HAU at the GermEval 2019 Shared Task on the Identification of Offensive Language in Microposts

System Description of Word List, Statistical and Hybrid Approaches

Johannes Schäfer[1], Tom De Smedt[2], and Sylvia Jaki[3]

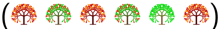
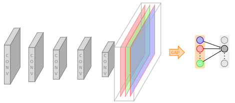
[1] Institute for Information Science and Natural Language Processing, Hildesheim
[2] Computational Linguistics Research Group, University of Antwerp
[3] Department of Translation and Specialized Communication, U. of Hildesheim

johannes.schaefer@uni-hildesheim.de,
tom.desmedt@uantwerpen.be,
jakisy@uni-hildesheim.de

October 8th, 2019

Motivation

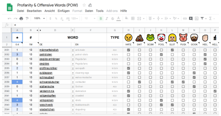
Best performing systems from last year:

Random forest () and CNN (

)

From our research:

Manually created/annotated word list



→ combination possibilities?

Overview

Overview

Overview POW List

Profanity and Offensive Words (POW)

- **Manually annotated** dictionary which allows for the quantitative analysis of hate speech in a dataset
- Decision to work with a dictionary - result of GermEval 2018
- List of **2852 words**, mainly taken from German Twitter Embeddings (Ruppenhofer, 2018)
- Words either often used tendentiously in political contexts or vulgar/offensive

POW List: Types of Words

Universiteit Antwerpen
Universität Hildesheim

Word classes (mostly)

- Nouns (*Lüge, Wesen, Arsch, Firlefanz*), incl. compounds (*Fremdenfeind, Lügenpresse*)
- Also: adjectives (*blöd, links-grün*) and participles (*verblendet*)
- Infinitives (*hetzen, spucken*) and imperatives (*lutsch, laber*)
- Interjections (*mimimi, boah*)

Separate entries (tokens)

- Declensions (*Dreckschwein, Dreckschweine*)
- Conjugations (*labern, laber, labert*)
- Spelling variations (*schreien/schrein, scheiß/scheiss/scheis/chice*)

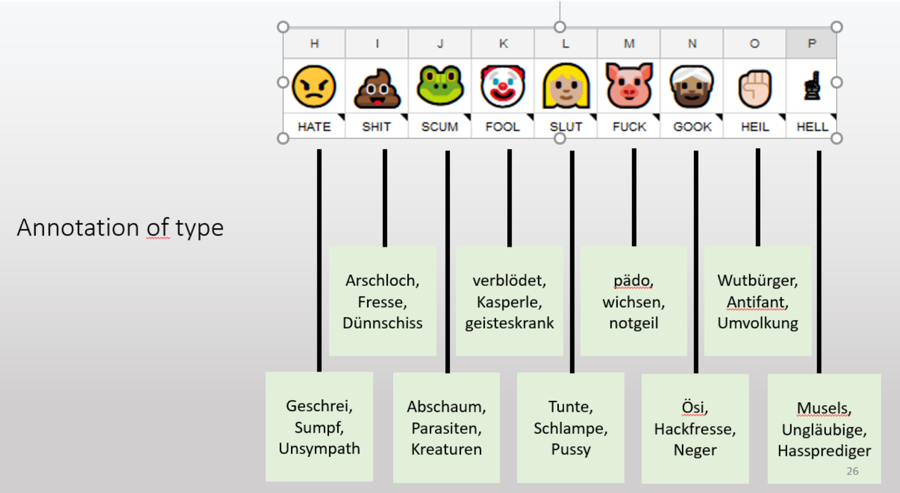
POW List: Annotation

Annotation of **intensity**

0. tendentious (*nichtmal, religiös, AfDler, Staub, Übergriffe*)
1. tendentious, sensational (*heulen, unkontrolliert, Extremisten*)
2. demeaning (*Schnauze, stupide, Systemparteien, antideutsch*)
3. **offensive (vulgar, racist)** (*verblödet, Dreck, Honk, Lügenpresse*)
4. **offensive (extremely so)** (*Hure, Untermenschen, Drecksau*)

