Identification and categorization of offensive language in German tweets

Kinga Gémes

kinga.gemes@tuwien.ac.at

12 May 2021

Warning

Warning! The following presentation contains foul language.

What is toxicity and offensive language?

Toxicity: An extremely harsh, malicious, or harmful quality. (Merriam-Webster dictionary[1])
Offensive: Something that is offensive upsets or embarrasses people because it is rude or insulting. (Collins dictionary[2])

[1] https://www.merriam-webster.com/
[2] https://www.collinsdictionary.com/

Classic approaches

- Dictionaries:

Classic approaches

- Dictionaries:
  - Slang

Classic approaches

- Dictionaries:
  - Slang
  - Emotion

Classic approaches

- Dictionaries:
  - Slang
  - Emotion
  - WordNet and other knowledge bases

Classic approaches

- Dictionaries:
  - Slang
  - Emotion
  - WordNet and other knowledge bases
- Problems:

Classic approaches

- Dictionaries:
  - Slang
  - Emotion
  - WordNet and other knowledge bases
- Problems:
  - l3375P3Ak (Leetspeak)

Classic approaches

- Dictionaries:
  - Slang
  - Emotion
  - WordNet and other knowledge bases
- Problems:
  - l3375P3Ak (Leetspeak)
  - Ever evolving language

Classic approaches

- Dictionaries:
  - Slang
  - Emotion
  - WordNet and other knowledge bases
- Problems:
  - l3375P3Ak (Leetspeak)
  - Ever evolving language
  - Sarcasm

Datasets

- GermEval2018 - 5009 + 3398 German tweets

Datasets

- GermEval2018 - 5009 + 3398 German tweets
- GermEval2019 - 3980 + 3031 German tweets

Datasets

- GermEval2018 - 5009 + 3398 German tweets
- GermEval2019 - 3980 + 3031 German tweets
- HASOC2019 - 3819 + 850 German tweets

Twitter data processing

- @username can be masked

Twitter data processing

- @username can be masked
- numbers, urls, dates can be masked

Twitter data processing

- @username can be masked
- numbers, urls, dates can be masked
- typo correction

Twitter data processing

- @username can be masked
- numbers, urls, dates can be masked
- typo correction
- emoticons can be replaced by their textual representations

Twitter data processing

- @username can be masked
- numbers, urls, dates can be masked
- typo correction
- emoticons can be replaced by their textual representations
- #ImportantHashtag - what should we do?

Twitter data processing

- @username can be masked
- numbers, urls, dates can be masked
- typo correction
- emoticons can be replaced by their textual representations
- #ImportantHashtag - what should we do?
  - cut it up by the camel case and remove the #

Twitter data processing

- @username can be masked
- numbers, urls, dates can be masked
- typo correction
- emoticons can be replaced by their textual representations
- #ImportantHashtag - what should we do?
  - cut it up by the camel case and remove the #
  - leave it as is, but remove the #

Twitter data processing

- @username can be masked
- numbers, urls, dates can be masked
- typo correction
- emoticons can be replaced by their textual representations
- #ImportantHashtag - what should we do?
  - cut it up by the camel case and remove the #
  - leave it as is, but remove the #
  - mask it completely

Datasets

- GermEval2018 - 5009 + 3398 German tweets
- GermEval2019 - 3980 + 3031 German tweets
- HASOC2019 - 3819 + 850 German tweets

Datasets

- GermEval2018 - 5009 + 3398 German tweets
- GermEval2019 - 3980 + 3031 German tweets
- HASOC2019 - 3819 + 850 German tweets

Subtasks

Datasets

- GermEval2018 - 5009 + 3398 German tweets
- GermEval2019 - 3980 + 3031 German tweets
- HASOC2019 - 3819 + 850 German tweets

Subtasks

- Binary classification - offense or not

Datasets

- GermEval2018 - 5009 + 3398 German tweets
- GermEval2019 - 3980 + 3031 German tweets
- HASOC2019 - 3819 + 850 German tweets

