Niels Pinkwart, Johannes Konert (Hrsg.): Die 17. Fachtagung Bildungstechnologien,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2019 45

Themenübergreifende Diskursklassifikation auf Basis von Word Embeddings und Sequenzfeatures

Experimente hin zu einer automatischen Diskursanalyse

Tim Steuer[1] und Christoph Rensing[2]

Abstract: Zur Beobachtung von kollaborativen Lernprozessen ist Diskursanalyse ein hilfreiches Werkzeug. Dazu wird der Textkorpus von Annotatoren händisch segmentiert und die Segmente nach ihrer Funktion klassifiziert. Dies ist zeitaufwendig und kostspielig. Automatische Modelle versprechen Zeitersparnis sowie Echtzeitanalysen des Diskurses. Diese könnten direktes Feedback, beispielsweise durch Visualisierungen, an die Lernenden ermöglichen. Automatische Modelle benötigen jedoch manuell annotierte Trainingsdaten. Außerdem sind sie meist vom Diskursvokabular abhängig und generalisieren schlecht über Themengrenzen hinweg. Die dadurch notwendige, häufige Neuerstellung von Trainingskorpora, verringert die Zeitersparnis durch Automatisierung und macht Echtzeit Analyse unmöglich. In dieser Arbeit wird ein Klassifikationsverfahren basierend auf Word Embeddings und Sequenz Features vorgestellt, welches vier Arten von Diskurssegmenten unterscheidet. Das Verfahren erreicht gute Evaluationsergebnisse, mit einer besseren Klassifikationsgüte als Verfahren aus verwandten Arbeiten (Cohens $\kappa > 0.7$). Außerdem generalisiert das Verfahren, auf dem Korpus, ohne weiteres Training von einem Themengebiet auf ein anderes. Dies würde die Notwendigkeit von themenspezifischen Trainingskorpora stark verringern.

Keywords: Diskursanalyse, Word Embeddings, Machine Learning

1 Motivation

In modernen E-Learning Szenarien ist textueller Diskurs allgegenwärtig. Lernende nutzen beispielsweise Diskussionsforen oder Chats, um zu diskutieren, sich zu organisieren oder um den Kontakt zu ihren Mitstreitern zu halten. Die so entstehenden Unterhaltungen sind mehr als reiner Informationsaustausch, sondern fördern den Lernprozess. Dies ist eine der Grundannahmen von sozio-konstruktivistischen Lerntheorien und von dem darauf aufbauenden Computer unterstütztem kooperativem Lernen (CSCL). Um zu verstehen, wie Teilnehmende in der Diskussion lernen, gibt es im Forschungsfeld CSCL Kodierungshandbücher, mit deren Hilfe man die unterschiedlichen Funktionen der einzelnen Diskussionsbeiträge sichtbar machen kann. Dazu wird jeder einzelne Diskussionsbeitrag zunächst segmentiert und die einzelnen Segmente werden dann jeweils ihrer Funktion

[1] Technische Universität Darmstadt, Fachgebiet Multimedia Kommunikation, Rundeturmstraße 10, 64283 Darmstadt, tim.steuer@kom.tu-darmstadt.de
[2] Technische Universität Darmstadt, Fachgebiet Multimedia Kommunikation, Rundeturmstraße 10, 64283 Darmstadt, christoph.rensing@kom.tu-darmstadt.de

doi:10.18420/delfi2019_234

nach klassifiziert. Hat man beispielsweise einen Diskussionsbeitrag: „Das war ein guter Punkt, jedoch bin ich nicht der Meinung, dass…“, so könnte man diesen Beitrag in zwei Segmente mit verschiedenen Klassen unterteilen: in ein Segment, „Das war ein guter Punkt“, das man als motivierend klassifizieren könnte und in ein zweites diskursspezifisches klassifizierbares Segment „jedoch bin ich nicht der Meinung, dass…“.

Nach der Annotation kann man mit Hilfe der annotierten Klassen Aussagen über die ablaufenden Gruppenprozesse oder über den Effekt von pädagogischen Interventionen auf das Gruppenverhalten treffen [We06]. Ist eine automatische Segmentierung und Klassifikation in Echtzeit möglich, so kann man über Visualisierungen oder Instruktionen auch kontinuierliches Feedback an die Lernenden geben (Learning Analytics).

Einen Diskussionskorpus manuell zu annotieren ist jedoch sehr aufwendig. Jeder einzelne Beitrag der Diskussion muss, um die Objektivität der Annotation zu gewährleisten, von mindestens zwei Ratern gelesen, segmentiert und dann klassifiziert werden. Streitfälle müssen diskutiert und aufgelöst werden, bis beide Rater eine substanzielle Übereinstimmung erreichen.

Eine computergestützte automatische Annotation des Korpus wäre vorteilhaft, um eine schnelle Annotation von großen Korpora zu ermöglichen oder Korpora fortlaufend während des Kurses zu annotieren.

