Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Genre classification on German novels

Lena Hettinger, Martin Becker,
Isabella Reger, Fotis Jannidis, Andreas Hotho

University of Würzburg

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Kallimachos.de

Library,
JMU Würzburg

OCR,
DFKI Kaiserslautern

German Philology,
JMU Würzburg

Computer Philology,
JMU Würzburg

Computational Linguistics,
FAU Erlangen-Nürnberg

Computer Science,
JMU Würzburg


Theodor Fontane
Effi Briest
Reclam
Franz Kafka
Der Proceß
Reclam

Task

- Regency novel
  - Regency romance
- Literary fiction
- Literary nonsense
- Mathematical fiction
- Metafiction
- Nonfiction novel
  - Bildungsroman
  - Biographical novel
    - Autobiographical novel
      - Semi-autobiographical novel
        - I novel
  - Slave narrative
    - Contemporary slave narrative
    - Neo-slave narrative
- Occupational fiction
  - Hollywood novel
  - Lab lit

en.wikipedia.org/wiki/List_of_writing_genres

Genres of the novel [edit]

- Antinovel
- British regional literature
- Campus novel
- Gay literature
- Newgate novel
- New
- Paranormal
- Picaresque novel
- Proletarian
- Psychological
- Sensation novel
- Speculative
- Social novel
- Supernatural
- Thriller
- Westerns

en.wikipedia.org/wiki/Novel

Agenda

- Task
- **Data**
- Feature extraction
- Genre classification
- Evaluation

Data

- Two genre classes:
  - Bildungsroman = educational novel
  - Gesellschaftsroman = social novel

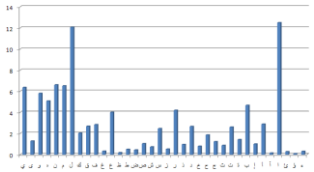
- 132 labeled novels (Labeled data set)
  - 32 were closely reviewed (Prototype data set)
- Entire corpus: close to 1700 novels
  - -> content features (LDA)

Agenda

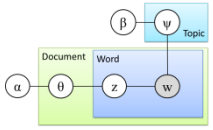
- Task
- Data
- **Feature extraction**
- Genre classification
- Evaluation

Features

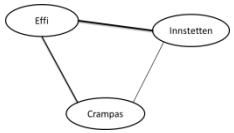
- **Stylometric features**



- Content features



- Social features



Stylometric

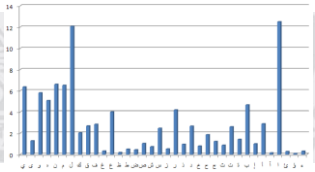
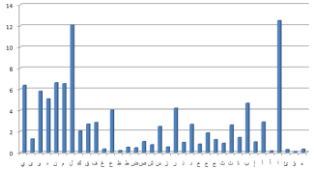


TABLE I. TOP TOKENS FOR CLOSE TO 1700 GERMAN NOVELS

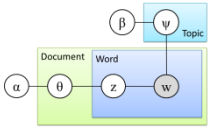
rank	words	frequency	rank	words	frequency
1	.	7.003.325	6	sie	2.312.731
2	,	5.026.058	7	er	2.188.019
3	und	4.050.873	8	zu	2.168673
4	die	3.745.769	9	≫	2.131.584
5	der	2.626.422	10	ich	1.945.178