Subtasks

- Binary classification - offense or not
- Fine-grained classification - offense categories

Datasets

- GermEval2018 - 5009 + 3398 German tweets
- GermEval2019 - 3980 + 3031 German tweets
- HASOC2019 - 3819 + 850 German tweets

Subtasks

- Binary classification - offense or not
- Fine-grained classification - offense categories
- Binary classification - explicit or implicit

GermEval

- **Abuse:** The tweet does not just insult a person but represents the stronger form of abusive language. By abuse we define a special type of degradation. This type of degrading consists in ascribing a social identity to a person that is judged negatively by a (perceived)majority of society. The identity in question is seen as a shameful, unworthy, morally objectionable or marginal identity. E.g. *Ich persönlich scheisse auf die grüne Kinderfickerpartei*

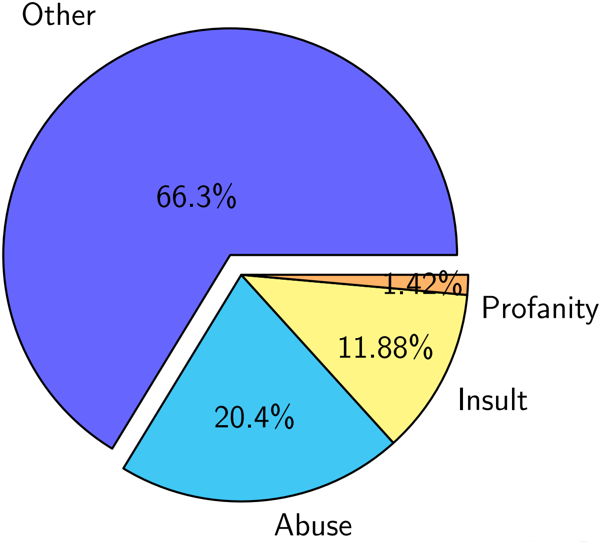
GermEval

- **Abuse:** The tweet does not just insult a person but represents the stronger form of abusive language. By abuse we define a special type of degradation. This type of degrading consists in ascribing a social identity to a person that is judged negatively by a (perceived)majority of society. The identity in question is seen as a shameful, unworthy, morally objectionable or marginal identity. E.g. *Ich persönlich scheisse auf die grüne Kinderfickerpartei*
- **Insult:** The tweet clearly wants to offend someone. E.g. *ein #Tatort mit der Presswurst #Saalfeld geht gar nicht #ARD*

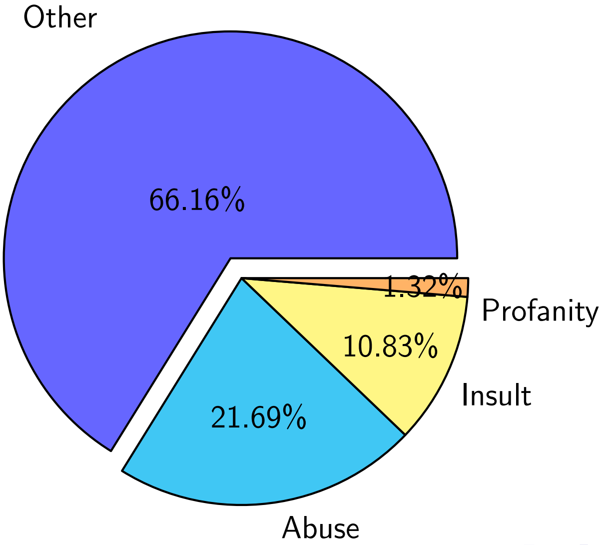
GermEval

- **Abuse:** The tweet does not just insult a person but represents the stronger form of abusive language. By abuse we define a special type of degradation. This type of degrading consists in ascribing a social identity to a person that is judged negatively by a (perceived)majority of society. The identity in question is seen as a shameful, unworthy, morally objectionable or marginal identity. E.g. *Ich persönlich scheisse auf die grüne Kinderfickerpartei*
- **Insult:** The tweet clearly wants to offend someone. E.g. *ein #Tatort mit der Presswurst #Saalfeld geht gar nicht #ARD*
- **Profanity:** Usage of profane words, however, the tweet clearly does not want to insult anyone. E.g. *Juhu, das morgige Wetter passt zum Tag SCHEIßWETTER*