POW List: Annotation of Types

Universiteit Antwerpen
Universität Hildesheim



POW List: Difficulties

Context-dependence

- Intensity (*honk, verrecken, hurensöhne*)
- Polarity (*bunt, willkommenskultur, fachkräfte*)

Type

- Lexial ambiguity (*geil, sack, fickt, würgen, schwuler, dödel, muschi*)
- Grammatical ambiguity (*quatsch, blase, leeren, ritze*)

⇒ Pragmatic solution:
Possibility for contextualisation by direct link to social media

POW List

Universiteit Antwerpen

Universität Hildesheim

Profanity & Offensive Words (POW)

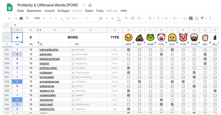
Datei Bearbeiten Ansicht Einfügen Format Daten Tools Add-ons Hilfe

	A	B	D	F	G	H	I	J	K	L	M	N	O	P
1	★	#	WORD		TYPE									
2	0-4	1M	DE	EN		HATE	SHIT	SCUM	FOOL	SLUT	FUCK	GOOK	HEIL	HELL
2051	1	13	männerfeindlich	anti-male	ADJ	✓				✓				
2052	3	13	patridioten	patriot idiots	NOUN				✓				✓	
2053	0	13	pegida-anhänger	Pegida-fan	NOUN								✓	
2054	0	13	pegidas	Pegida-fans	NOUN								✓	
2055	2	13	profilneurotiker	sociopath	NOUN	✓								
2056	1	13	putzlappen	cleaning rags	NOUN	✓						✓		
2057	1	13	schnauzbart	walrus mustache	NOUN							✓		
2058	4	13	schwanzlutscher	cocksucker	NOUN		✓			✓	✓			
2059	0	13	sizilianische	Sicilian	ADJ							✓		
2060	1	13	verdammnis	damnation	NOUN	✓								✓
2061	1	13	vollgestopften	crowded	ADJ	✓								
2062	4	13	vollspacken	idiots	NOUN				✓					
2063	3	13	waschweib	blabbermouth	ADJ				✓	✓				
2064	2	13	wiederlichen	disgusting	ADJ	✓								

Overview

Overview

System HAU2: POW List Lookup

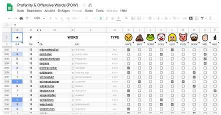


$\rightarrow$ Tweet

- **Motivation**:
  Word lists are very explainable (cf. "black boxes") and precise
- **Method**:
  - For each message, check if it has words that are also in the POW list
  - Compute the sum of the score of those words $>$ threshold $\Rightarrow$ offensive
  - Mapping of intensity annotation (0-4 in POW list):
    $0 \rightarrow 0.1, \quad 1 \rightarrow 0.25, \quad 2 \rightarrow 0.5, \quad 3/4 \rightarrow 1.0$
  - **For example**:
    "Ungebildetes, kulturloses Gesindel führt Deutschland vor!"
    $\rightarrow$ *ungebildet* (0.5) + *gesindel* (1.0) $= 1.5 > 0.95 \Rightarrow$ offensive

System HAU2: POW List Lookup

Universiteit Antwerpen
Universität Hildesheim



$\rightarrow$ Tweet

- **Motivation**:
  Word lists are very explainable (cf. "black boxes") and precise
- **Method**:
  - For each message, check if it has words that are also in the POW list
  - Compute the sum of the score of those words $>$ threshold $\Rightarrow$ offensive
  - Mapping of intensity annotation (0-4 in POW list):
    $0 \rightarrow 0.1, \quad 1 \rightarrow 0.25, \quad 2 \rightarrow 0.5, \quad 3/4 \rightarrow 1.0$
  - **For example**:
    "Ungebildetes, kulturloses Gesindel führt Deutschland vor!"
    $\rightarrow$ *ungebildet* $(0.5) +$ *gesindel* $(1.0) = 1.5 > 0.95 \Rightarrow$ offensive
- **Results**:
  Low recall for OFFENSE: 37.11% (lexicon should be expanded)

