Mulit-Label Branchenklassifikation von Web-Texten

Dominik Mottl

Hochschule Darmstadt,
Fachbereiche Mathematik und Naturwissenschaften
& Informatik

Problemstellung

In den Zeiten von *Big Data* mit immer schneller wachsenden Datenmengen gewinnt die Extraktion von Informationen aus Daten immer mehr an Bedeutung. Die Herausforderung dabei ist, dass der Großteil der Daten unstrukturiert in Form von Texten, Audio- und Video-Formaten vorliegt. Das *World Wide Web* stellt eine der wichtigsten öffentlich zugänglichen Informationsquellen der heutigen Zeit dar. Insbesondere in unstrukturierten Texten, wie Nachrichtenartikeln oder Wikipedia-Artikeln, wird Wissen gesammelt und dargestellt. Den unstrukturierten Informationen stehen strukturierte Datenquellen wie *Knowledge Bases* gegenüber, in denen Wissen in strukturierter Form aufbereitet gesammelt wird.

Ziel der Arbeit ist es, aufzuzeigen, wie Informationen aus unstrukturierten und frei zugänglichen Web-Texten mit Hilfe von Methoden aus Knowledge Bases kombiniert werden können, um neues Wissen zu Firmen zu generieren. Diese so neu erlangten Informationen könnten dazu genutzt werden, vorhandenes Wissen zu Firmen in Knowledge Bases zu erweitern und so einen besseren Überblick über die Wirtschaft zu erhalten. Es werden Nachrichtenartikel von Webseiten aus dem *Common-Crawl-Datensatz* extrahiert und mit Wissen aus der Knowledge Base *DBpedia* angereichert. Die so erzeugten Daten beschreiben ein Klassifikationsproblem, in dem Texte, in denen Firmen auftreten, nach Branchen klassifiziert werden. Da Firmen in meheren Branchen gleichzeitig tätig sein und gleichzeitig mehrere Firmen in einem Text auftreten können, handelt es sich um ein Multi-Label Klassifikationsproblem. Dieses soll mit verschiedenen Modellen und Methoden aus dem *Machine Learning*, bzw. dem *Natural Language Processing* gelöst werden. Für die Modelle und Methoden wird Skalierbarkeit gefordert, sodass sie auf eine verteilte Architektur übertragen und so auf einer großen Datenmenge angewendet werden können.

Vorgehen

Das Klassifikationsproblem lässt sich auf mehrere Schritte bzw. Schichten herunterbrechen. Eine grafische Darstellung des gesamten Klassifikationsprozesses dieser Arbeit ist in Abbildung 1 zu sehen. Die erste Schicht ist die Daten-Schicht, in der sich die Rohdaten, die Knowledge Base DBpedia und die Common-Crawl-Daten befinden. Die zweite Schicht ist die Datenaufbereitung. Hier werden zur Vorbereitung der Daten die Tools *Stanford CoreNLP* und *DBpedia Spotlight* eingesetzt, um Entitäten in Nachrichtenartikeln zu erkennen und mit der Knowledge Base DBpedia abzugleichen. Die Daten werden in verschiedenen Datensätzen aufbereitet. Zum einen werden die Daten nach kompletten Nachrichtenartikeln und nach einzelnen Paragraphen aufbereitet, zum anderen wird beim Entity Linking mit einem Konfidenz-Parameter variiert, sodass ein Datensatz mit niedriger und ein Datensatz mit hoher Konfidenz entsteht. Die Klassenstruktur der DBpedia-Branchen wird auf eine Klassenstruktur übertragen, die sich von der Nachrichtenseite *Marketwired* ableitet. Der Kern der Arbeit liegt in der Modellbildungs- und der Auswertungsschicht. Hier wird mit den Daten aus den darüber liegenden Schichten ein Data Mining Prozess durchgeführt.