Aus einer technischen Perspektive handelt es sich hierbei um ein Segmentierungs- beziehungsweise Klassifizierungsproblem. Beiden Problemklassen ist gemein, dass Lösungsansätze typischerweise auf überwachtem maschinellem Lernen aufbauen, also einen manuell vorannotierten Korpus nutzen, um die statistischen Regelmäßigkeiten zu lernen, die charakteristisch für eine Klasse oder ein Segment sind. Anschließend lässt sich das Wissen über die so gelernten Regelmäßigkeiten nutzen, um Datensätze mit ähnlicher Struktur zu segmentieren oder klassifizieren. Die statistischen Muster werden dabei bei Texten in der Regel auf Basis des Vokabulars gelernt (z.B. Bag-of-Words Modelle) [Go17]. Daraus folgt, dass die Repräsentation des Vokabulars ausschlaggebend für den Algorithmus ist. Ändert sich das Vokabular, zum Beispiel, weil man anstelle eines Themas aus der Physik ein Thema aus der Biologie diskutiert, so kommen im Diskurs viele Wörter vor, die nicht in dem Trainingskorpus waren und dass verschlechtert die Klassifizierungsergebnisse des Algorithmus [Go17].

Diese Arbeit leistet nun einen Beitrag zur Lösung der oben genannten Probleme, indem wir einen neuen Klassifizierungsansatz für bereits vorsegmentierte, englische Textpassagen vorstellen. Wir arbeiten mit vier verschiedenen Klassen (Tabelle 1): Off-Topic, Organisation, Motivation und Diskursbeitrag. Alle vier Klassen sind pädagogisch relevant. So ist eine Unterscheidung zwischen Off-Topic und Diskursbeiträgen wichtig, um als Tutor schnell zwischen Diskursbeiträgen, die meine Aufmerksamkeit erfordern, und anderen Beiträgen zu unterscheiden. Gleichzeitig gibt es Evidenz dafür, dass sowohl Motivation als auch Selbstorganisation einer Gruppe wichtige Indikatoren für die Effektivität der Gruppe darstellen [Ro18] [Xi11].

Klasse	Beispielsegment aus Korpus
Diskursbeitrag	In my view metacognition, that is, being able to understand, analyze, and control one's cognitive processes, is important in any kind of learning from individual (solo) learning to collaborative and CSCL. Thus, metacognitive knowledge and skills are necessary for effective learning.
Off-Topic	"A problem shared is a problem halved, a joy that's shared is a joy made double.", very nice quote indeed. I would like to express another personal quote like saying, "A problem shared is a problem doubled" (p.s. this is of course personal view in the example of some countries in the modern world, some countries share their problems and as a result the problem in some cases is doubled not solved :))
Organisation	What if we start simply by brainstorming: why is metacognition important or unimportant in CSCL?
Motivation	you are awesome! danke :)

Tab 1: Beispiele für die vier unterschiedenen Textklassen aus dem verwendeten Korpus

Nachfolgend präsentieren wir ein Verfahren, das folgende Beiträge zum bestehenden Stand der Forschung leistet:

- Das Verfahren kann vier pädagogisch wichtige Klassen mit einer für empirische Studien genügenden Güte (Cohens $\kappa > 0.7$) voneinander abgrenzen.
- Das Verfahren verbessert die Klassifizierungsgüte im Vergleich zu in verwandten Arbeiten verwendeten Standardverfahren gemessen an F_1 Maß sowie Cohens κ.
- Das Verfahren benutzt eine vorab trainierte Vektorrepräsentation von Wörtern, sogenannte Word Embeddings, um keine direkten Abhängigkeiten zum Trainingsvokabular aufzubauen. Aufgrund dessen vermuten wir eine bessere Generalisierbarkeit über Diskursthemengrenzen hinweg. Wir präsentieren erste Evaluationsergebnisse, die auf den Erfolg solch einer themenübergreifende Generalisierung hindeuten.

2 Verwandte Arbeiten

In den letzten Jahren wurden verschiedene Verfahren zum automatischen Klassifizieren von Forumsdiskussionen präsentiert. Je nach Zielsetzung arbeiten die Verfahren dabei mit verschiedenen Klassen und klassifizieren Segmente, Beiträge oder Threads. Nachfolgend beschreiben wir prototypisch einige Verfahren und unterscheiden im Folgenden zunächst

danach, ob das Verfahren ohne weiteres Training auf andere Korpora generalisiert oder nicht.

Ein Verfahren, das auf ganzen Forenposts arbeitet und ohne Trainingskorpus auskommt, wird von Ezen-Can et al. vorgestellt [Ez15]. Die Autoren arbeiten mit 550 Posts von 155 Lernenden in einem MOOC und nutzen den k-medoids Clustering Algorithmus, um Posts nach Dialogakten zu gruppieren. Anschließend verwenden sie Latent Dirichlet Allocation, um die Cluster mit Stichwörtern zu beschriften. Während das Verfahren aufgrund seines trainingslosen Clustering-Algorithmus leicht zu generalisieren ist, unterscheidet es sich von dem hier vorgestellten Verfahren vor allem in zwei Aspekten. Erstens, die Autoren können nicht a priori festlegen, welche Gemeinsamkeiten die gefundenen Gruppen haben, denn dies hängt vom Clustering auf dem konkreten Korpus ab. Zweitens arbeitet das Verfahren auf ganzen Posts und nicht auf Segmenten.

