Hate Speech Detection

Group 15

Carsten Gieshoff, Zhaolin Li, David Pomerenke

Problem

Problem

Hate Speech

@JuanYeez shut yo beaner ass up spic and hop your faggot ass back across the border little nigga

@MDreyfus @NatFascist88 Shit your ass your moms pussy u Jew bastard. Ur times coming. Heil Hitler!

My advice of the day: If your a tranny… go fuck your self!

Offensive Language

When you realize how curiosity is a bitch #CuriosityKilledMe

Why you worried bout that other hoe? Cuz that other hoe aint worried bout another hoe

I knew Kendrick Lamar was onto something when he said "I call a bitch a bitch, a hoe a hoe, a woman a woman"

Problem

Hate Speech

@JuanYeez shut yo beaner ass up spic and hop your faggot ass back across the border little nigga

@MDreyfus @NatFascist88 Shit your ass your moms pussy u Jew bastard. Ur times coming. Heil Hitler!

My advice of the day: If your a tranny... go fuck your self!

Offensive Language

When you realize how curiosity is a bitch #CuriosityKilledMe

Why you worried bout that other hoe? Cuz that other hoe aint worried bout another hoe

I knew Kendrick Lamar was onto something when he said “I call a bitch a bitch, a hoe a hoe, a woman a woman”

HateBase Data Set

Approach

1. Improve existing approach (Davidson & al 2017)
    - Word-based vs character-based tokenization
    - Different machine learning models
    - Different weighting of categories

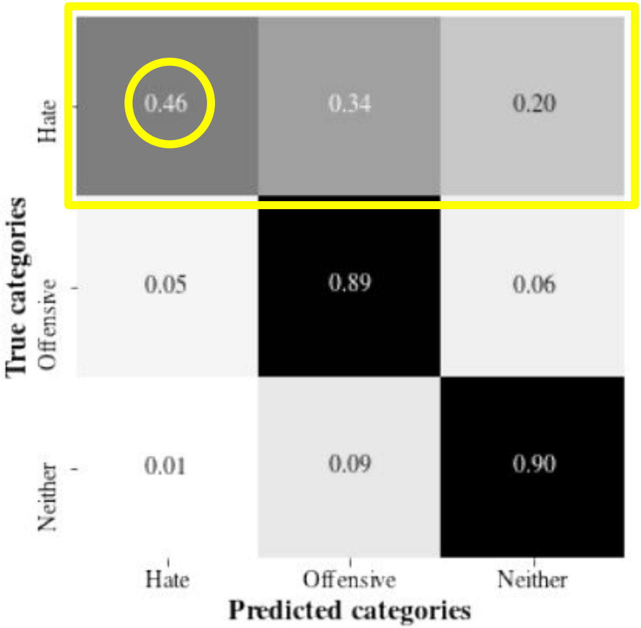
Approach

1. Improve existing approach (Davidson & al 2017)
   - Word-based vs character-based tokenization
   - Different machine learning models
   - Different weighting of categories
2. Build classifier with transformer language model
   - BERT
   - BERTweet (RoBERTa)

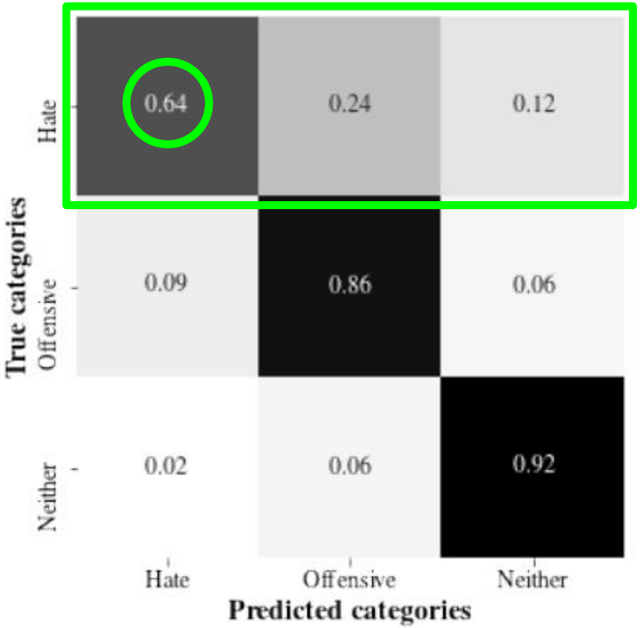
Approach

1. Improve existing approach (Davidson & al 2017)
   - Word-based vs character-based tokenization
   - Different machine learning models
   - Different weighting of categories
2. Build classifier with transformer language model
   - BERT
   - BERTweet (RoBERTa)
3. Transfer classification from English to other languages
   - mBART-50 for translation from Spanish, Italian, Turkish
   - Evaluation on detoxify data set

