Automatische Verfahren zur Bewertung der Relevanz von Dokumenten für geisteswissenschaftliche Forschungsfragen

André Blessing
Universität Stuttgart
andre.blessing@ims.uni-stuttgart.de

Melanie Dick
Universität Hildesheim
melaniedick@gmx.net

Ulrich Heid
Universität Hildesheim
heid@uni-hildesheim.de

Abstract

In vielen Projekten der Digital Humanities werden große Textmengen im Hinblick auf eine Forschungsfrage ausgewertet. Das interdisziplinäre Projekt *eIdentity* (BMBF, FKZ. 01UG1234) widmet sich beispielsweise der Frage nach multiplen kollektiven Identitäten in internationalen Debatten um Krieg und Frieden seit dem Ende des Kalten Krieges. Damit sprachtechnologische Werkzeuge überhaupt auf das dort verwendete mehrsprachige Zeitungskorpus angewendet werden können, muss dieses zunächst von nicht für die Forschungsfrage relevanten Artikeln („off-topic-Artikel“) bereinigt werden. Nur so kann sichergestellt werden, dass nur Texte in die Auswertung einfließen, die Gegenstand der Forschungsfrage sind.

Viele Digital Humanities-Studien verwenden zur off-topic-Filterung lediglich Metadaten, wie zum Beispiel den Namen der Quelle, das Veröffentlichungsdatum oder den Autor. Für die Bereinigung des eIdentity-Korpus genügen diese Informationen aber nicht: es muss zusätzlich eine inhaltliche Filterung vorgenommen werden. Kantner et al. (2011) erstellten dazu manuell Schlagwortlisten um relevante und nicht relevante Artikel zu identifizieren. Die Erstellung solcher Listen ist allerdings zeitaufwendig und in der Praxis stellten sich diese als nicht vollständig heraus. Die Aufgabe der off-topic-Filterung ist ähnlich wie sogenannte „Spam-Filter“ für e-Mail; allerdings kann ein Spam-Filter anhand großer Mengen von Daten trainiert werden, weil der Nutzer in der Regel alle Nachrichten manuell nach Relevanz bewertet. In Digital Humanitites-Projekten ist die Anzahl der zu klassifizierenden Texte dafür zu groß; es braucht also Klassifikationsverfahren, die schon auf kleinen Mengen annotierter Texte gute Ergebnisse liefern.

In unserer Arbeit stellen wir einen neuen Ansatz vor; er geht aus von einer manuell annotierten Grundmenge von für die Forschungsfrage als relevant beziehungsweise irrelevant annotierten Artikeln. Ein Problem bei der Annotation ist die Auswahl der für das Training des Klassifikators nützlichen Artikel. Wenn zufällig ausgewählt wird, kann es sein, dass die Auswahl nicht repräsentativ für die zu klassifizierende Textmenge ist: wenn z.B. die Mehrheit aller Texte für die Forschungsfrage relevant ist, würden zu viele relevante und zu wenig irrelevante Texte annotiert. Hier kann durch die Nutzung von „Topic Modelling“ mit Latent Dirichlet Allocation (LDA; vgl. Blei et al. 2003) sichergestellt werden, dass eine besser nutzbare Auswahl getroffen wird.

Im ersten Schritt, der Feature-Extraktion, werden zunächst Merkmale aus Textdokumenten extrahiert. Diese extrahierten Merkmale werden in einem zweiten Schritt an einen Klassifikator übergeben, welcher die Artikel als relevant oder irrelevant kategorisiert (vgl. Abbildung 1).

