Maschinelles Lernen mit Titel- und Normdaten

Alexander Gerling und Andreas Heß

Agenda

- Problemstellung: Automatische Erschließung
- Maschinelles Lernen
- Verfahren
- Daten
- Evaluation
- Ausblick

Automatische Erschließung

Link zu diesem Datensatz	http://d-nb.info/112262364X
Titel	Java Programmierung im Bereich Internet of Things – Verwendung eines Raspber-ry Pi zur Erstellung von digitalen Einkaufslisten – Kann mit Hilfe eines Raspberry Pi das Einkaufen revolutioniert werden? / Eduard Franke ; Betreuer: Andreas Hess
Person(en)	Franke, Eduard (Verfasser) Heß, Andreas (Akademischer Betreuer)
Verlag	Furtwangen : Hochschule Furtwangen
Zeitliche Einordnung	Erscheinungsdatum: 2016
Umfang/Format	Online-Ressource
Hochschulschrift	Bachelorarbeit, Furtwangen, Hochschule Furtwangen, 2016
Persistent Identifier	URN: urn:nbn:de:bsz:fn1-opus4-11450
URL	https://opus.hs-furtwangen.de/frontdoor/index/index/docId/1145 (Verlag) (kostenfrei zugänglich)
Sprache(n)	Deutsch (ger)
Anmerkungen	Langzeitarchivierung gewährleistet
Sachgruppe(n)	004 Informatik ; 640 Hauswirtschaft und Familienleben

Wichtiger Hinweis!

- Wir sind (Wirtschafts-) Informatiker, keine Bibliothekare

Maschinelles Lernen

A computer program is said to learn

- from experience **E**
- with respect to some class of tasks **T**
- and performance measure **P**

if its performance at task **T**, as measured by **P**, improves with experience **E**.

Tom Mitchell

Maschinelles Lernen

- **T**: Sachgruppe automatisch vergeben
- **E**: Intellektuell vergebene Sachgruppen
  (für andere Inhalte aus der Vergangenheit)
- **P**: z.B. Anteil korrekte Klassifikationen (Accuracy), Precision, Recall, F1

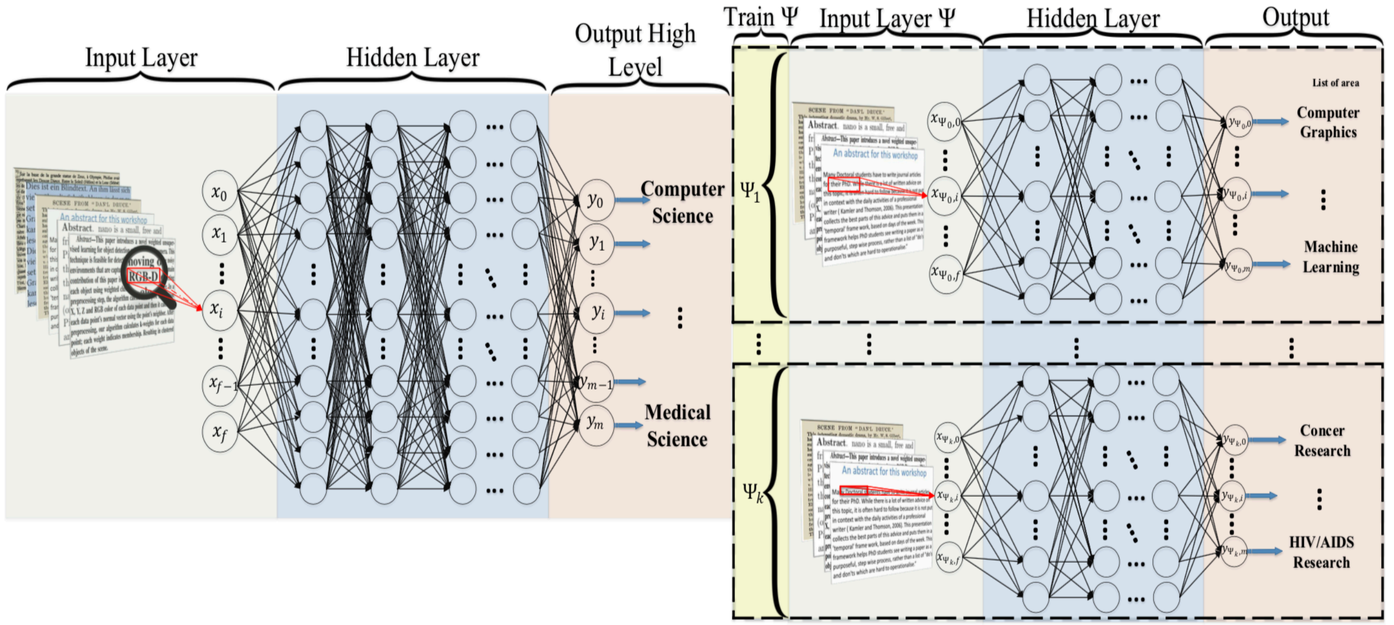
Verfahren

- Klassisch: Naive Bayes (statistisch), Support Vector Machines
- Im Trend: Neuronale Netze