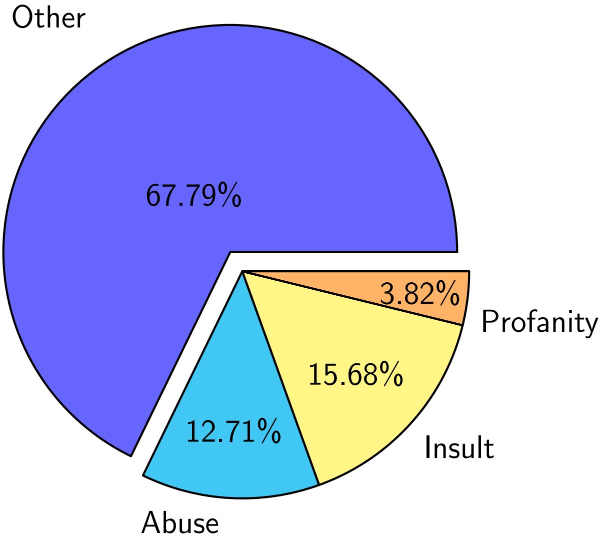
GermEval2018 data distribution



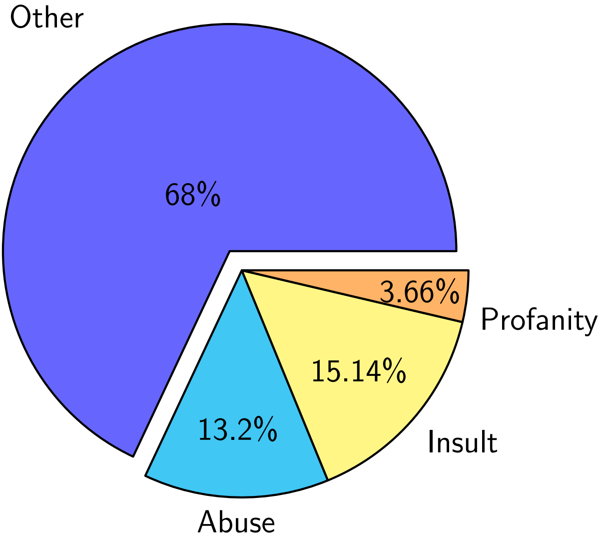
GermEval2018 data distribution



GermEval2019 data distribution



GermEval2019 data distribution



HASOC - Hate Speech and Offensive Content Identification in Indo-European Languages

- **Hate speech:** Describing negative attributes or deficiencies to groups of individuals because they are members of a group (e.g. all poor people are stupid). Hateful comment toward groups because of race, political opinion, sexual orientation, gender, social status, health condition or similar.

HASOC - Hate Speech and Offensive Content Identification in Indo-European Languages

- **Hate speech:** Describing negative attributes or deficiencies to groups of individuals because they are members of a group (e.g. all poor people are stupid). Hateful comment toward groups because of race, political opinion, sexual orientation, gender, social status, health condition or similar.
- **Offensive:** Posts which are degrading, dehumanizing, insulting an individual, threatening with violent acts.

HASOC - Hate Speech and Offensive Content Identification in Indo-European Languages

- **Hate speech:** Describing negative attributes or deficiencies to groups of individuals because they are members of a group (e.g. all poor people are stupid). Hateful comment toward groups because of race, political opinion, sexual orientation, gender, social status, health condition or similar.
- **Offensive:** Posts which are degrading, dehumanizing, insulting an individual, threatening with violent acts.
- **Profanity:** Unacceptable language in the absence of insults and abuse. This typically concerns the usage of swearwords (Scheiße, Fuck etc.) and cursing (Zur Hölle! Verdammt! etc.) are categorized into this category.