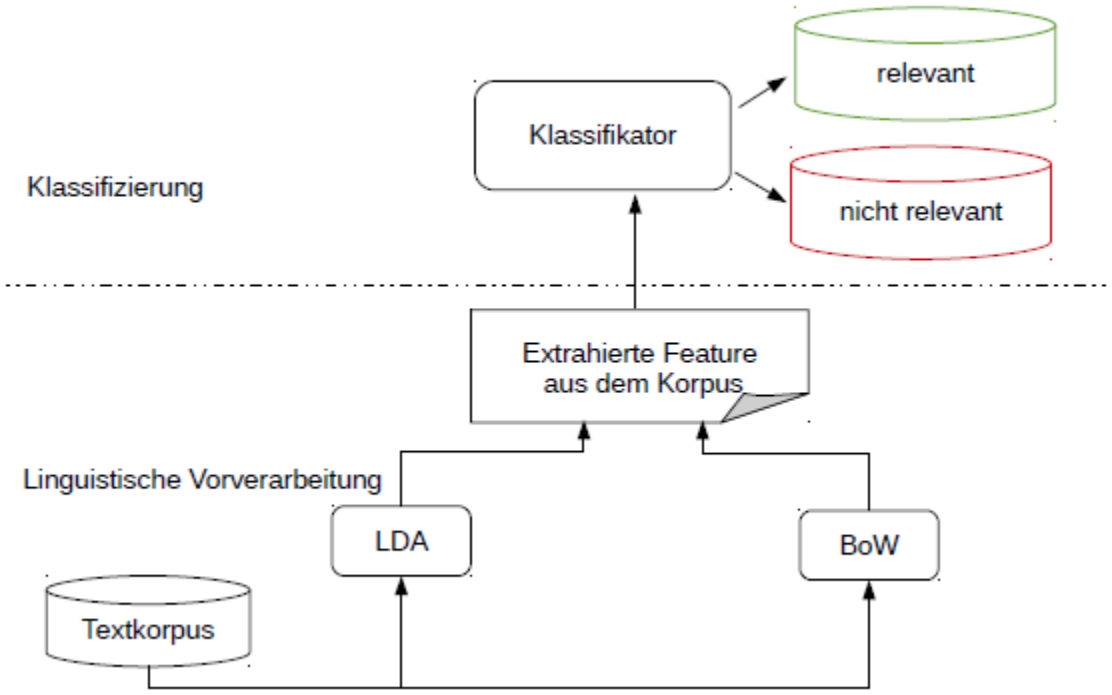

Abbildung 1: Zweistufiges Klassifikator-Modell

Für die Featureextraktion wird in unserem Experiment neben dem gängigen Bag-of-Words (BOW)-Modell, Topic Modelling durch LDA, ein generatives Wahrscheinlichkeitsmodell, eingesetzt und die Ergebnisse werden verglichen. Mit LDA können die Artikel entsprechend den vom System bestimmten „Topics“ vorsortiert werden. Ein „Topic“ soll mittels eines Wortclusters im abstrakten Sinne ein Themengebiet beschreiben. Der Klassifikator kann nun diese Topics explorieren (vgl. Abbildung 2) und anhand ihrer „Schlagwörter“ eine für den Forschungsgegenstand relevantere Auswahl kodieren. In unserem Forschungsprojekt konnten so sehr gut irrelevante Artikel zum Thema Sport, historische Konflikte, Buch- oder Filmkritik aufgespürt und als nicht relevant annotiert werden.