Ein Korpus unabhängiges Verfahren, das nicht auf Segment- oder Postebene arbeitet, sondern ganze Threads klassifiziert, wurde von Rossi und Gnawali entwickelt [Ro14]. Die Autoren nutzen zwar überwachte Klassifikation, achten bei ihren Eingabedaten aber darauf, dass diese nicht von der Textsprache abhängen. Stattdessen werden quantitative Werte wie beispielsweise die Anzahl der Wörter oder das soziale Netzwerk der im Thread Beteiligten herangezogen. Die Autoren klassifizieren sechs Klassen (General Discussion, Assignments, Meetups, Lectures Logistics und Feedback). Dabei erreichen sie je nach Klasse ROC AUC Werte zwischen 0.58 und 0.89. Das Verfahren unterscheidet sich vom hier präsentierten vor allem darin, dass komplett auf Informationen aus dem Vokabular verzichtet wird. Weiterhin arbeitet es auf ganzen Threads anstelle von Segmenten.

Weiterhin gibt es Verfahren, die vom Vokabular des Trainingskorpus abhängig sind. Wise et al. haben eine automatische Klassifikation von MOOC Threads vorgestellt, die Off-Topic von themenspezifischen Threads abgrenzt [Wi16]. Dabei benutzen die Autoren Bag-of-Words Features mit Unigrammen sowie Bigrammen und eine Support Vector Machine (SVM) als Klassifikationsverfahren. Die Autoren erreichen auf ihren zwei Klassen eine Treffgenauigkeit von 0.86 und ein Cohens κ von 0.64 und können somit drei-Viertel aller Threads in die korrekte Klasse einteilen. Die Unterschiede zu diesem Artikel liegen darin, dass die Autoren ihr Verfahren direkt an das Vokabular der Threads koppeln und dass die Autoren nur zwei Klassen unterscheiden.

Die Arbeit von Rosé et al. ist Korpus abhängig und arbeitet auf Segmenten [Ro08]. Die Zielsetzung der Autoren war es herauszufinden, welche Ansätze vielversprechend für die komplette Automatisierung eines Kodierungshandbuches sind. Dabei beschäftigt sich die Arbeit zu einem Großteil mit der Klassifizierung in den sieben Dimensionen des Kodierungshandbuches von Weinberger und Fischer [We06]. Die Autoren benutzen eine Vielzahl von typischen Features wie Unigramme, Bigramme oder Satzzeichen aber konstruieren auch neue, für Segmente wichtige Features wie die Ähnlichkeit zu den Vorgängersegmenten, die Posttiefe oder die Klassen der Vorgängersegmente. Eine Evaluation dieser neuen Features zeigt, dass diese die Klassifizierungsgüte signifikant verbessern. Außerdem testen die Autoren verschiedene Klassifizierungsalgorithmen auf ihren

Daten, unter anderem Algorithmen, die explizit auf die sequenzielle Struktur von Segmenten ausgelegt sind. Die Ergebnisse zeigen jedoch, dass die Wahl des Klassifizierungsalgorithmus nicht ausschlaggebend für die Klassifizierungsgüte ist und es keine signifikanten Unterschiede zwischen einem SVM Ansatz und spezialisierten Algorithmen gibt. Die Arbeit unterscheidet sich von der hier vorgestellten durch ihren Fokus auf ein komplettes Kodierungshandbuch und durch ihre direkte Abhängigkeit vom Vokabular über die Unigramm und Bigramm Features.

3 Begründung und Beschreibung unseres Ansatzes

Das konkrete Klassifizierungsproblem befindet sich im folgenden Kontext. Gegeben ist ein mit vier Klassen vorannotierter, englisch sprachiger Korpus auf Segmentebene. Gesucht ist ein Verfahren, das es ermöglicht, einem nicht im Korpus enthaltenem Segment die richtige Klasse zuzuweisen. Dabei soll es möglichst gut generalisieren, also unabhängig von der konkreten Wortwahl des Korpus sein.