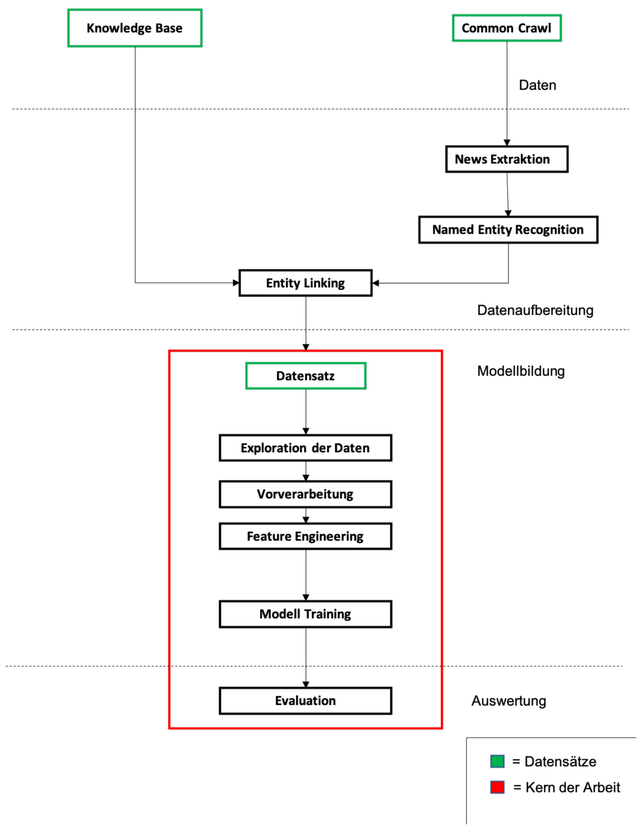

Abbildung 1: Aufbau des Klassifikationsproblems

Die Schicht der Modellbildung ist in Abbildung 2 im Detail dargestellt. Aus den untersuchten und aufbereiteten Daten werden verschiedene Features extrahiert. Zu diesen zählen *Bag-of-Words*-Vektoren, *Tf-idf*-Vektoren und die Darstellung von Texten als Sequenz mehrdimensionaler Vektoren unter Verwendung von vortrainierten *Word Embeddings*, die im Fall der Masterarbeit auf dem Google-News-Dataset trainiert wurden. Außerdem wird das Topic-Modell *Latent Dirichlet Allocation* dazu eingesetzt, neue Features zu generieren, die den Text auf Basis einer reduzierten Feature-Dimensionalität repräsentieren. Auf den extrahierten Features werden zum Lösen des Multi-Label Branchenklassifikationsproblems mehrere Modelle trainiert. Mit dem *Binary-Relevance*-Ansatz werden die Modelle *Naive Bayes*, *logistische Regression* und *lineare Suport Vector Machine* trainiert. Außerdem werden *Entscheidungsbäume* und *Convolutional Neural Networks* eingesetzt. Bewertet werden die Modelle mit Label-basierten und Beispiel-basierten Metriken. Zu den Label-basierten Metriken zählen das *micro-averaging* und das *macro-averaging*. Die verwendeten Beispiel-basierten Metriken sind die *Exact Match Ratio*, *Hamming-Loss*-Metrik und die Metriken *Multi-Label-Accurarcy/-Precision/-Recall*.