Features

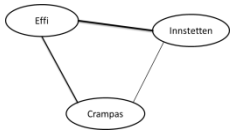
- Stylometric features



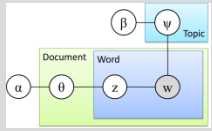
- **Content features**

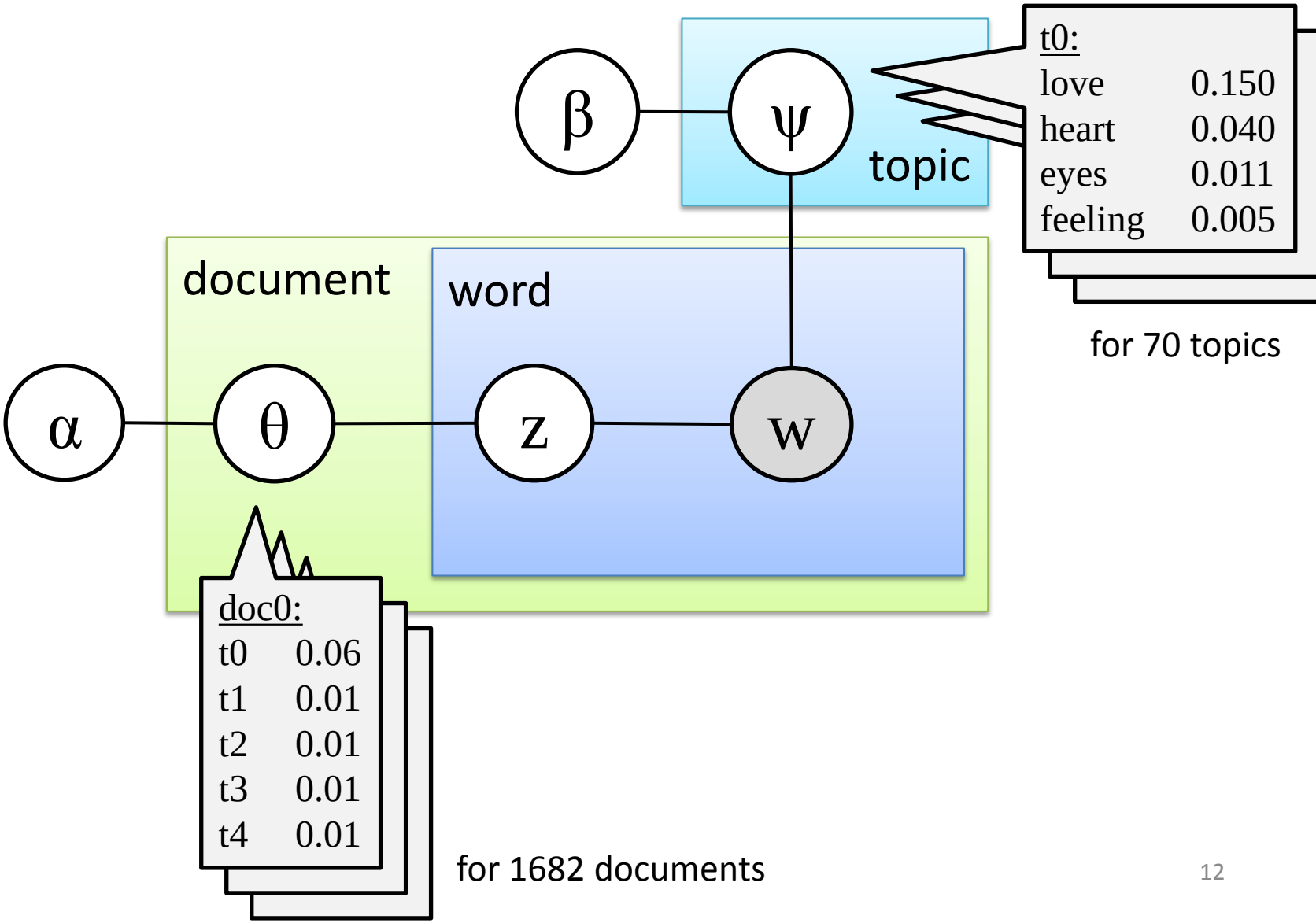


- Social features

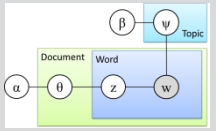


Content based: LDA





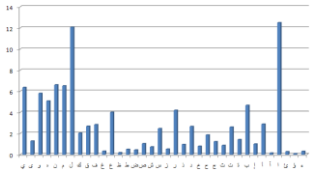
Discriminative topics



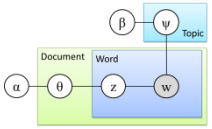
„Society“	„Emotions“
WOMAN/MRS	LOVE
MAN/MR	LIFE
PARIS	SELF
MADAME	HEART
FRANKEN	MOTHER
LOVE	SOUL
MAN	FATHER
WOMEN	WORLD
	EYES