Overview

Universiteit Antwerpen
Universität Hildesheim

System HAU3: Random Forest

Universiteit Antwerpen
Universität Hildesheim

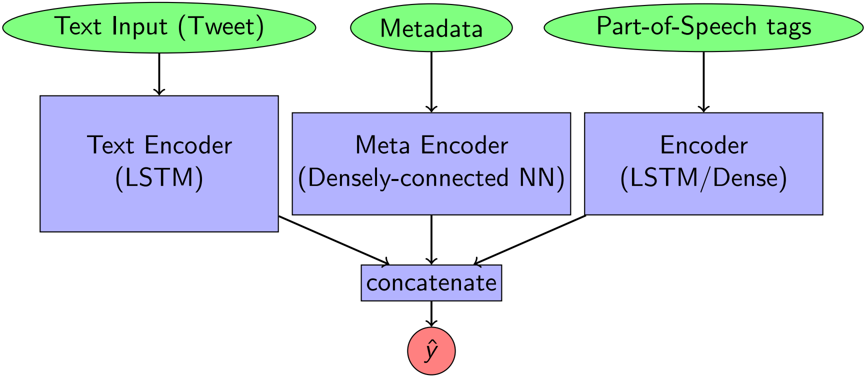
- **Motivation**:
  among last year's best systems, use as **comparative baseline**
- **Python algorithm**: `https://github.com/textgain/grasp`
- **Features**: character trigrams + word unigrams
- 100 trees, each with a random subset of 750 features

Overview

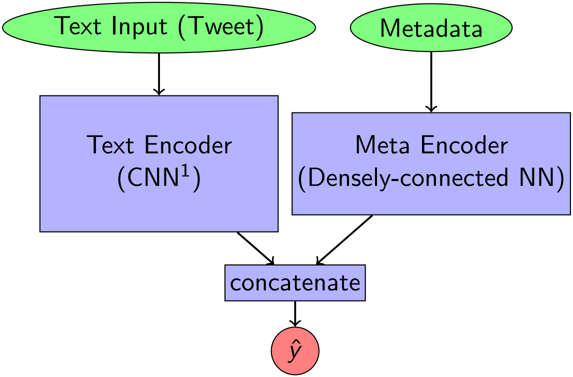
Universiteit Antwerpen Universität Hildesheim

Starting Point: NN Architecture

Schäfer (2018) at GermEval 2018; extended from Founta et al. (2018)

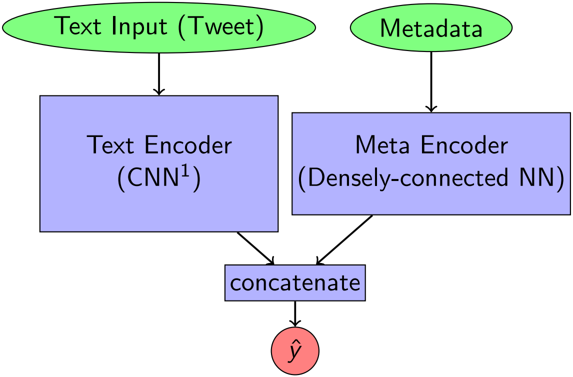


Our Basic NN Architecture for GermEval 2019



[1] CNN configuration as described in Schäfer and Burtenshaw (2019)

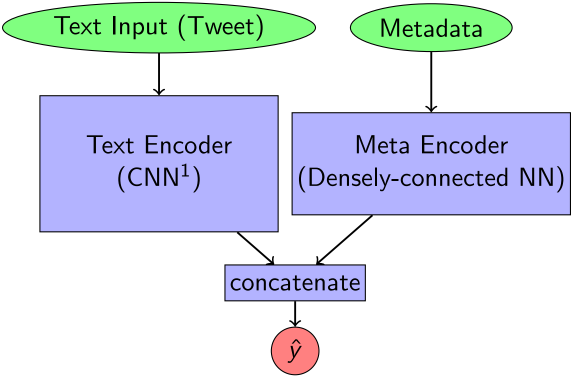
Our Basic NN Architecture for GermEval 2019