HASOC data distribution

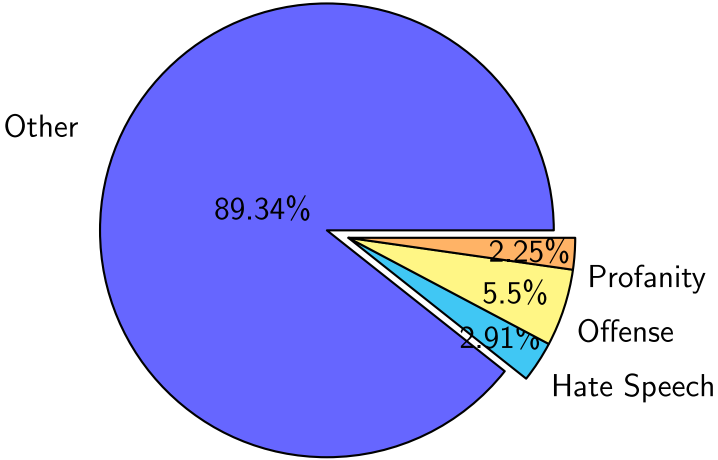


Figure: Train distribution

HASOC data distribution

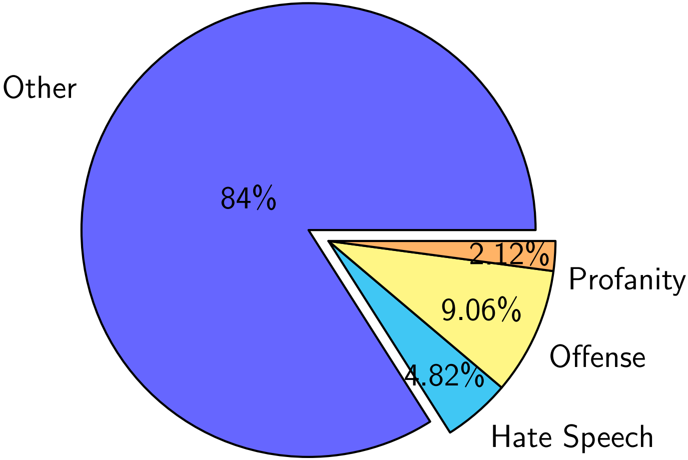


Figure: Test distribution

Leader board on GermEval 2018

Team	Other	Abuse	Insult	Profanity	Average
uhhLT	84.85	53.25	39.46	29.63	**52.71**
TUWienKBS	85.8	52.4	43.71	20.34	**51.42**
uhhLT	84.26	51.96	40.18	15.58	**48.44**
uhhLT	82.88	46.1	21.12	3.92	**43.04**
InriaFBK	83.29	41.34	32.89	4.88	**41.77**

Leader board on GermEval 2019

Team	Other	Abuse	Insult	Profanity	Average
upb	86.57	50.79	38.89	26.21	**53.59**
FoSIL	84.22	49.37	45.2	24	**52.74**
FoSIL	84.95	49.21	42.16	22.7	**52.67**
bertZH	86.66	50.07	44.37	28.27	**52.64**
upb	84.9	49.79	41.37	28.4	**52.48**

Leader board on HASOC 2019

Team	Macro F1	Weighted F1
LSV-UdS	**34.68**	77.49
LSV-UdS	**27.85**	58.29
HateMonitors	**27.69**	75.37
3Idiots	**27.58**	77.79
Cs	**27.4**	75.7

Most popular and successive approaches

- uhhLT (Wiedemann et al., 2018): BiLSTM-CNN model

Most popular and successive approaches

- uhhLT (Wiedemann et al., 2018): BiLSTM-CNN model
- TUWienKBS (Montani, 2018): Word2vec and ensemble machine learning model

Most popular and successive approaches

- uhhLT (Wiedemann et al., 2018): BiLSTM-CNN model
- TUWienKBS (Montani, 2018): Word2vec and ensemble machine learning model
- upb (Paraschiv and Cercel, 2019): pre-trained BERT with last six layers replaced with an output layer after binary classification