Dabei ist der hier verwendete Ansatz nicht unüberwacht, wie beispielsweise das Verfahren von Ezen-Can et al., da die Klassen a priori definiert sind und unbekannte Segmente in derselben Art und Weise annotiert werden sollen, wie bereits bekannte. Außerdem verwendet der hier gezeigte Ansatz sprachabhängige Features, im Gegensatz zu Rossi und Gnawali, da davon auszugehen ist, dass das Vokabular für die Klassen im Korpus wichtige Information beinhaltet. Grundlegend stützt sich unser Klassifizierungsalgorithmus auf den Einsatz von vortrainierten Word Embeddings zusammen mit einer SVM (RBF Kernel). Sowohl Rose et al. als auch Wise et al. haben gezeigt das eine SVM auf ähnlichen Klassifizierungsproblemen gute Ergebnisse mit Bag-of-Words Features erzielen kann. Es sind allerdings verschiedene Schwächen der Bag-of-Words Ansätze bekannt:

Die meisten Klassifikationsalgorithmen, so wie auch SVMs, erreichen bessere Ergebnisse, wenn die Anzahl der Features deutlich kleiner als die Anzahl der Trainingsdaten ist [Ha09]. Bei Bag-of-Words-Ansätzen wird jedes Wort im Vokabular als One-Hot kodierter Vektor φ dargestellt. Der Bag-of-Words (BOW) eines Textsegmentes *S* ist dann eine Linearkombination dieser Wortvektoren wobei man $a_{\mathrm{i}} = 1$ im einfachsten Falle wählt:

$$BOW_{segment} = \sum_{\varphi_i \epsilon S} a_i \varphi_i \tag{1}$$

Dabei wächst die Dimension des Feature-Vektorraumes linear mit der Anzahl an unterschiedlichen Wörtern im Vokabular. Dies führt bei größeren Texten oftmals zu einer sehr großen Anzahl Features und dies beeinflusst die Güte der Klassifikation nachteilig. Deshalb versucht man in diesen Ansätzen oftmals das Vokabular zu beschränken. Dazu filtert man zum Beispiel alle Wörter, die sehr selten oder sehr häufig in einem Text vorkommen heraus, um nur noch für das Trainingsset aussagekräftige Wörter zu behalten [Go17]. Dies verschlechtert allerdings dann wieder die Generalisierbarkeit des Klassifikationsalgorithmus.

Ein weiterer Nachteil von Bag-of-Words ist, dass Wörter, die eigentlich semantisch ähnlich sind, in ihrer Vektorrepräsentation orthogonal zueinanderstehen. Hat man beispielsweise die Wörter "jedoch" und "wohingegen" besitzen die Vektoren den gleichen Abstand wie „jedoch" und „Hund". Der Klassifikationsalgorithmus kann deshalb die semantische Ähnlichkeit der Eingaben nicht ausnutzen. Verwendet man Word Embeddings anstelle von Bag-of-Words werden diese Nachteile abgemildert:

Die Technik der Word Embeddings wurde ursprünglich im Zusammenhang mit neuronalen Netzen entwickelt und erlaubt, wie auch die Bag-Of-Words Ansätze, Wörter als Vektoren zu repräsentieren und als Features in Klassifizierungsalgorithmen zu verwenden [Co11]. Die Konstruktionsvorschrift der Embeddings betrachtet jedoch nicht jedes Wort als eine separate Dimension, sondern das Wort in seinem Kontext [Co11]. Bei der Konstruktion wird eine fixe Dimension für die Embeddingvektoren festgelegt und die einzelnen Wörter werden dann in diese Dimension eingebettet [Co11]. Wählt man zum Beispiel eine Dimension von 300, wird für jedes Eingabewort ein Vektor der Dimension 300 berechnet. Durch die Art, wie während der Einbettung der Kontext einbezogen wird, liegen Vektoren von Wörtern, die semantisch ähnlich sind, dicht beieinander.

Um nun einen Text *S* mit *k* Wörtern mithilfe dieser Embeddingvektoren zu repräsentieren, berechnen wir einen Continuous-Bag-of-Words (CBOW) [Go17] erneut als Linearkombination der Embeddingvektoren, wobei *f* eine Abbildung ist, die jedem Wort *w* des vortrainierten Vokabulars seinen n-dimensionalen Embeddingvektor zuordnet, im einfachsten Fall mit $a_i = 1$:

$$\mathrm{CBOW}_{\mathrm{segment}} = \frac{1}{k} \sum_{w \in S} a_i f(w) \tag{2}$$

Der Continuous-Bag-of-Words adressiert nun die zwei oben genannten Probleme des Bag-of-Words. Zunächst wächst die Dimension des Continous-Bag-of-Words nicht mehr linear mit der Anzahl der unterschiedlichen Wörter im Vokabular, sondern ist konstant, was vorteilhaft für den Klassifikationsalgorithmus ist. Außerdem besitzen ähnliche Wörter, aufgrund der Konstruktion, ähnliche Vektoren. Dadurch liegen Texte aus semantisch ähnlichen Wörtern in ihrer Vektorrepräsentation näher beieinander, was die Klassifikation erleichtert. So haben Collobert et al. gezeigt, dass Word Embeddings aufgrund dieser Eigenschaft eine Vielzahl von Sprachverarbeitungstasks verbessern [Co11].

