Sophie Burkhardt, Jannis Brugger, Nicolas Wagner, Zahra Ahmadi, Kristian Kersting und Stefan Kramer. „Rule extraction from binary neural networks with convolutional rules for model validation“. In: *Frontiers in Artificial Intelligence* 4 (2021), S. 90.

Literaturverzeichnis

[Rad+19] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei und Ilya Sutskever. „Language Models are Unsupervised Multitask Learners“. In: 2019.

[KW13] Diederik P Kingma und Max Welling. „Auto-encoding variational Bayes“. In: *arXiv preprint arXiv:1312.6114* (2013).

[MYB16] Yishu Miao, Lei Yu und Phil Blunsom. „Neural variational inference for text processing“. In: *ICML*. 2016, S. 1727–1736.

[SS17] Akash Srivastava und Charles Sutton. „Autoencoding Variational Inference For Topic Models“. In: *ICLR*. 2017.

[Zha+18] Hao Zhang, Bo Chen, Dandan Guo und Mingyuan Zhou. „WHAI: Weibull Hybrid Autoencoding Inference for Deep Topic Modeling“. In: *ICLR*. 2018.

[Joo+19] Weonyoung Joo, Wonsung Lee, Sungrae Park und Il-Chul Moon. „Dirichlet Variational Autoencoder“. In: *CoRR* abs/1901.02739 (2019). arXiv: `1901.02739`.

[FMM18] Mikhail Figurnov, Shakir Mohamed und Andriy Mnih. „Implicit Reparameterization Gradients“. In: *NeurIPS*. 2018, S. 439–450.

[BK19a] Sophie Burkhardt und Stefan Kramer. „Decoupling Sparsity and Smoothness in the Dirichlet Variational Autoencoder Topic Model“. In: *Journal of Machine Learning Research* 20.131 (2019), S. 1–27.

[BK19b] Sophie Burkhardt und Stefan Kramer. „Multi-Label Classification using Stacked Hierarchical Dirichlet Processes with Reduced Sampling Complexity“. In: *KAIS* (2019), S. 1–23.

[ZSE17] Shengjia Zhao, Jiaming Song und Stefano Ermon. „Learning hierarchical features from deep generative models“. In: *ICML*. JMLR. org. 2017, S. 4091–4099.

[GEB16] L. A. Gatys, A. S. Ecker und M. Bethge. „Image Style Transfer Using Convolutional Neural Networks“. In: *CVPR*. Juni 2016, S. 2414–2423.

[SLY15] Kihyuk Sohn, Honglak Lee und Xinchen Yan. „Learning Structured Output Representation using Deep Conditional Generative Models“. In: *NIPS*. Hrsg. von C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama und R. Garnett. Curran Associates, Inc., 2015, S. 3483–3491.

[Yan+15] Xinchen Yan, Jimei Yang, Kihyuk Sohn und Honglak Lee. „Attribute2Image: Conditional Image Generation from Visual Attributes“. In: *arXiv preprint arXiv:1512.00570* (2015).

[Mak+16] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow und Brendan Frey. *Adversarial Autoencoders*. arXiv:1511.05644 [cs]. Mai 2016.

[Zha+17] Jake Zhao, Yoon Kim, Kelly Zhang, Alexander M Rush und Yann LeCun. „Adversarially regularized autoencoders“. In: *arXiv preprint arXiv:1706.04223* (2017).

[Gri+05] Thomas Griffiths, Mark Steyvers, David Blei und Joshua Tenenbaum. „Integrating Topics and Syntax“. In: *Advances in Neural Information Processing Systems*. Hrsg. von L. Saul, Y. Weiss und L. Bottou. Bd. 17. MIT Press, 2005.

[HG06] Bo-June (Paul) Hsu und James Glass. „Style & Topic Language Model Adaptation Using HMM-LDA“. In: *EMNLP*. EMNLP '06. Sydney, Australia: Association for Computational Linguistics, 2006, S. 373–381. ISBN: 1-932432-73-6.