Most popular and successive approaches

- uhhLT (Wiedemann et al., 2018): BiLSTM-CNN model
- TUWienKBS (Montani, 2018): Word2vec and ensemble machine learning model
- upb (Paraschiv and Cercel, 2019): pre-trained BERT with last six layers replaced with an output layer after binary classification
- FoSIL (Schmid et al., 2019): FastText with SVM and radial kernel

Most popular and successive approaches

- uhhLT (Wiedemann et al., 2018): BiLSTM-CNN model
- TUWienKBS (Montani, 2018): Word2vec and ensemble machine learning model
- upb (Paraschiv and Cercel, 2019): pre-trained BERT with last six layers replaced with an output layer after binary classification
- FoSIL (Schmid et al., 2019): FastText with SVM and radial kernel
- bertZH (Graf and Salini, 2019): pre-trained BERT classifier

Most popular and successive approaches

- uhhLT (Wiedemann et al., 2018): BiLSTM-CNN model
- TUWienKBS (Montani, 2018): Word2vec and ensemble machine learning model
- upb (Paraschiv and Cercel, 2019): pre-trained BERT with last six layers replaced with an output layer after binary classification
- FoSIL (Schmid et al., 2019): FastText with SVM and radial kernel
- bertZH (Graf and Salini, 2019): pre-trained BERT classifier
- LSV-UdS (Ruiter, Rahman, and Klakow, 2019): 10 fold ensemble BERT classification after binary classification

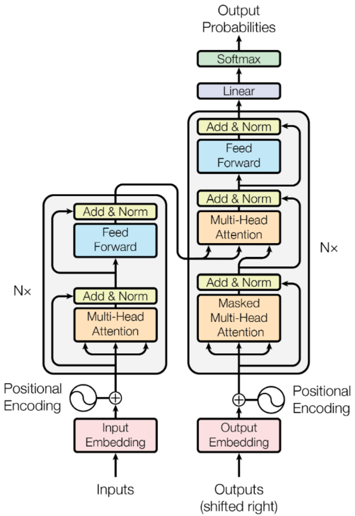
Most popular and successive approaches

- uhhLT (Wiedemann et al., 2018): BiLSTM-CNN model
- TUWienKBS (Montani, 2018): Word2vec and ensemble machine learning model
- upb (Paraschiv and Cercel, 2019): pre-trained BERT with last six layers replaced with an output layer after binary classification
- FoSIL (Schmid et al., 2019): FastText with SVM and radial kernel
- bertZH (Graf and Salini, 2019): pre-trained BERT classifier
- LSV-UdS (Ruiter, Rahman, and Klakow, 2019): 10 fold ensemble BERT classification after binary classification
- HateMonitors (Saha et al., 2019): SVM and Gradient boosted trees

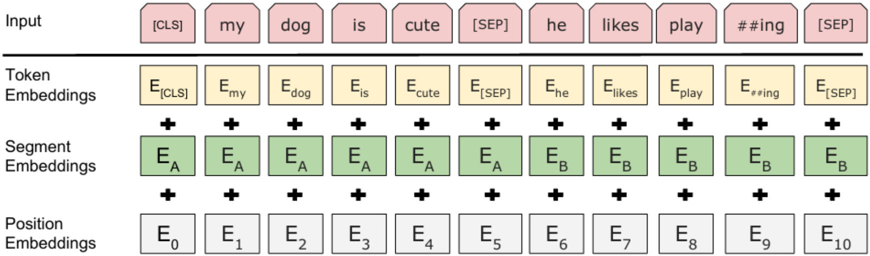
Most popular and successive approaches

- uhhLT (Wiedemann et al., 2018): BiLSTM-CNN model
- TUWienKBS (Montani, 2018): Word2vec and ensemble machine learning model
- upb (Paraschiv and Cercel, 2019): pre-trained BERT with last six layers replaced with an output layer after binary classification
- FoSIL (Schmid et al., 2019): FastText with SVM and radial kernel
- bertZH (Graf and Salini, 2019): pre-trained BERT classifier
- LSV-UdS (Ruiter, Rahman, and Klakow, 2019): 10 fold ensemble BERT classification after binary classification
- HateMonitors (Saha et al., 2019): SVM and Gradient boosted trees
- 3Idiots (Mishra, 2019): BERT classifier