Schließlich ist die Nutzung der Embeddings noch mit einem weiteren wichtigen Vorteil verbunden. Es wurde gezeigt, dass wenn die Embeddings einmal auf einem ausreichend großen Korpus trainiert wurden, ihre semantische Ähnlichkeit auf andere Korpora und somit andere Klassifikationsaufgaben übertragbar ist [Co11]. Durch die Nutzung vortrainierter Embeddings (GloVe Common Crawl 300 [Pe14]), die auf einem sehr großen Kor-

pus trainiert wurden, können deshalb auch Wörter, die nicht im Trainingskorpus vorkommen, später als aussagekräftiges Feature genutzt werden, denn es existieren dafür ebenfalls Embeddingvektoren.

Neben den Embeddings nutzt der hier beschriebene Ansatz einige weitere Features, um den Kontext, in dem die einzelnen Segmente geäußert wurden, zu modellieren. Für jedes Segment werden zunächst einige einfache Kennzahlen betrachtet: Die Posttiefe, die Anzahl der Fragezeichen sowie die Anzahl der direkten Zitate innerhalb des Segmentes. Außerdem werden temporale Äußerungen mithilfe des Temporal Tagger Heideltime [St12] annotiert und gezählt, um die Anzahl der explizit genannten Zeitangaben herauszufinden, zur besseren Klassifikation von Organisationsbeiträgen.

Weiterhin kommen, ähnlich zu Rose et al. Sequenzfeatures zum Einsatz. Dazu werden die Kosinus-Ähnlichkeiten des Continuous-Bag-of-Words des aktuellen Segmentes mit denen der vorherigen verglichen, um eine Änderung des Diskursverlaufs zu detektieren. Außerdem wird das Sentiment des aktuellen Segments, sowie der vorherigen, berechnet, um motivationale Abschnitte besser zu detektieren. Abschließend werden ebenfalls die Klassen der vorherigen Segmente miteinbezogen, da es gut möglich ist, dass es typische Abfolgen von Klassen gibt (z.B. Motivation immer vor Diskurs). Dazu wird Stanford CoreNLP verwendet [Ma14]. Insgesamt erhalten wir so die in Tabelle 2 gezeigten Features. Die genutzte SVM wird mit dem Pythonframework scikit-learn und einem RBF Kernel trainiert. Dabei skalieren wir γ automatisch und gewichten die Kostenfunktion mit der Klassenhäufigkeit.

	Wortrepräsentation	**Postkennzahlen**	**Sequenzrepräsentation (jeweils letzte 15)**
Art	GloVe-Common-Crawl Embeddings	Posttiefe, Frageanzahl, Zitatanzahl, Zeitangaben, aktuelles Sentiment	vorherige Sentimente, Klassen, Kosinus Ähnlichkeit
Dimension	300	5	45

Tab. 2: Klassifikationsmerkmale mir ihrer Dimension. Aufteilung in sprachliche Features, segmentspezifische Features und sequenzspezifische Features

4 Korpus

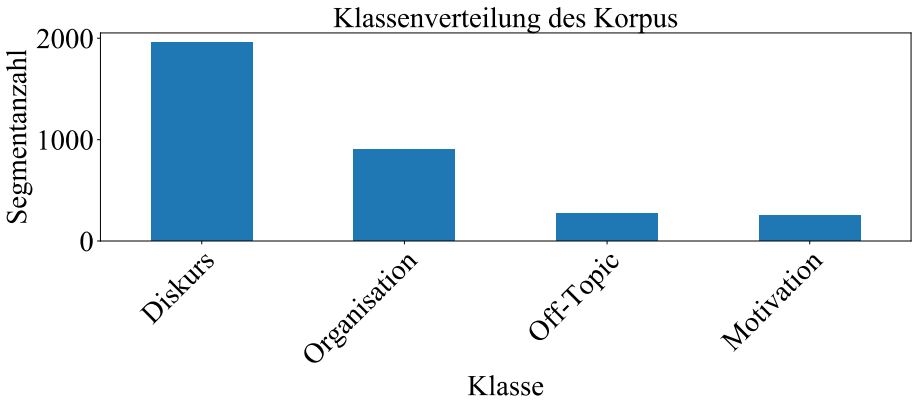

Abb. 1: Verteilung der einzelnen Klassen im Evaluationskorpus

Um unser Klassifikationsverfahren zu evaluieren, nutzen wir einen vorannotierten Korpus. Dieser Korpus wurde während zwei Durchläufen eines sechswöchigen online Universitätskurses dreier Universitäten in den Jahren 2015 und 2016 erstellt. Die Arbeitssprache des Kurses war Englisch und die Studierenden diskutieren jeweils in Kleingruppen aus drei bis fünf Mitgliedern online auf Facebook. Dabei wurde eines von drei Themen jeweils für vier Wochen von den Teilnehmern diskutiert: „CSCL Skripte“, „Motivation und Emotion im Lernen“ sowie „Metakognition“. Die quantitative und qualitative Teilnahme an der Diskussion wurde benotet.