Features

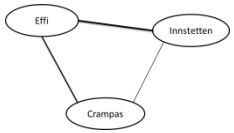
- Stylometric features



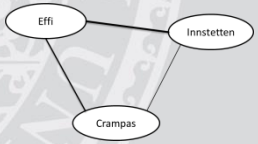
- Content features



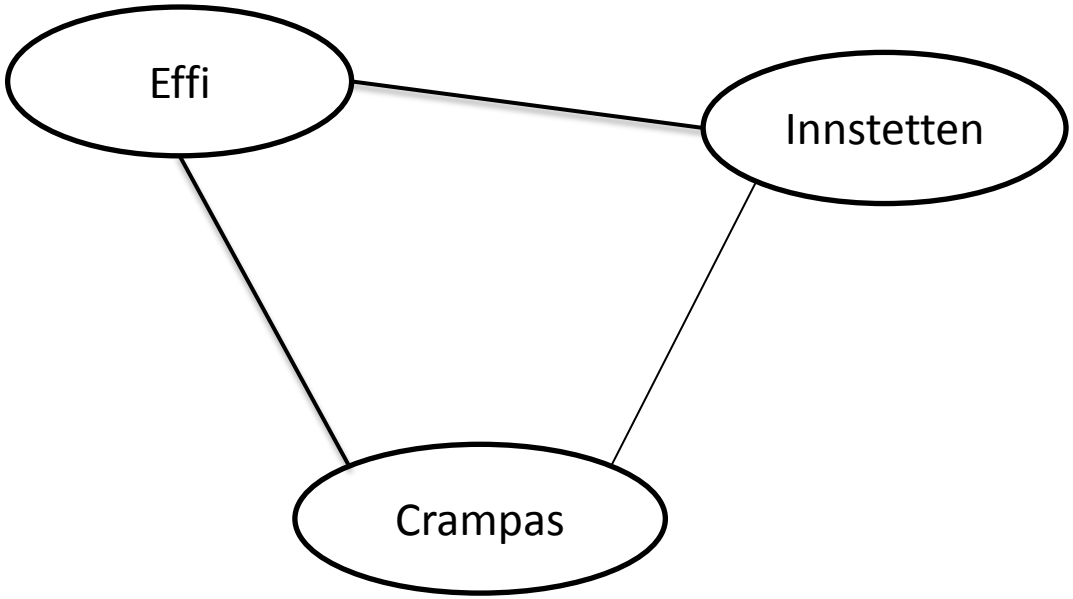
- **Social features**



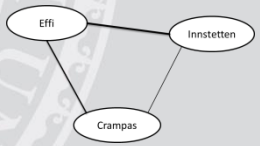
Character graph



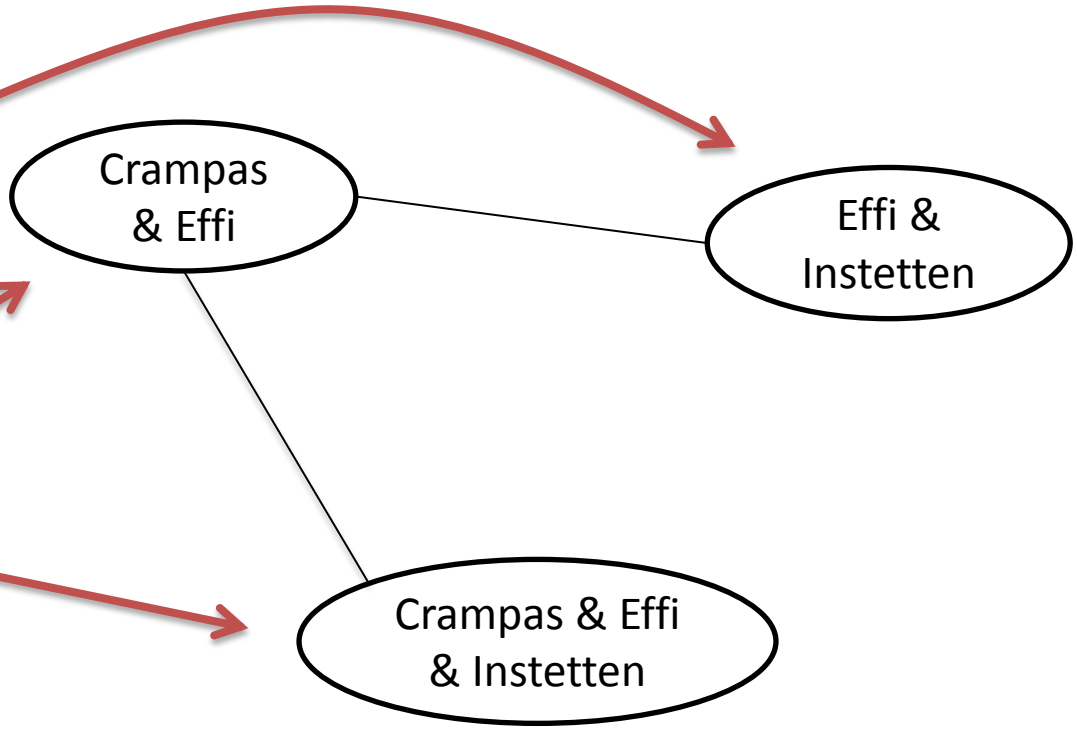
[...]
»... Ich muß dir nämlich sagen, **Effi**, daß Baron **Innstetten** eben um deine Hand angehalten hat.«
[...]
Effi sagte zu dem neben ihr sitzenden Major von **Crampas**:
[...]
Trotzdem schien **Innstetten** auf **Crampas**' scherzhafte Bemerkungen antworten zu wollen, was denn **Effi** bestimmte, lieber direkt einzugreifen.