Why is BERT so popular? (Vaswani et al., 2017)



Why is BERT so popular? (Devlin et al., 2019)



Why is BERT so popular? (Devlin et al., 2019)

- WordPiece tokenizer and embeddings with 30,000 token vocabulary

Why is BERT so popular? (Devlin et al., 2019)

- WordPiece tokenizer and embeddings with 30,000 token vocabulary
- Architecture: multi-layer bidirectional Transformer encoder

Why is BERT so popular? (Devlin et al., 2019)

- WordPiece tokenizer and embeddings with 30,000 token vocabulary
- Architecture: multi-layer bidirectional Transformer encoder
  - Large model: 24 Transformer layer

Why is BERT so popular? (Devlin et al., 2019)

- WordPiece tokenizer and embeddings with 30,000 token vocabulary
- Architecture: multi-layer bidirectional Transformer encoder
  - Large model: 24 Transformer layer
  - Base model: 12 Transformer layer

Why is BERT so popular? (Devlin et al., 2019)

- WordPiece tokenizer and embeddings with 30,000 token vocabulary
- Architecture: multi-layer bidirectional Transformer encoder
  - Large model: 24 Transformer layer
  - Base model: 12 Transformer layer
- Pre-training on un-labeled data

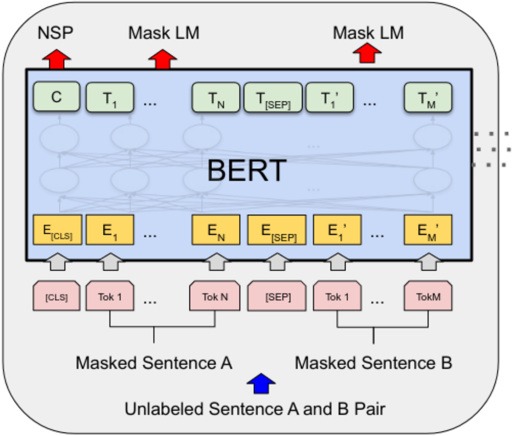
Why is BERT so popular? (Devlin et al., 2019)

- WordPiece tokenizer and embeddings with 30,000 token vocabulary
- Architecture: multi-layer bidirectional Transformer encoder
  - Large model: 24 Transformer layer
  - Base model: 12 Transformer layer
- Pre-training on un-labeled data
  - Masked LM

Why is BERT so popular? (Devlin et al., 2019)

- WordPiece tokenizer and embeddings with 30,000 token vocabulary
- Architecture: multi-layer bidirectional Transformer encoder
  - Large model: 24 Transformer layer
  - Base model: 12 Transformer layer
- Pre-training on un-labeled data
  - Masked LM
  - Next Sentence Prediction

Why is BERT so popular? (Devlin et al., 2019)



Pre-training

Why is BERT so popular?

- BERT (and its relatives) proves to be a strong model on sequence classification and sequence tagging problems

[3] https://huggingface.co/bert-base-multilingual-cased

[4] https://huggingface.co/bert-base-german-cased

Why is BERT so popular?

- BERT (and its relatives) proves to be a strong model on sequence classification and sequence tagging problems
- `bert-base-multilingual-cased`[3]

[3] https://huggingface.co/bert-base-multilingual-cased

[4] https://huggingface.co/bert-base-german-cased

Why is BERT so popular?

- BERT (and its relatives) proves to be a strong model on sequence classification and sequence tagging problems
- `bert-base-multilingual-cased`[3]
- `bert-base-german-cased`[4]

[3] https://huggingface.co/bert-base-multilingual-cased
[4] https://huggingface.co/bert-base-german-cased

Why is BERT so popular?