Results: Different Weighting

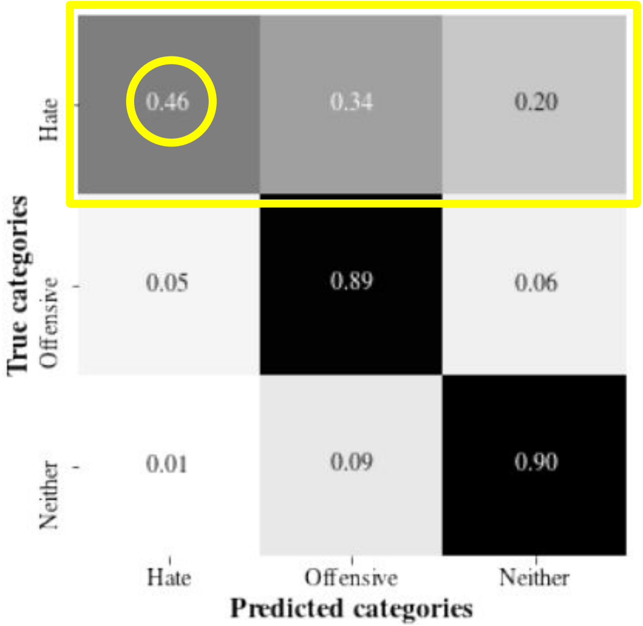


(a) $E = 1$

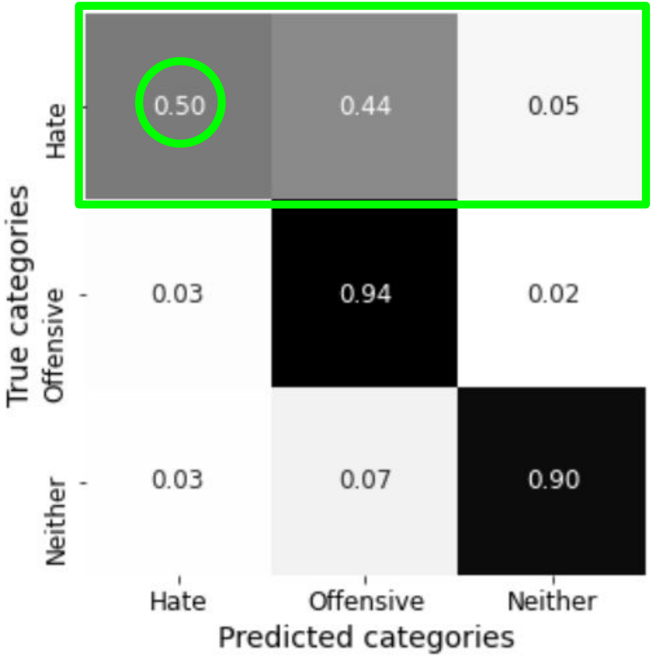


(c) $E = 1.5$

Results: Original vs BERT

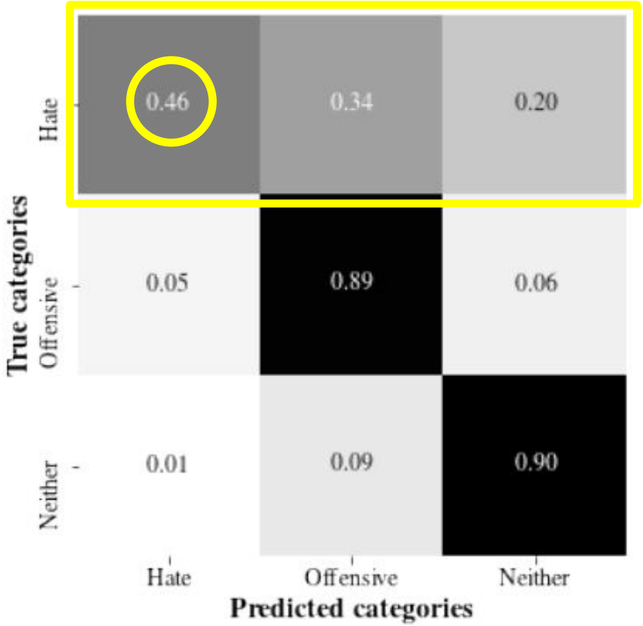


True categories \ Predicted categories	Hate	Offensive	Neither
Hate	0.46	0.34	0.20
Offensive	0.05	0.89	0.06
Neither	0.01	0.09	0.90