Verwandte Arbeit: HDLTex

- Hierarchische Klassifikation
  ➔ klingt eigentlich gut für Sachgruppen

- Bringt nur leider nichts

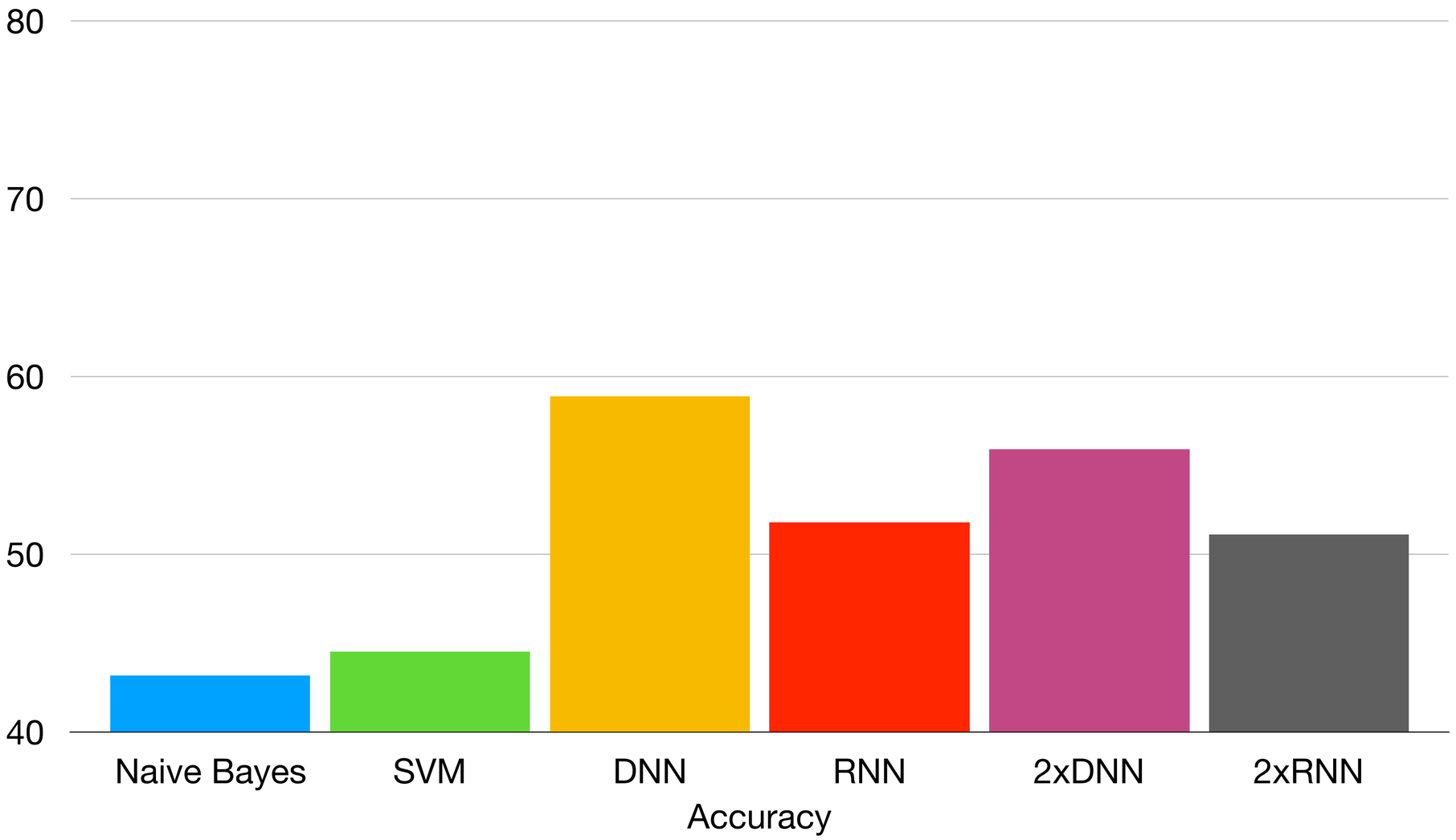


Kowsari et al.: HDLTex: Hierarchical Deep Learning for Text Classification

Korpus "DNB 1000"