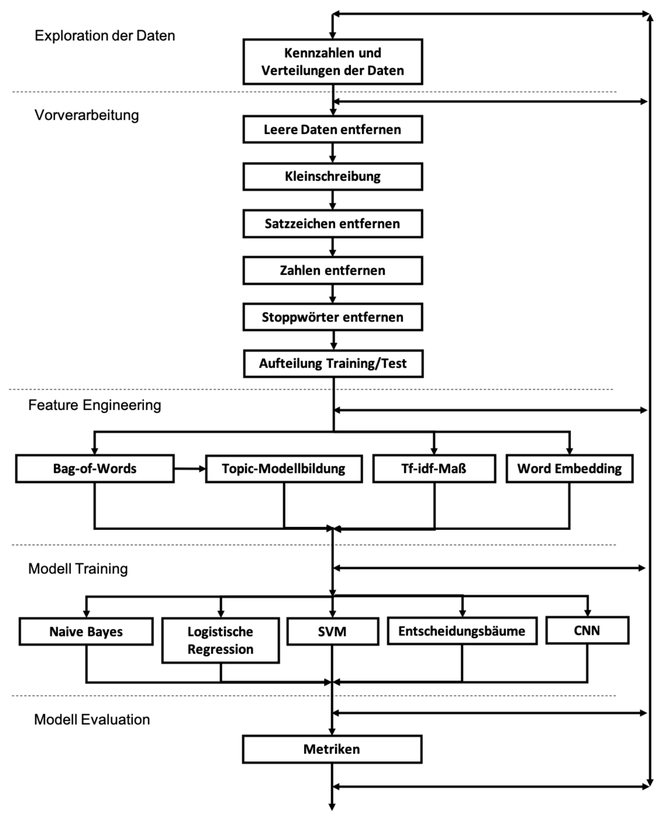

Abbildung 2: Aufbau der Modellbildung

Umsetzung

Das Klassifikationsproblem wurde in der Arbeit mit *Python 3.6.6* umgesetzt. Dabei wurden weiterhin Technologien verwendet, die lokal ausgeführt wurden oder bei den Anbietern der Daten betrieben werden. Eine Übersicht der wichtigsten verwendeten Technologien und deren Zusammenspiel ist in Abbildung 3 dargestellt.

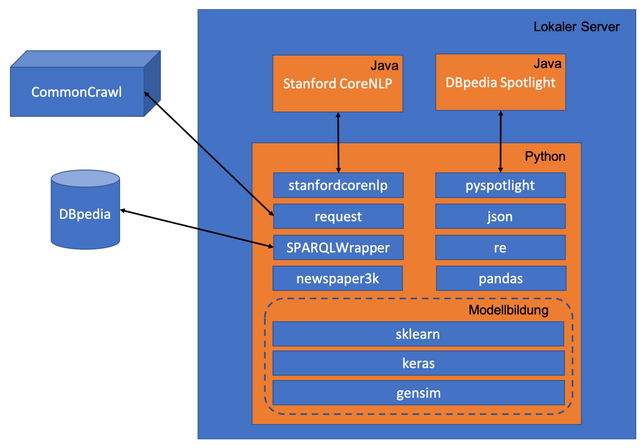

Abbildung 3: Architektur der verwendeten Technologien

Ergebnisse

Die Ergebnisse der Modelle zeigen, dass sie auf den aufbereiteten Daten und dem erstellten Klassifikationsproblem dazu geeignet sind, Branchen für die gegebenen Nachrichtentexte vorherzusagen. Daraus ergibt sich die Möglichkeit, mit Nachrichtentexten, in denen sich unbekannte Entitäten befinden, einen Rückschluss auf die Branche der Entität zu erhalten. Dadurch lassen sich Knowledge Bases um weitere Informationen erweitern. Als bestes Modell hat sich ein lineares SVM-Modell herausgestellt, das auf Bag-of-Words-Vektoren trainiert wurde. Das Modell erreicht eine Exact Match Ratio von $91{,}79\%$. Für die Multi-Label-, Micro- und Macro-Metriken erreicht es jeweils F1-Werte, die größer sind als $0{,}9$. Zur Untersuchung des Common-Crawl-Datensatzes auf für das Klassifikationsproblems relevante Daten wurde die Schicht der Datenaufbereitung auf ein *Apache Spark*-Cluster auf *Amazon AWS* übertragen. Schätungsweise $39{,}6\%$ der Records im betrachteten November-2018-Crawl enthalten Firmen und $31{,}5\%$ der Records enthalten Firmen, die in DBpedia bekannt sind. Dies sind geschätzt $986.720.000$, bzw. $784.616.000$ relevante Records.

Ausblick

Eine Weiterführung dieser Arbeit könnte die Umsetzung des Klassifikationsproblems auf einer größeren Datenmenge darstellen. Dabei könnten die Schritte der Datenaufbereitungs- und Modellbildungspipeline auf ein Spark-Cluster übertragen werden, um dort parallel ausgeführt zu werden. Außerdem könnte eine größere Anzahl an DBpedia-Branchen bei er Klassifikation einbezogen werden. Dabei sollten jedoch auftretende Korrelationen zwischen den Labeln berücksichtigt und einbezogen werden. Weitere Ansätze zum Lösen des Klassifikationsproblems bieten supervised Modelle aus dem Bereich Latent Dirichlet Allocation und Methoden aus dem Bereich Neuroale Netze und Deep Learning.