[GL15] Yaroslav Ganin und Victor Lempitsky. „Unsupervised Domain Adaptation by Backpropagation“. In: *ICML*. Hrsg. von Francis Bach und David Blei. Bd. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, Juli 2015, S. 1180–1189.

[Gan+16] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand und Victor Lempitsky. „Domain-adversarial Training of Neural Networks“. In: *Journal of Machine Learning Research* 17.1 (Jan. 2016), S. 2096–2030. ISSN: 1532-4435.

[Pet+19] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller und Sebastian Riedel. „Language Models as Knowledge Bases?“ In: *EMNLP*. 2019.

[RSR15] Tim Rocktäschel, Sameer Singh und Sebastian Riedel. „Injecting logical background knowledge into embeddings for relation extraction“. In: *NAACL*. 2015, S. 1119–1129.

[Joh+19] Vineet John, Lili Mou, Hareesh Bahuleyan und Olga Vechtomova. „Disentangled Representation Learning for Non-Parallel Text Style Transfer“. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Juli 2019, S. 424–434. DOI: 10.18653/v1/P19-1041.

[Bha+23] Asmita Bhat, Nooshin Haji-Ghassemi, Deepak Nagaraj und Sophie Fellenz. „Constraint-Based Parameterization and Disentanglement of Aerodynamic Shapes using Deep Generative Models“. In: *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*. 2023.

[Ble+03] David M Blei, Thomas L Griffiths, Michael I Jordan, Joshua B Tenenbaum u. a. „Hierarchical topic models and the nested Chinese restaurant process.“ In: *NIPS*. Bd. 16. 2003.

[Dua+21] Zhibin Duan, Dongsheng Wang, Bo Chen, Chaojie Wang, Wenchao Chen, Yewen Li, Jie Ren und Mingyuan Zhou. „Sawtooth Factorial Topic Embeddings Guided Gamma Belief Network“. In: *International Conference on Machine Learning*. PMLR. 2021, S. 2903–2913.

[Iso+20] Masaru Isonuma, Junichiro Mori, Danushka Bollegala und Ichiro Sakata. „Tree-Structured Neural Topic Model“. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Juli 2020, S. 800–806. DOI: 10.18653/v1/2020.acl-main.73.

[RBH15] Michael Röder, Andreas Both und Alexander Hinneburg. „Exploring the space of topic coherence measures“. In: *Proceedings of the eighth ACM international conference on Web search and data mining*. 2015, S. 399–408.

[BK19c] Sophie Burkhardt und Stefan Kramer. „Decoupling Sparsity and Smoothness in the Dirichlet Variational Autoencoder Topic Model.“ In: *J. Mach. Learn. Res.* 20.131 (2019), S. 1–27.

[DRB20] Adji B Dieng, Francisco JR Ruiz und David M Blei. „Topic modeling in embedding spaces“. In: *Transactions of the Association for Computational Linguistics* 8 (2020), S. 439–453.

[NF24] Mayank Nagda und Sophie Fellenz. „Putting Back the Stops: Integrating Syntax with Neural Topic Models“. In: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. 2024.

[Jam+24] Charu James, Mayank Nagda, Nooshin Haji Ghassemi, Marius Kloft und Sophie Fellenz. „Evaluating Dynamic Topic Models“. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*. 2024.

[Li+21] Weichen Li, Patrick Abels, Zahra Ahmadi, Sophie Burkhardt, Benjamin Schiller, Iryna Gurevych und Stefan Kramer. „Topic-Guided Knowledge Graph Construction for Argument Mining“. In: *2021 IEEE International Conference on Big Knowledge, ICBK 2021, Auckland, New Zealand, December 7-8, 2021*. Hrsg. von Lei Chen und Baltasar Fernández-Manjón. IEEE, 2021, S. 315–322. DOI: 10.1109/ICKG52313.2021.00049.