- BERT (and its relatives) proves to be a strong model on sequence classification and sequence tagging problems
- `bert-base-multilingual-cased`[3]
- `bert-base-german-cased`[4]
  - Published on Jun 14th, 2019

[3] https://huggingface.co/bert-base-multilingual-cased
[4] https://huggingface.co/bert-base-german-cased

Why is BERT so popular?

- BERT (and its relatives) proves to be a strong model on sequence classification and sequence tagging problems
- `bert-base-multilingual-cased`[3]
- `bert-base-german-cased`[4]
  - Published on Jun 14th, 2019
  - Trained on German Wikipedia, OpenLegalData and News data

[3] https://huggingface.co/bert-base-multilingual-cased
[4] https://huggingface.co/bert-base-german-cased

Why is BERT so popular? - Syntactic knowledge (Rogers, Kovaleva, and Rumshisky, 2020)

- BERT representations are hierarchical rather than linear, like syntactic trees

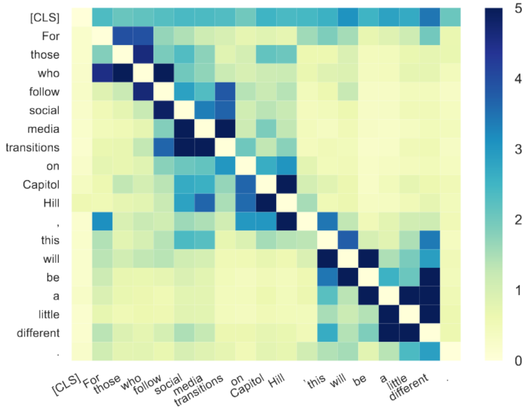
Why is BERT so popular? - Syntactic knowledge (Rogers, Kovaleva, and Rumshisky, 2020)

- BERT representations are hierarchical rather than linear, like syntactic trees
- BERT embeddings encode information about pos, syntactic chunks and roles

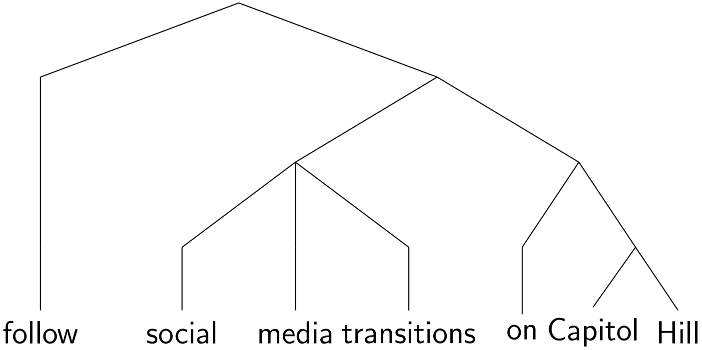
Why is BERT so popular? - Syntactic knowledge (Rogers, Kovaleva, and Rumshisky, 2020)

- BERT representations are hierarchical rather than linear, like syntactic trees
- BERT embeddings encode information about pos, syntactic chunks and roles
- BERT does not store this information in its self-attention weights, but it can be recovered from the token representations

Why is BERT so popular? - Syntactic knowledge (Rogers, Kovaleva, and Rumshisky, 2020)



Why is BERT so popular? - Syntactic knowledge (Rogers, Kovaleva, and Rumshisky, 2020)



Why is BERT so popular? - Semantic knowledge (Rogers, Kovaleva, and Rumshisky, 2020)

- BERT encodes information about semantic roles, entity types, relations

Why is BERT so popular? - Semantic knowledge (Rogers, Kovaleva, and Rumshisky, 2020)

- BERT encodes information about semantic roles, entity types, relations
- BERT struggles with numbers; it does not form a good representation of floating point numbers and fails to generalize

Why is BERT so popular? - Semantic knowledge (Rogers, Kovaleva, and Rumshisky, 2020)