- ML improvements: early stopping; class weights

[1] CNN configuration as described in Schäfer and Burtenshaw (2019)

Our Basic NN Architecture for GermEval 2019

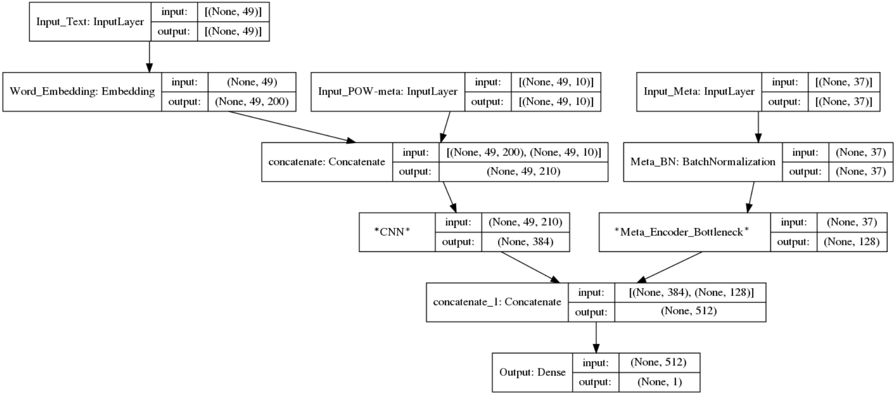


- ML improvements: early stopping; class weights

→ **POW list features?**

[1] CNN configuration as described in Schäfer and Burtenshaw (2019)

HAU1: CNN + POW List Model



Results on the GermEval Training Dataset

Universiteit Antwerpen
Universität Hildesheim

Average scores from 3-fold cross validation (values in %):

System configuration	Accuracy	F_1-score		
		OTHER	OFFENSE	m.-avg.
CNN	76.25	83.02	60.47	71.98
CNN + meta	76.10	82.23	63.43	72.84
CNN + $meta_{POW}$	78.15	83.77	**66.56**	75.17
CNN_{POW} + meta	76.67	82.62	64.45	73.56
CNN_{POW} + $meta_{POW}$	**78.87**	**84.62**	66.21	**75.46**

Overview

Overview System Runs HAU1-3 for Tasks 1-3

F_1-scores on the GermEval 2019 test dataset

Subtask I (OL detection):

HAU2 (POW list lookup)	68.13%
HAU3 (random forest)	69.75%
HAU1 (CNN+meta including POW)	70.46%

Subtask II (fine-grained OL detection):

HAU3 (random forest)	40.80%
HAU1 (CNN+meta including POW)	45.34%

Subtask III (implicit/explicit):

HAU1 (CNN+meta including POW)	69.3%

Conclusion

Based on our results:

- Simple word list lookup approach is not that bad!
- Statistical ML approaches (CNN here) improve considerably when combining it with word list

Outlook

Future Work:

- Normalization
- Other neural approaches, e.g. contextualized character embeddings
- Linguistic features
- Outlook: further collaboration in EU-project *DeTACT* *(Detect Then ACT: Taking Direct Action against Online Hate Speech by Turning Bystanders into Upstanders*)

References

- Josef Ruppenhofer. 2018. *German Twitter Embeddings*. `http://www.cl.uni-heidelberg.de/english/research/downloads/resource_pages/GermanTwitterEmbeddings/GermanTwitterEmbeddings_data.shtml`.
- Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. *Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language*. 14th Conference on Natural Language Processing KONVENS 2018.
- Johannes Schäfer. 2018. *HIIwiStJS at GermEval-2018: Integrating Linguistic Features in a Neural Network for the Identification of Offensive Language in Micropost*, In Proceedings of the Workshop Germeval 2018 – Shared Task on the Identification of Offensive Language. Vienna, Austria. September 21, 2018.
- Antigoni-Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2018. *A unified deep learning architecture for abuse detection.* CoRR, abs/1802.00385.
- Johannes Schäfer and Ben Burtenshaw. 2019. *Offence in Dialogues: A Corpus-Based Study*. Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2019), pages 1085-1093, Varna, Bulgaria, September 2-4, 2019.