Spezifikationsvortrag - Graphbasierte Methode zur Generierung von deutschen Lexika für Hate Speech

Fabian Düker, Uli Steinbach

Universität Heidelberg, Institut für Computerlinguistik
Softwareprojekt, SoSe 2018

Prof Dr. Katja Markert

12.06.2018

Übersicht I

Übersicht

Autom. Erstellung eines Lexikons für die Erkennung von Abusive Words

Anwendung auf Germeval 2018 Task I $\Rightarrow$ Binäre Klassifikation von 5000 Tweets

Problemstellung

- Problem: Hatespeech ist in ständiger Veränderung begriffen (Neologismen, Ambiguität, Kontext/Domäne)
- Wiegand et al. 2016: Erstellung eines englischen Lexikons mit guten Ergebnissen auf cross-domain Evaluation
- Erstellung eines kontextunabhängigen Lexikons für das Deutsche
- Anwendung des Lexikons bei der (binären) Erkennung von Tweets

Lösungsansatz

- manuelle Erstellung Baselexikon aus SentiWS neg. Sentiment-Lexikon
- Halbautomatische Erweiterung des Baselexikons mit deutschen Schimpwörtern
- Autom. Erweiterung mittels graphbasiertem Label-Propagation-Algorithmus
- Anwendung auf Germeval 2018 Datenset und Evaluation

Erstellung des Baselexikons

SentiWS

- Extraktion negativer Wörter aus SentiWS
- 686 Nomen
- 420 Verben
- 708 Adjektive
- Problem: Zu wenige explizite Schimpfwörter
- Lösung: Mehr Schimpfwörter hinzufügen

Erstellung des Baselexikons

- Genius API: Erstellung eines Deutschrapkorpus
- Deutschrap: zeitgemäße Verwendung von Schimpfwörtern (genrespezifisch, aber auch politisch, rassistisch, sexistisch)
- autom. Extraktion von Kandidaten mittels syntaktischer Pattern

Erstellung des Baselexikons

- Automatischer Abgleich aller Nomen im Rapkorpus
- mit Schimpfwortliste aus dem Internet
- mit Pattern "du [NN]"
- etwa 280 potentielle Schimpfwörter
- manuelles Aussortieren von false positives (z.B. "Rapper")
- Auswahl der 200 häufigsten Schimpfwörter
- Erweiterung durch beleidigende Adjektive
- Suche nach Pattern "du [ADJ] Schimpfwort"
- etwa 280 potentiell beleidigende Adjektive

Erstellung des Baselexikons

- Lemmatisierung mit IWNLP
- Lemmatisierung der nicht erkannten Adjektive von Hand
- Beseitigung von Duplikaten
- Finales Baselexikon:
- 887 Nomen
- 413 Verben
- 824 Adjektive
- - 2124 Wörter

Rapkorpus

- Texte von 30 Rappern (Auswahl angelehnt an Daten-Journalismus Studie des BR/PULS zum Thema "Diskriminierung im Deutschrap" aus dem Jahr 2016)
- Bushido
- Chakuza
- K.I.Z.
- Kay One
- Kollegah & Farid Bang
- Prinz Pi
- Bass Sultan Hengzt
- Fler
- Azad
- Kool Savas
- ...

Rapkorpus

Auszug der extrahierten Schimpfwörter vor Handselektion

- *Rapper*
- *Kopf*
- **Arsch**
- **Bitch**
- **Schwanz**
- **Scheiße**
- *Gangster*
- *Block*
- *Baby*
- **Nutte**

Graph-basierter Ansatz für autom. Erweiterung des Baselexikons

- Erstellung von pos + neg seed-Liste mit annotierten Schimpwörtern aus Baselexikon (+) und häufigsten Wörtern (-)
- Graph mit Kanten zw. Wörtern auf Basis von Kosinusähnlichkeit zwischen Word Embeddings Vektoren (auf Twitter Korpus trainiert)
- Propagierung der Seed-Labels auf ungelabelte Knoten/Wörter mittels graphbasiertem Label-Propagation-Algorithmus (Adsorption Algorithmus, Talukdar/Crammer 2009)