- BERT encodes information about semantic roles, entity types, relations
- BERT struggles with numbers; it does not form a good representation of floating point numbers and fails to generalize
- BERT does not form a generic idea of named-entities

Why is BERT so popular? - World knowledge (Rogers, Kovaleva, and Rumshisky, 2020)

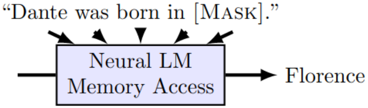
- BERT can be competitive with methods relying on knowledge bases for some relation types

Why is BERT so popular? - World knowledge (Rogers, Kovaleva, and Rumshisky, 2020)

- BERT can be competitive with methods relying on knowledge bases for some relation types
- BERT struggles with pragmatic inference, role-based event knowledge, and abstract attributes of objects

Why is BERT so popular? - World knowledge (Rogers, Kovaleva, and Rumshisky, 2020)

- BERT can be competitive with methods relying on knowledge bases for some relation types
- BERT struggles with pragmatic inference, role-based event knowledge, and abstract attributes of objects
- BERT cannot reason based on world-knowledge



Twitter data processing for BERT

- @username $\rightarrow$ [USER]

Twitter data processing for BERT

- @username → [USER]
- numbers → [NUM], urls → [URL], dates → [DATE]

Twitter data processing for BERT

- @username → [USER]
- numbers → [NUM], urls → [URL], dates → [DATE]
- emoticons should be replaced by their textual representations because of the WordPiece tokenizer

Twitter data processing for BERT

- @username $\rightarrow$ [USER]
- numbers $\rightarrow$ [NUM], urls $\rightarrow$ [URL], dates $\rightarrow$ [DATE]
- emoticons should be replaced by their textual representations because of the WordPiece tokenizer
- #ImportantHashtag $\rightarrow$ cut it up by the camel case and remove the #

Thank You for Your Attention!

Devlin, Jacob et al. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv: 1810.04805 [cs.CL].

Graf, Tim and Luca Salini (2019). "bertZH at GermEval 2019: Fine-Grained Classification of German Offensive Language using Fine-Tuned BERT". In: *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*. Erlangen, Germany: German Society for Computational Linguistics & Language Technology, pp. 434–437.

Mishra, Shubhanshu (2019). "3Idiots at HASOC 2019: Fine-tuning Transformer Neural Networks for Hate Speech Identification in Indo-European Languages". In: *FIRE*.

Montani, J. P. (2018). "GermEval 2018: German Abusive Tweet Detection". In: *Proceedings of the GermEval 2018 Workshop*.

Paraschiv, Andrei and Dumitru-Clementin Cercel (2019). "UPB at GermEval-2019 Task 2: BERT-Based Offensive Language Classification of German Tweets". In: *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*. Erlangen, Germany: German Society for Computational Linguistics & Language Technology, pp. 398–404.

Rogers, Anna, Olga Kovaleva, and Anna Rumshisky (2020). "A Primer in BERTology: What We Know About How BERT Works". In: *Transactions of the Association for Computational Linguistics* 8, pp. 842–866. DOI: 10.1162/tacl_a_00349. URL: https://www.aclweb.org/anthology/2020.tacl-1.54.

Ruiter, Dana, Md. Ataur Rahman, and D. Klakow (2019). "LSV-UdS at HASOC 2019: The Problem of Defining Hate". In: *FIRE*.

Saha, Punyajoy et al. (2019). *HateMonitors: Language Agnostic Abuse Detection in Social Media*. arXiv: 1909.12642 [cs.SI].

Schmid, Florian et al. (2019). "FoSIL - Offensive language classification of German tweets combining SVMs and deep learning techniques". In: *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*. Erlangen, Germany: German Society for Computational Linguistics & Language Technology, pp. 382–386.

Vaswani, Ashish et al. (2017). *Attention Is All You Need*. arXiv: 1706.03762 [cs.CL].

Wiedemann, Gregor et al. (2018). *Transfer Learning from LDA to BiLSTM-CNN for Offensive Language Detection in Twitter*. arXiv: 1811.02906 [cs.CL].