- Hochschulschriften
- 34097 Katalogeinträge in 57 Sachgruppen
- 20 Sachgruppen gedeckelt auf 1000 Instanzen
- 19 Sachgruppen mit < 1000 und > 200 Instanzen
- 18 Sachgruppen mit ≦ 200 und ≧ 100 Instanzen
- Sachgruppen mit < 100 Instanzen entfernt
- Split in 80% Training- und 20% Testdaten

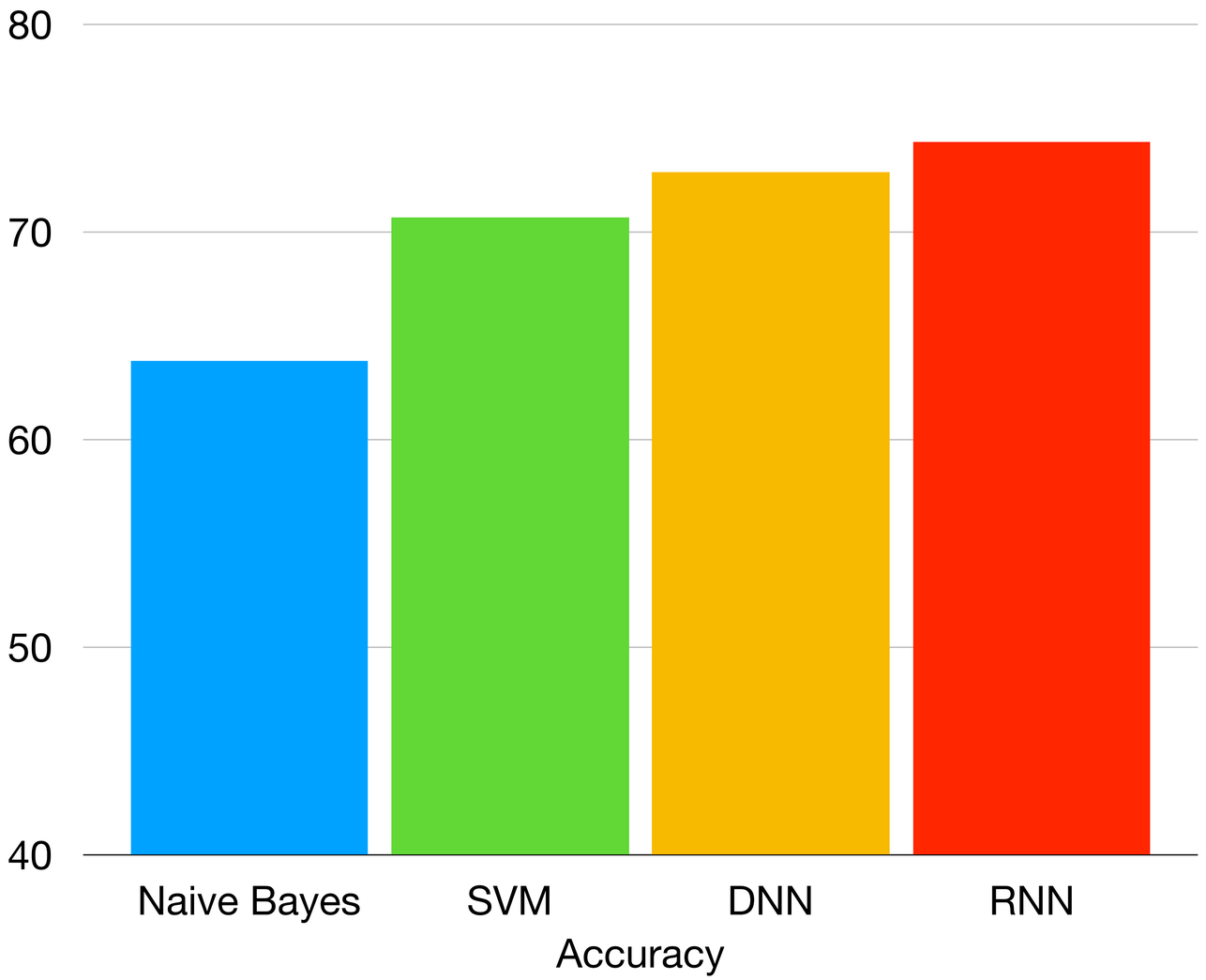
Evaluation "DNB 1000"