Insgesamt besteht der Korpus aus 3246 einzelnen Segmenten, die von zwei Annotatoren nach einem Kodierungshandbuch annotiert wurden. Die Interrater-Reliabilität betrug dabei Cohens $\kappa = 0.85$. Für die Evaluation wurden die Annotationen auf die vier Klassen Off-Topic, Organisation, Motivation und Diskurs zusammengefasst. Dabei zeigt sich (Abbildung 1) eine starke Ungleichverteilung der Klassen, wobei Beiträge zum inhaltlichen Diskurs dominieren. Nachfolgend wird deshalb als Evaluationsmetrik neben Cohens κ ebenso das Makro F1 Maß sowie Precision und Recall angegeben, um die Ungleichverteilung der Klassen zu berücksichtigen.

5 Evaluation

Zu evaluieren sind drei Fragestellungen. Erstens: Ist es mit unserem Verfahren möglich, die vier gegebenen Klassen ausreichend gut zu klassifizieren? Wir gehen in Übereinstimmung mit Rose et al. davon aus, dass wir dieses Ziel erreichen, wenn wir ein Cohens $\kappa >$

0.7 erreichen, welches auch in empirischen Studien als beachtliche Übereinstimmung angesehen wird.

Zweitens: Schneidet unser Verfahren besser auf den Daten ab, als die Standardverfahren? Wir testen dazu ob normale Bag-of-Words Verfahren auf den Daten ein niedrigeres F1-Maß erreichen, als das Continuous-Bag-of-Words Verfahren. Dabei vergleichen wir mit einem Bag-of-Words der das gesamte Vokabular benutzt und einem Bag-of-Words, der das TF-IDF Maß nutzt [Go17], um hoch- und niederfrequente Wörter herauszufiltern. Dies ist ein typischer Standardansatz, um das oben skizzierte Problem der hohen Dimensionalität des Bag-of-Words zu lösen.

Drittens: Generalisiert unser Verfahren auf unbekannten Themen besser als Bag-of-Words Ansätze? Wir zerteilen unseren Korpus bei der Evaluation immer entlang der Themengrenzen des Kurses und trainieren mit zwei der drei Themen. Dabei ist die Anzahl der Segmente pro Thema ungefähr gleich groß. Das dritte Thema wird dann als Testmenge genutzt. Sofern unser Verfahren gut generalisiert, sollten auf allen Testmengen die gleichen F1 Maße herauskommen, denn es sollte egal sein, ob der Algorithmus mit Themen (1,2) trainiert wird und auf 3 generalisiert oder mit Themen (2,3) trainiert wird und auf 1 generalisiert. Gleichzeitig sollten die Bag-of-Words Ansätze eine größere Streuung der F1 Maße aufweisen, da sich das Vokabular in den einzelnen Themen unterscheidet und so die Generalisierbarkeit erschwert.

Abbildung 2 zeigt die Evaluationsergebnisse, wobei der Korpus entlang der Themengrenzen geteilt wurde, auf zwei Themen trainiert und auf dem verbleibendem evaluiert wurde. Da Off-Topic Beiträge ungleichmäßig viel in der ersten Hälfte des Kurses vorkommen, haben wir diese gleichmäßig auf die Trainings- und Testmenge verteilt, um zu gewährleisten, dass die einzelnen Klassen in jeder Menge ungefähr gleichermaßen verteilt sind. Bei den Ergebnissen wurde jeweils die Wortrepräsentationsfeatures getauscht, die Postkennzahlen sowie die Sequenzkennzahlen blieben gleich.

Der ungefilterte und ungewichtete Bag-of-Words Modell erreicht ein Makro F1-Maß zwischen 0.59 und 0.65 (Precision: 0.68-0.71 Recall: 0.62-0.66) je nach Datenaufteilung. Dabei nutzt er das gesamte Vokabular der Trainingsdaten, im Schnitt um die 6000 Worte bei jeweils ca. 2000 Trainingsdaten. Das TF-IDF Modell mit Häufigkeitsfilter (0.05 bis 0.95) erreicht Makro F1-Maße zwischen 0.67 und 0.72 (Precision: 0.72-0.74 Recall: 0.64-0.69) wobei nach dem Filter 120 Features verbleiben. Das Continuous-Bag-of-Words Modell erreicht ein konstantes Makro F1-Maß von 0.75 (Precision: 0.78-0.81 Recall: 0.72-0.77) bei einer Embedding-Dimension von 300.