Graph-basierter Ansatz für autom. Erweiterung des Baselexikons

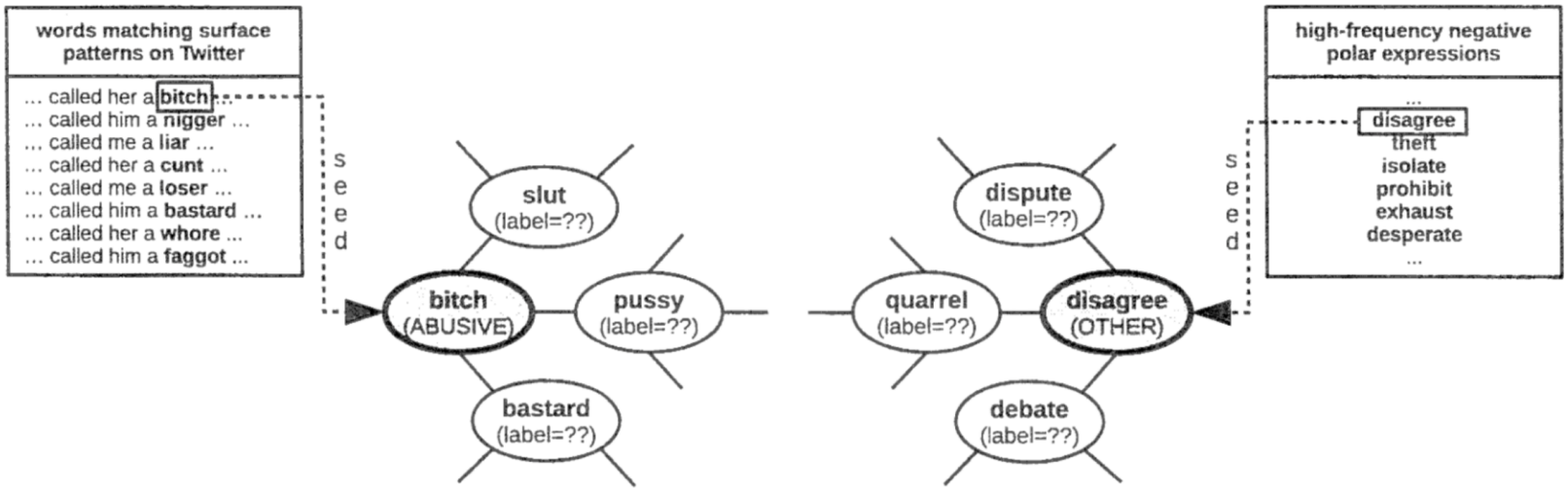

Figure: Label Propagation Graph aus Wiegand et al. 2018

Transduktives Lernen (statt induktiv) Menge von labeled und Menge von unlabeled Instanzen Goal: Label the unlabeled instances! mehrere Ansätze/Varianten: Random Walk Variante: kontrollierter Random Walk über den Graphen G kontrolliert über 3 Wahrscheinlichkeiten für jeden Knoten (inj, cont, abnd)

Anwendung und Evaluation auf Germeval Datenset

- Germeval 2018: deutschsprachiger Twitter Korpus ⇒ 5000 annotierte Tweets bzgl. beleidigender Sprache
- Task I: binäre Klassifikation der Tweets
- **beleidigend:**
  `Für mich ist die #katholische #Kirche nur ein Sammelbecken von #Verbrechern #Kinderschändern und #Kriminellen`
- **other:**
  `Endlich hat Kurz einen Verbündeten aus Frankreich, der auch die LBR ungesetzliche Einwanderung von jungen Afrikanern unterbinden will`

Anwendung und Evaluation auf Germeval Datenset

- Unser Ansatz für die Verwendung des Baselexikons und warum könnte das besser sein ?

Baseline

- Baseline 1: Majority Baseline
- Baseline 2: Unigram und Bigram SVM
- Baseline 3: Feature Selection (Mutual Information) SVM
- Preprocessing: Autosarkasmus-SP (SoSe 2016): Alle Tweets wurden tokenisiert, normalisiert und pos-getagged. Zusätzlich Lemmatisierung und stopword removal
- 10-Fold Cross Validation mit random-seed für bessere Vergleichbarkeit

Baseline

- Unigram und Bigram SVM: Input-Features sind die auf der Dokument-Term Matrix berechneten tf-idf Werte für Uni- bzw. Bigramme.
- Feature Selection Algorithmus: Berechnung des mutual information score (Manning et. al 2011) zwischen Label und Wort $\Rightarrow$ 1500 Wörter mit den höchsten mi-scores als Input-Features für SVM Klassifizierer.

Evaluation der Baseline

Table: Baseline: tf-idf unigram SVM

Fold	#train	pos	neg	tok/tw	#test	pos	neg	tok/tw	F1
1	4500	1524	2976	146.94	500	160	340	144.124	0.7026952348344236
2	4500	1560	2940	146.73	500	124	376	146.05	0.7430423747345847
3	4500	1525	2975	146.73	500	159	341	146.0	0.6822048536263124
4	4500	1505	2995	146.57	500	179	321	147.512	0.6637390819218496
5	4500	1498	3002	145.94	500	186	314	153.154	0.6466493970138341
6	4500	1504	2996	147.0	500	180	320	143.568	0.6675527979162865
7	4500	1504	2996	146.61	500	180	320	147.106	0.6563892890626787
8	4500	1519	2981	146.16	500	165	335	151.13	0.6681861534976389
9	4500	1504	2996	147.08	500	180	320	142.86	0.650127611518916
10	4500	1513	2987	146.83	500	171	329	145.104	0.6737827241885149
									Total Accuracy: 0.68 (+/- 0.05)