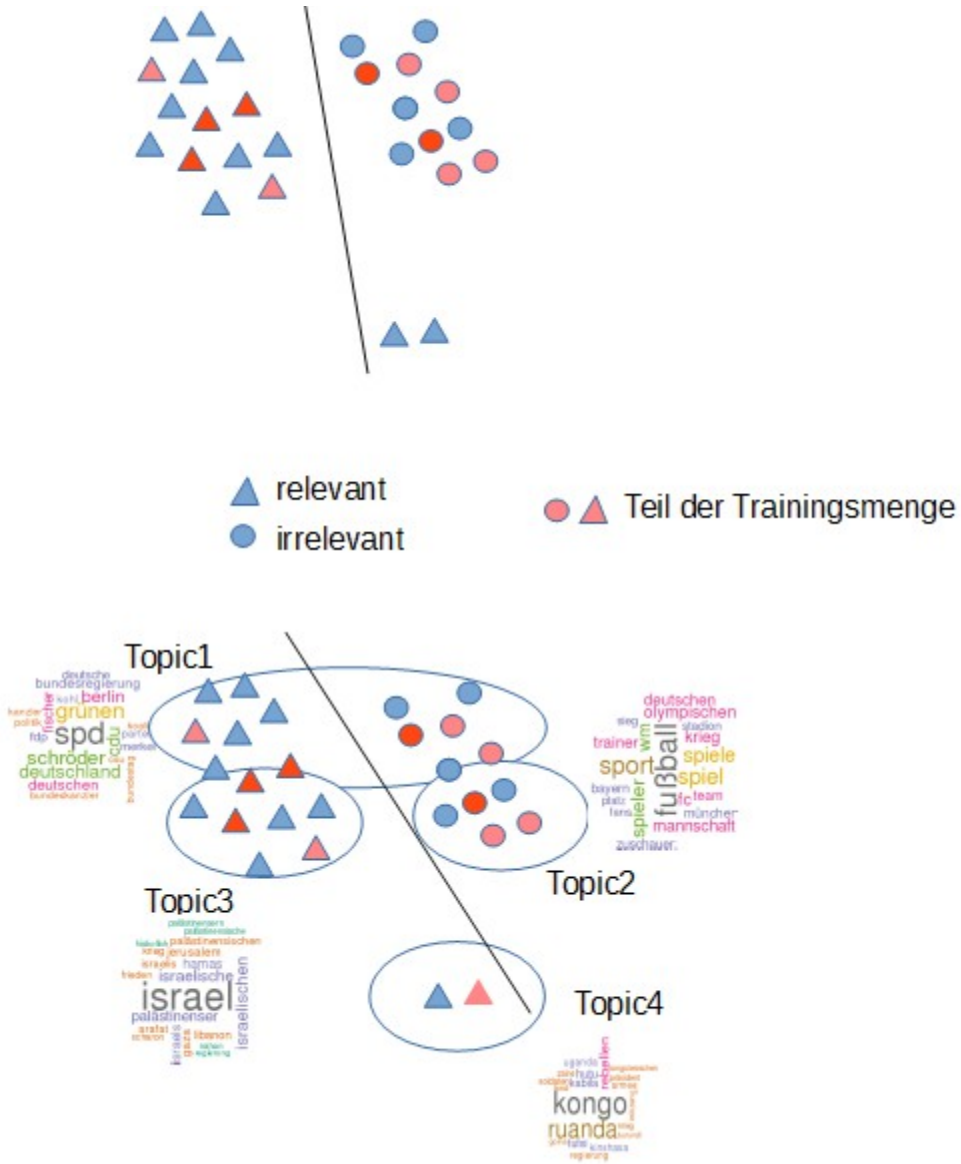

Abbildung 2: Exploration der Topics: oben: Zufallsauswahl; unten: LDA Topic-Modellierung

Im Rahmen der Vorverarbeitung wird zunächst eine Stopwortliste auf das Korpus angewendet. Häufig auftretende Wörter der geschlossenen Wortklassen werden damit als Merkmal ausgeschlossen. Als Baseline wird ein BoW-Modell verwendet, welches die Häufigkeit aller Wörter eines Dokuments als Hauptmerkmal für den Klassifikator aufbereitet. LDA hingegen lässt auf Wortebene für jedes Wort eine anteilige Zuordnung zu mehreren Topics zu. Die Information über die Verteilung über eine zuvor definierte Anzahl von Topics (beispielsweise 150) dient als Grundlage für den Klassifikator. LDA übergibt im Vergleich zum BoW-Modell Informationen an den Klassifikator, welche dieser wesentlich effizienter verarbeiten kann. Im zweiten Schritt folgt die Klassifikation in relevante und irrelevante Artikel (vgl. Abbildung 1).

Erste Ergebnisse haben gezeigt, dass das Topic Modelling für die Textklassifikation geeignet ist. Die manuell bewerteten Artikel (ca. 70 Dokumente – mehrfach bewertet) zeigten Accuracy-Werte von durchschnittlich 0,89. Erste Accuracy-Werte aus der automatischen Klassifikation (ca. 2.500 Dokumente – meist einfach

bewertet) mit LDA sowie BoW in der Vorverarbeitung, bewegen sich auch zwischen 0,8 und 0,9: die automatische Annotation liefert also die gleiche Qualität wie die menschlichen Annotatoren. Der Erfolg der Klassifizierung (Accuracy) beim LDA ist allerdings stark von der Anzahl der manuell ausgewählten Topics abhängig.

Weitere Variationen der beiden Verfahren wie beispielsweise eine Wortselektion mittels Part-of-Speech (POS)-Tagging sowie die Verwendung eines Stemmers, werden jeweils zusammen mit der linguistischen Vorverarbeitung eingesetzt: so kann deren Einfluss auf die Qualität der Ergebnisse des Klassifikators analysiert werden und die bestmögliche Merkmalsextraktion für die Entscheidung über die Artikelrelevanz kann bestimmt werden.

Literatur

David M. Blei, Andrew Y. Ng, Michael I. Jordan (2003): *Latent dirichlet allocation.* IN: The Journal of Machine Learning Research 3, S. 993-1022.

Cathleen Kantner, Amelie Kutter, Andreas Hildebrandt, Mark Püttcher (2011): *How to get rid of the Noise in the Corpus: Cleaning Large Samples of Digital Newspaper Texts, International Relation Online Working Paper.* 2011/2, Juli 2011, Stuttgart: Universität Stuttgart.

eIdentity (2014): *Multiple kollektive Identitäten in internationalen Debatten um Krieg und Frieden seit dem Ende des Kalten Krieges. Sprachtechnologische Werkzeuge und Methoden für die Analyse mehrsprachiger Textmengen in den Sozialwissenschaften (eIdentity).* URL: http://www.uni-stuttgart.de/soz/ib/forschung/Forschungsprojekte/eIdentity.html Stand: 08.10.2014.