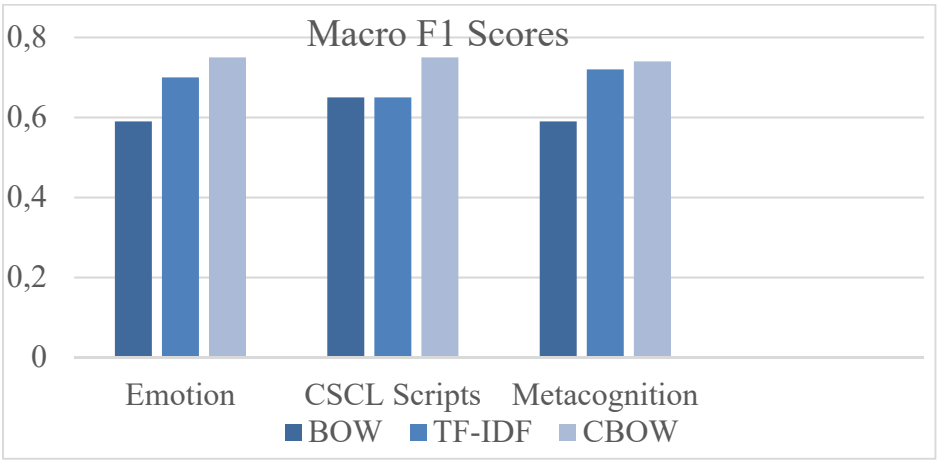

Abb. 2: Evaluationsergebnisse: die Beschriftung einer Balkengruppe bezeichnet die genutzte Testmenge. Die Bag-of-Words SVM beinhaltet das gesamte Vokabular (ca. 6000 Wörter), die TF-IDF SVM nutzt einen Wort Häufigkeitsfilter von min. 0.05 und max. 0.95 (ca. 120 Wörter) und der Continuous -Bag-of-Words SVM nutzt 300-dimensionale GloVe Embeddings.

6 Diskussion und Ausblick

Bezüglich der zu evaluierenden Fragestellungen lässt sich folgendes feststellen: Die erste Fragestellung, ob das Verfahren es schafft, die vier Klassen mit ausreichender Güte auseinanderzuhalten, können wir positiv beantworten. Wir haben mit $F1 > 0.7$ und Cohens $\kappa > 0.7$ eine Klassifizierungsgüte erreicht, die den sozialwissenschaftlichen Standards gerecht wird. Zunächst heißt dies, dass es dieses Verfahren einen guten Ausgangspunkt bildet, um Segmentierungsalgorithmen für eine vollautomatische Annotation zu entwickeln, da die prinzipielle automatische Unterscheidbarkeit der Klassen gegeben ist. Im Idealfall bedeutet dies auch, dass man den Klassifizierungsalgorithmus als zweiten Rater eines Korpus betrachten kann und somit nur noch mit einem Rater manuell, statt mit zweien annotieren muss. Diese Idee sollte jedoch in weiteren Arbeiten überprüft werden, denn im Gegensatz zu menschlichen Ratern bringt ein voll automatisierter Algorithmus kein wirkliches semantisches Verständnis über die Daten mit, sondern arbeitet auf den gelernten Mustern, was eine starke Vereinfachung der Semantik darstellt. Dadurch kann sowohl die Reliabilität als auch die Validität des Kodierungsschemas beeinträchtigt werden, was zumindest auf einem Teil der automatisch annotierten Daten wiederum zu testen ist [Ro08].

Die zweite Fragestellung, ob sich ein auf Embeddings basierendes Verfahren auf dem gegebenen Korpus besser schlägt als typische Referenzverfahren aus der Literatur, lässt sich ebenfalls positiv beantworten. Dabei schneidet das auf Embeddings basierende Verfahren in allen Evaluationsszenarien besser ab. Außerdem erreicht nur das Embedding basierte

Verfahren eine sozialwissenschaftlich ausreichende Klassifikationsgüte in allen Aufteilungen. Dabei schneiden die Embeddings auch besser ab als die TF-IDF gewichteten und gefilterten Bag-of-Words Vektoren, obwohl die Dimension des Eingaberaumes, dort geringer war. Wir vermuten, dass dies damit zusammenhängt, dass die TF-IDF Repräsentation hier für die Klassifikation nützliche Wörter verwirft, da diese zu selten oder häufig vorkommen. Um solche Fehler zu vermeiden, kann man komplexere Feature-Engineering Methoden anwenden, und beispielsweise nur die Wörter als Feature verwenden, die auch eine gute Korrelation mit den Ausgangsklassen haben. Dabei besteht immer der Nachteil, dass solche Verfahren die Features abhängig vom Korpus auswählen, was die Generalisierbarkeit verschlechtert. Ein auf vortrainierten Embeddings basierender Ansatz benötigt hingegen überhaupt kein Feature-Engineering und bildet ein großes Vokabular ab. Darum zeigen unsere Ergebnisse, dass vortrainierte Embeddings auch in einer spezialisierten Domäne wie E-Learning eine nützliche Feature Darstellung sein können, die ohne großen Aufwand gute Ergebnisse liefert.