[LDF23] Weichen Li, Rati Devidze und Sophie Fellenz. „Learning to Play Text-based Adventure Games with Maximum Entropy Reinforcement Learning“. In: *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*. 2023.

[Li+24] Weichen Li, Rati Devidze, Waleed Mustafa und Sophie Fellenz. „Ethics in Action: Training Reinforcement Learning Agent for Moral Decision-making In Text-based Adventure Games“. In: *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2024.

[Vol+24] Luisa Vollmer, Sophie Fellenz, Fabian Jirasek, Heike Leitte und Hans Hasse. „KnowTD - An Actionable Knowledge Representation System for Thermodynamics“. In: *Journal of Chemical Information and Modeling* 64.15 (2024). PMID: 39042488, S. 5878–5887. DOI: 10.1021/acs.jcim.4c00647. eprint: https://doi.org/10.1021/acs.jcim.4c00647.

[Jir+23] Fabian Jirasek, Sophie Fellenz, Robert Bamler, Michael Bortz, Marius Kloft, Stephan Mandt und Hans Hasse. „Making thermodynamic models of mixtures predictive by machine learning: matrix completion of pair interactions“. In: *ECML/PKDD, Workshop on Neuro-Explicit AI and Expert-informed Machine Learning for Engineering and Physical Sciences*. 2023.

[Spe+24] Thomas Specht, Mayank Nagda, Sophie Fellenz, Stephan Mandt, Hans Hasse und Fabian Jirasek. „HANNA: Hard-constraint Neural Network for Consistent Activity Coefficient Prediction“. In: *Chem. Sci.* (2024), S. -. DOI: 10.1039/D4SC05115G.

[Nag+24] Mayank Nagda, Phil Ostheimer, Thomas Specht, Frank Rhein, Fabian Jirasek, Marius Kloft und Sophie Fellenz. *SetPINNs: Set-based Physics-informed Neural Networks*. 2024. arXiv: 2409.20206 [cs.LG].

[She+20] Tianxiao Shen, Jonas Mueller, Regina Barzilay und Tommi Jaakkola. „Educating Text Autoencoders: Latent Representation Guidance via Denoising“. In: *Proceedings of the 37th International Conference on Machine Learning*. ICML'20. JMLR.org, 2020.

[Ost+23] Phil Ostheimer, Mayank Kumar Nagda, Marius Kloft und Sophie Fellenz. „A Call for Standardization and Validation of Text Style Transfer Evaluation“. In: *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, Juli 2023, S. 10791–10815. DOI: 10.18653/v1/2023.findings-acl.687.

[Ost+24] Phil Ostheimer, Mayank Nagda, Marius Kloft und Sophie Fellenz. „Text Style Transfer Evaluation Using Large Language Models“. In: *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-Coling)*. 2024.

[Dev+18] Jacob Devlin, Ming-Wei Chang, Kenton Lee und Kristina Toutanova. „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding“. In: *ArXiv* abs/1810.04805 (2018).

[Lyu+21] Yiwei Lyu, Paul Pu Liang, Hai Pham, Eduard H. Hovy, Barnabás Póczos, Ruslan Salakhutdinov und Louis-Philippe Morency. „StylePTB: A Compositional Benchmark for Fine-grained Controllable Text Style Transfer“. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*. Hrsg. von Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty und Yichao Zhou. Association for Computational Linguistics, 2021, S. 2116–2138. DOI: 10.18653/v1/2021.naacl-main.171.

[Ouy+22] Long Ouyang u. a. „Training language models to follow instructions with human feedback“. In: *Advances in Neural Information Processing Systems*. Hrsg. von S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho und A. Oh. Bd. 35. Curran Associates, Inc., 2022, S. 27730–27744.

[Lew+20] Patrick Lewis u. a. „Retrieval-augmented generation for knowledge-intensive NLP tasks“. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS '20. Vancouver, BC, Canada: Curran Associates Inc., 2020. ISBN: 9781713829546.