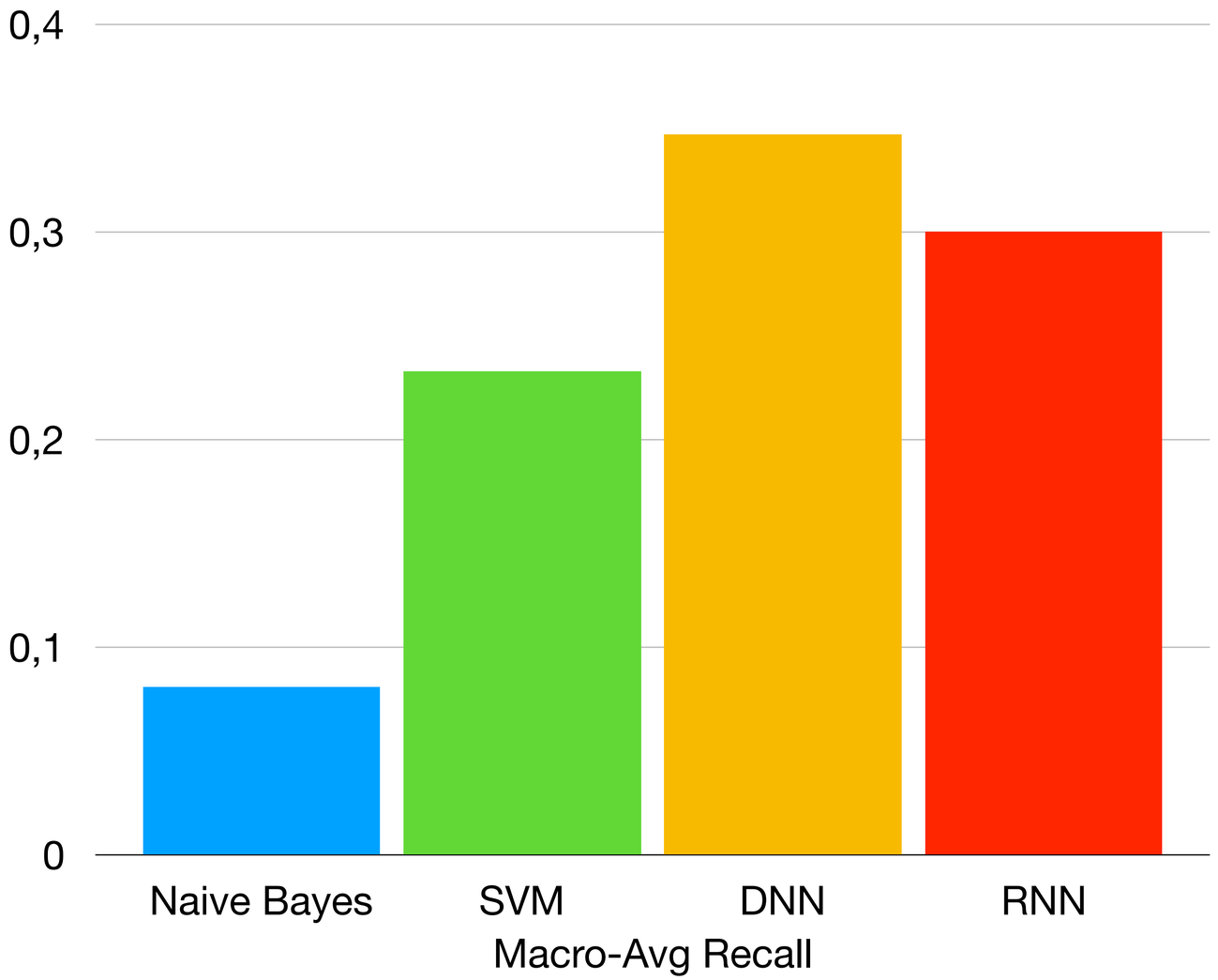
Korpus "DNB 2018"

- Hochschulschriften aus 2017 und 2018, nur Titel
- 35407 Katalogeinträge aus den Top 50 Sachgruppen
- 21730 Einträge in Sachgruppe 610
- zweitgrößte Sachgruppe 340 nur noch 1612 Einträge
- Split in 80% Training- und 20% Testdaten

Evaluation "DNB 2018"



Evaluation "DNB 2018"



Fazit

- Neuronale Netze für automatische Sachgruppenvergabe machbar
- sowohl Deep Neural Networks als auch Recurrent Neural Networks besser als Naive Bayes und SVM, besonders wenn Klassen ungleich verteilt sind
- hierarchische Klassifikation bringt nichts

Ausblick

- Vergleichbarkeit mit aktuell in der DNB verwendeten Verfahren herstellen
- Multi-Label-Klassifikation: Vorläufige Experimente vielversprechend
- Einbeziehung von Hintergrundwissen (z.B. Autor, Hochschule…)

Ergänzung:
Experimentelles Setup

Naive Bayes und SVM

- Implementierung von WEKA
- Preprocessing: keine Stoppwortliste, binäre Features
- SVM: SMO mit 1-vs-all-Konfiguration

Neuronale Netze

- Keras mit TensorFlow oder PlaidML-Backend
- 20 Epochen
- DNN: 1 hidden layer, 768 Neuronen, Dropout 0,5
- RNN: Glove Word Embeddings 100 Dimensionen, GRU, Dropout 0,2