Bezüglich der dritten Fragestellung gibt es Evidenz, dass unser Embedding basiertes Verfahren besser generalisiert als die Bag-of-Words Verfahren. Dabei schwankt das F1 Maß hier bei den verschiedenen Trainings und Testmengen kaum, wohingegen sowohl bei TF-IDF als auch beim klassischen Bag-of-Words mindestens einer der drei Evaluationsdatenpunkte in der Klassifikationsgüte abweicht. Diese Ergebnisse deuten auf eine bessere Generalisierbarkeit durch die Continuous-Bag-of-Words Features hin. Für endgültige Aussagen liegen jedoch zu wenige unterschiedliche Themen im Evaluationskorpus vor. Außerdem unterscheiden sich Trainings- und Testmenge zwar bezüglich ihrer Themen, jedoch kommen die Themen immer noch aus den gleichen Wissenschaftsgebieten. Die Diskussionskultur unterscheidet sich jedoch beachtlich in verschiedenen Wissenschaftsfeldern und es ist nicht klar, ob unsere Ergebnisse, auch auf weiter auseinanderliegende Themen, wie beispielsweise Diskussion über biologische Themen, generalisieren. Weitergehende Arbeiten sollten deshalb mit einem zweiten annotierten Korpus arbeiten, der mindestens ein Thema aus einer anderen Wissenschaftsrichtung behandelt. So könnte man testen, ob die vier Klassen wirklich komplett unabhängig vom diskutierten Themengebiet sind.

Ein wichtiger Anknüpfpunkt für weitere Arbeiten ist die Segmentierung. Zwar kann das hier präsentierte Verfahren die Klassifikation vornehmen, doch benötigt es dazu fertige Segmente. Da ein menschlicher Rater in der Regel beim Segmentieren gleichzeitig klassifiziert, spart das Verfahren im Falle eines einzigen Raters keinen Aufwand. Deshalb ist die Segmentierung der nächste Schritt hin zu einer vollautomatischen Lösung. Außerdem wichtig wäre eine feinere Klassifizierung der einzelnen Segmente wie sie auch in den Kodierungshandbüchern vorkommen, um eine genauere Analyse des Diskurses zu ermöglichen. Dazu muss zunächst evaluiert werden, inwieweit sich die feineren Klassen überhaupt automatisch unterscheiden lassen.

Danksagung Wir möchten uns beim Lehrstuhl für Bildungstechnologie und Wissensmanagement der Universität des Saarlandes sowie Eirini Papanastasiu für die Zusammenstellung und Zurverfügungstellung des Korpus bedanken.

Literaturverzeichnis

[Co11] Collobert, R.; Weston, J., Leon; B., Karlen; M., Kavukcuoglu, K.; & Kuksa, P.: Natural Language Processing (Almost) from Scratch. The Journal of Machine Learning Research, 12, S. 2493-2537, 2011.

[Ez15] Ezen-Can, A.; Boyer, K. E.; Kellogg, S.; Booth, S.: Unsupervised modeling for understanding MOOC discussion forums: A learning analytics approach. In Proc. of LAK. S. 146–150, 2015.

[Go17] Goldberg, Y.: Neural Network Methods for Natural Language Processing. 2017.

[Ha09] Hastie, T.; Tibshirani, R.; Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2. Auflage, Springer Series in Statistics, 2008.

[Ma14] Manning, C. D.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, S. J.; McClosky, D.: The Stanford CoreNLP Natural Language Processing Toolkit. In Proc. of Annual Meeting of the Association for Computer Linguistics: System Demonstrations, S. 55–60, 2014.

[Pe14] Pennington, J.; Socher, R.; Manning, C. D.: GloVe: Global Vectors for Word Representation. In Proc. Of EMNLP. S. 1532–1543, 2014.

[Ro08] Rosé, C.; Wang, Y. C.; Cui, Y.; Arguello, J.; Stegmann, K.; Weinberger, A.; Fischer, F.: Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. International Journal of Computer-Supported Collaborative Learning 3/3, S. 237–271, 2008.

[Ro14] Rossi, L. A.; Gnawali, O.: Language independent analysis and classification of discussion threads in Coursera MOOC forums. In Proc. of IEEE IRI. S. 654–661, 2014.

[Ro18] Romero, M.; Lambropoulos, N.: Internal and External Regulation to Support Knowledge Construction and Convergence in Computer Supported Collaborative Learning (CSCL). Electronic Journal of Research in Education Psychology 9/23, S. 309–330, 2018.

[St12] Strötgen, J.; Gertz, M.: Temporal Tagging on Different Domains: Challenges, Strategies, and Gold Standards. In Proc. of LREC, S. 3746–3753, 2012.

[We06] Weinberger, A.; Fischer, F.: A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. Computers and Education, 46/1, S. 71–95, 2006.

[Wi16] Wise, A. F.; Cui, Y.; Vytasek, J.: Bringing order to chaos in MOOC discussion forums with content-related thread identification. In Proc. of LAK. S. 188–197, 2016.

[Xi11] Xie, K.; Ke, F.: The role of students' motivation in peer-moderated asynchronous online discussions. British Journal of Educational Technology 42/06, S. 916–930, 2011.