Evaluation der Baseline

Table: Baseline: tf-idf bigram SVM

Fold	#train	pos	neg	tok/tw	#test	pos	neg	tok/tw	F1
1	4500	1524	2976	146.94	500	160	340	144.124	0.6713317445366677
2	4500	1560	2940	146.73	500	124	376	146.05	0.7143231005141465
3	4500	1525	2975	146.73	500	159	341	146.0	0.6607936275023581
4	4500	1505	2995	146.57	500	179	321	147.512	0.6407256780069586
5	4500	1498	3002	145.94	500	186	314	153.154	0.6004567364407795
6	4500	1504	2996	147.0	500	180	320	143.568	0.636733440679629
7	4500	1504	2996	146.61	500	180	320	147.106	0.6211176439430055
8	4500	1519	2981	146.16	500	165	335	151.13	0.6254687762688433
9	4500	1504	2996	147.08	500	180	320	142.86	0.6184350158030838
10	4500	1513	2987	146.83	500	171	329	145.104	0.6386443278943279
									Total Accuracy: 0.64 (+/- 0.06)

Evaluation der Baseline

Table: Baseline: Feature-Selection m. Mutual Information

Fold	#train	pos	neg	tok/tw	#test	pos	neg	tok/tw	F1
1	4500	1524	2976	146.94	500	160	340	144.124	0.7506333288282201
2	4500	1560	2940	146.73	500	124	376	146.05	0.8159589743589742
3	4500	1525	2975	146.73	500	159	341	146.0	0.7661802897341697
4	4500	1505	2995	146.57	500	179	321	147.512	0.761125609472951
5	4500	1498	3002	145.94	500	186	314	153.154	0.717136006424597
6	4500	1504	2996	147.0	500	180	320	143.568	0.7715594209711858
7	4500	1504	2996	146.61	500	180	320	147.106	0.7808015386464718
8	4500	1519	2981	146.16	500	165	335	151.13	0.7812137931034483
9	4500	1504	2996	147.08	500	180	320	142.86	0.7738412698412699
10	4500	1513	2987	146.83	500	171	329	145.104	0.7651150793650793
Total Accuracy: 0.77 (+/- 0.05)									

- Paketierung: Aufsplittung des Gesamtpakets in kleinere Module
- Python-Pogrammierung (objektorientierte Implementierung)
- "split und shared tasks"

Modul		Uli	Fabian
Erstellung Baselexikon			
	Extraktion negativer Wörter aus SentiWS		x
	Erstellung Rapkorpus	x	
	PAT-basierte Extraktion von NNs/ADJs	x	x
	Autom. Abgleich von Nomen mit Schimpfwortliste und Patterns		x
	Handselektion der beleidigenden Nomen		x
	Lemmatisierung des Lexikons mit IWNLP Lemmatizer		x
	Korrektur und Restlemmatisierung des Lexikons		x
	Skript für Erstellung eines Annotations-Testsets	x	
	Skript für autom. Auswertung d. Annotations-Testsets		x
Erstellung Baselines			
	Implementation manuelle 10-Fold Cross Validation	x	
	Integration Autosarkasmus-Tweet Preprocessing	x	
	Erweitertes Preprocessing (Lemmatisierung, Stopwörter)	x	
	Implementation Unigram/Bigram SVM Baseline	x	
	Implementation Mutual Information (Manning et. al 2011)	x	
	Implementation Feature Selection (MI) SVM Baseline	x	
	Evaluation und Output	x	
Erstellung Word-Similarity Graph			
	Word Embeddings auf Twitter Daten	x	x
	Erstellung des Wortähnlichkeitsgraphen auf Basis von Kosinusähnlichkeiten	x	x
	Unknown Words Handhabung (character-level embeddings?)	x	x
	Erstellung der Seed Listen (pos + neg)	x	x
Label Propagation			
	Implementation Adsorption Algorithmus (Talukdar 2008)	x	x
	Erweiterung des Baselexikons mit Output	x	x
Anwendung und Evaluation			
	Test auf Germeval Daten	x	x
	Verbesserungen und Erweiterungen	x	x
	Visualisierung des Outputs	x	x
	Präsentation der Ergebnisse	x	x
	Abschlussbericht	x	x

Literatur

- Talukdar, Crammer 2009 - New Regularized Algorithms for Transductive Learning
- Velikovich et al. 2010 - The Viability of Web-derived Polarity Lexicons
- Wiegand et al. 2018 - Inducing a Lexicon of Abusive Words - A Feature-Based Approach
- Manning et al. 2008 - Introduction to Information Retrieval
- Autosarkasmus SWP 2016 - https://gitlab.cl.uni-heidelberg.de/hoff/GermanTwitterPreprocessing
- BR/PULS Studie 2016 - https://www.br.de/puls/musik/so-homophob-frauenfeindlich-rassistisch-und-behindertenfeindlich-ist-deutschrap-100.html

Ressourcen

- **Schimpfwortliste** http://www.hyperhero.com/de/insults.htm
- **SentiWS** http://wortschatz.uni-leipzig.de/en/download/
- **Genius API** https://genius.com/
- **spaCy** https://spacy.io/