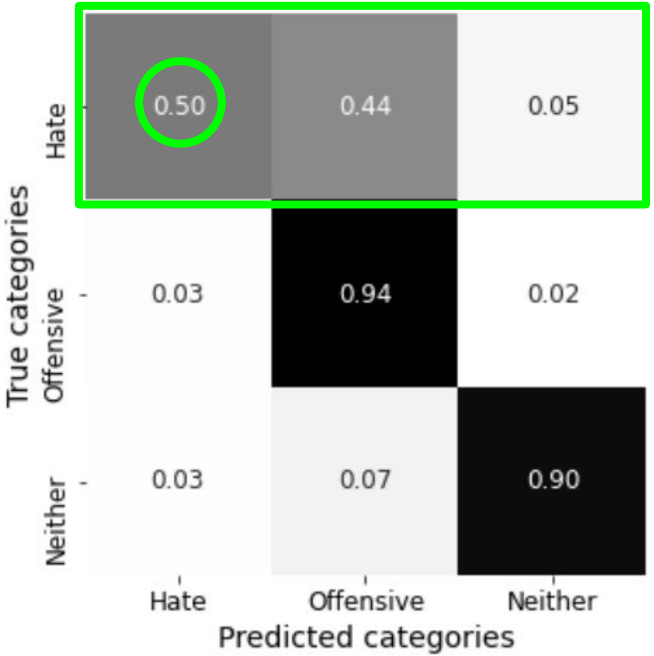
True categories \ Predicted categories	Hate	Offensive	Neither
Hate	0.50	0.44	0.05
Offensive	0.03	0.94	0.02
Neither	0.03	0.07	0.90

Results: Original vs BERT



True categories \ Predicted categories	Hate	Offensive	Neither
Hate	0.46	0.34	0.20
Offensive	0.05	0.89	0.06
Neither	0.01	0.09	0.90

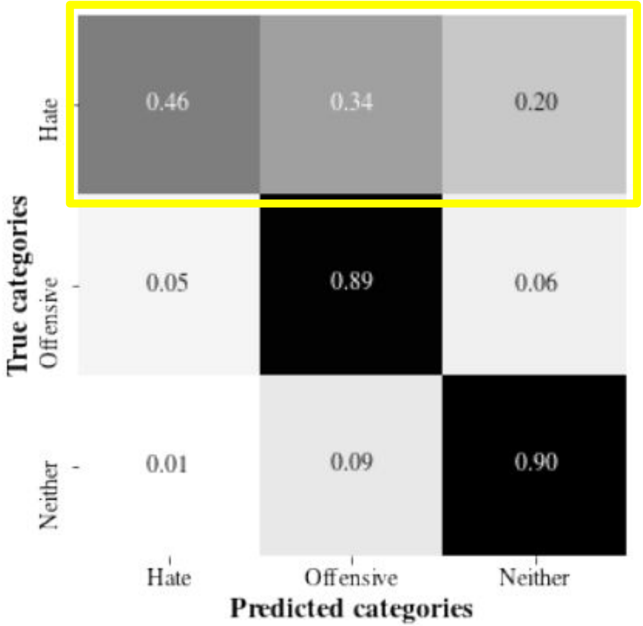
84% accuracy



True categories \ Predicted categories	Hate	Offensive	Neither
Hate	0.50	0.44	0.05
Offensive	0.03	0.94	0.02
Neither	0.03	0.07	0.90