[...]



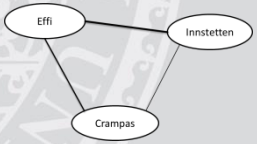
Interaction graph



[...]
»... Ich muß dir nämlich sagen, **Effi**, daß Baron **Innstetten** eben um deine Hand angehalten hat.«
[...]
Effi sagte zu dem neben ihr sitzenden Major von **Crampas**:
[...]
Trotzdem schien **Innstetten** auf **Crampas**' scherzhafte Bemerkungen antworten zu wollen, was denn **Effi** bestimmte, lieber direkt einzugreifen.
[...]



Social features



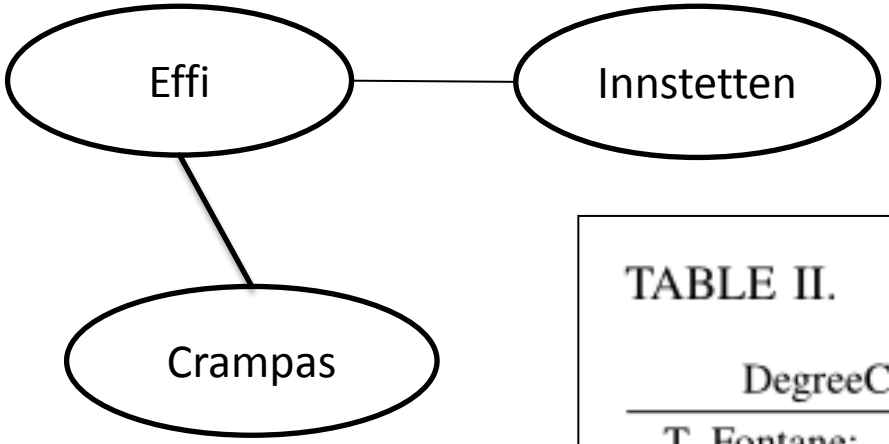
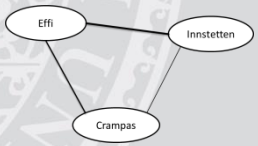