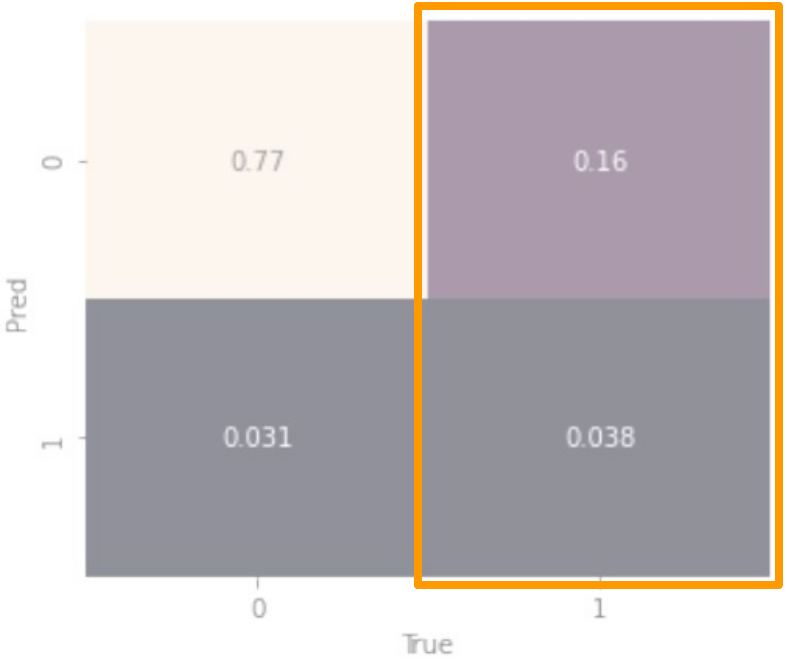
91% accuracy

Results: Original vs transfer to Italian



True categories \ Predicted categories	Hate	Offensive	Neither
Hate	0.46	0.34	0.20
Offensive	0.05	0.89	0.06
Neither	0.01	0.09	0.90

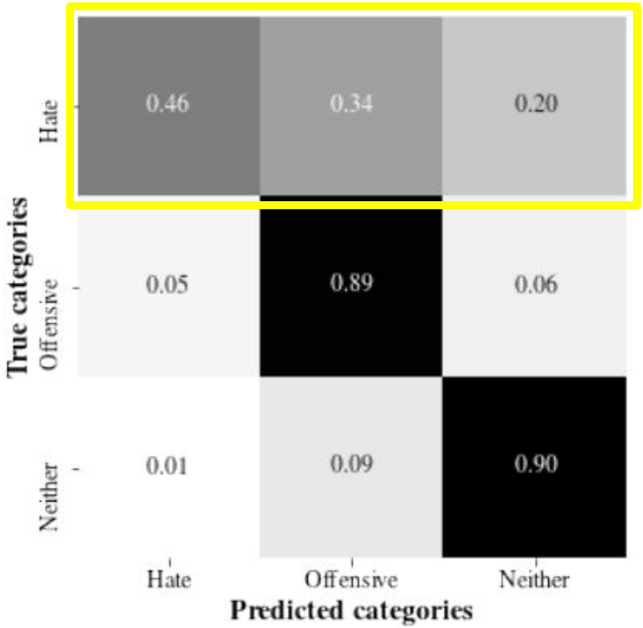
84% accuracy



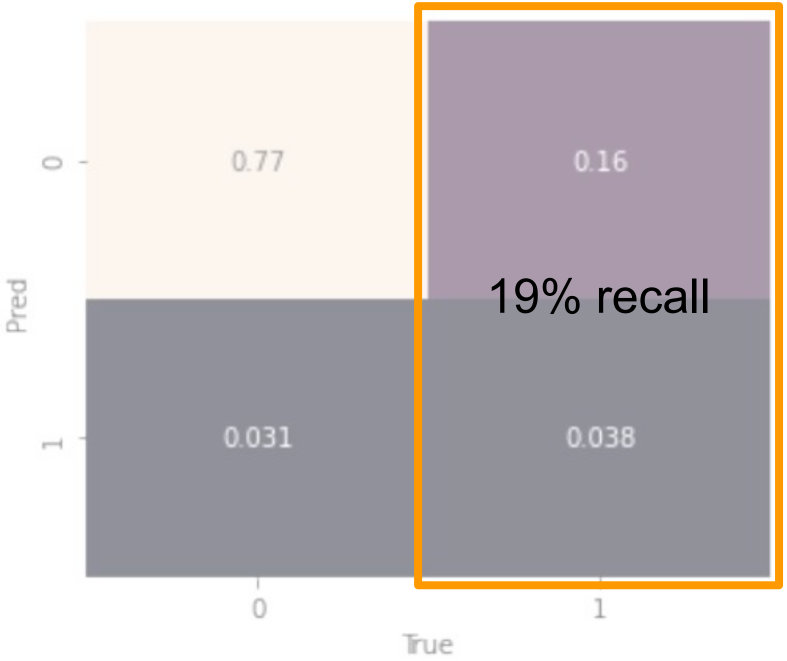
Pred \ True	0	1
0	0.77	0.16
1	0.031	0.038

81% accuracy

Results: Original vs transfer to Italian



84% accuracy



19% recall

81% accuracy

Hate Speech Detection

Carsten Gieshoff, Zhaolin Li, David Pomerenke

carstengieshoff/HS-Detection-Project

- We improve hate speech recall by aggressive weighting
- Our BERT-based model beats the baseline from 2017
- We perform multilingual detection based on only English training data
- Future work: Improving recall using data augmentation