TABLE II. EXAMPLE FOR CENTRALITY VALUES

DegreeCentrality		EigenvectorCentrality	
T. Fontane: Effi Briest	C. Brontë: Jane Eyre	T. Fontane: Effi Briest	C. Brontë: Jane Eyre
0.366	0.174	0.222	0.206
0.140	0.153	0.208	0.145
0.087	0.111	0.154	0.103
0.076	0.097	0.115	0.095
0.070	0.090	0.077	0.094
0.058	0.083	0.066	0.090
0.058	0.076	0.053	0.078
0.052	0.076	0.049	0.068

Social features



Let $C = \{c_1^{mg}, \dots, c_{10}^{mg}\}$ be the 10 highest centrality values for measure m and graph type g

- Averages - Consider only the highest centrality value:
  - $average_{mg}(c_1^{mg})$
  - $a_c = average_m(c_1^{mc})$
  - $a_i = average_m(c_1^{mi})$
  - $a_c - a_i$
  - $a_c - (1 - a_i)$

  5 features

- Differences - $\forall\, m \in \{d, b, c, e\}, g \in \{c, i\}$
  - $c_1^{mg} - c_2^{mg}$, $c_2^{mg} - c_3^{mg}$, $c_3^{mg} - c_4^{mg}$

  24 features

- Power Law - $\forall\, m \in \{d, b, c, e\}, g \in \{c, i\}$:
  - Fit $f(c) = a * c^b$ to $c \in C$ and extract parameters a, b

  16 features

Agenda

- Task
- Data
- Feature extraction
- **Genre classification**
- Evaluation

Feature sets

- All
- Only
  - Stylometric (100 Features)
  - Content (70 Features)
  - Social (45 Features)
- Combined
  - Stylometric + content
  - Stylometric + social
  - Content + social

Classifier

- Baseline: majority vote

- K-nearest Neighbour
- Naive Bayes
- Rule-based learning
  - Mixed Fuzzy Rule Formation
- Decision tree (C4.5)
- Neural Network (MLP)
- Support Vector Machine
  - LibSVM, linear Kernel

- Task
- Data
- Feature extraction
- Genre classification
- **Evaluation**

Accuracy (Labeled, 132)

	MV	kNN	NB	Rule	Tree	pTree	NN	SVM
All		0.64	0.70	*0.51*	0.65	0.65	0.70	0.73
Stylo		0.64	0.64	*0.54*	0.61	0.60	0.64	0.69
Content		0.67	0.69	*0.49*	0.67	0.69	0.69	0.81
Social	0.58	*0.58*	0.61	*0.42*	*0.52*	*0.56*	*0.59*	0.59
Stylo+ Content		0.67	0.69	*0.50*	0.67	0.65	0.71	0.74
Stylo+ Social		0.64	0.63	*0.48*	0.61	0.60	0.64	0.71
Content + Social		*0.57*	0.69	*0.52*	0.65	0.67	0.69	0.78

Results

- Accuracy of over 80%
- Contrary to assumptions in literary studies, content features instead of number of characters works best
- Stylometric and social features perform worst in this setting

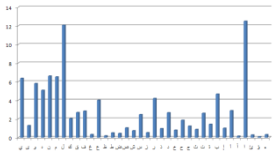
Further goals

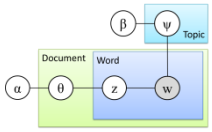
- Optimize parameters of classifiers & features
- Further investigate social features
- Develop new topic models which integrate genre
- Collect more labeled data

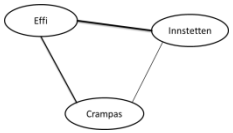
Thanks for your attention

Dataset: dmir.org/genre-data

Features:







Results: Accuracy of more than 80%,
discriminative features differ from literature

Lena Hettinger
Martin Becker
Prof. Andreas Hotho

INFORMATION RETRIEVAL DATA MINING GROUP

KALLIMACHOS
Isabella Reger
Prof. Fotis Jannidis