HS-EMO: Analyzing Emotions in Hate Speech

Johannes Schäfer and Elina Kistner

Institute for Information Science and Natural Language Processing
University of Hildesheim
johannes.schaefer@uni-hildesheim.de

Hate Speech and Emotions

- Related work: hate speech correlates with negative emotions [1]
- This work: empirical analysis with joint learning as goal
  1. Annotation of hate speech data with emotion labels
  2. Analysis of annotated dataset (1,000 instances)
  3. Joint learning experiment

Data

- Initial dataset: HASOC 2021 [2] (3,843 English Twitter posts), annotated for **Hate and Offensive (HOF) categories** vs. NONE:
  - PRFN: profanity, curse words
  - OFFN: offensive language (such as insults of individuals)
  - HATE: hate speech content (attack because of group membership)
- **Annotation of emotion categories** [3]: Anger, Disgust, Sadness, Joy, Fear, Surprise
- Examples:

Anger — HATE
„After killing half the world with #WuhanVirus these morons now blaming us“

Surprise — HATE
„What the heck the bjp is doing... destroying people life? #ResignPMmodi #BjpDestroyedIndia #BJP #prayaraj“

Joy — OFFN
„C'mon Twitter.. do your thing. Make this greedy twat so embarrassed he hates to go to a game again.“

Disgust — OFFN
„@ndtv Shameless PM. What else can we say? #ShameOnModi #Resign_PM_Modi #ResignPMmodi“

Sadness — PRFN
„@Andrea Sorry to hear you're having a tough enough time, then some twat makes it worserer (I'm sure that's a word) Stay strong Andrea 🙏“

Fear — HATE
„This is the first time in my life I am not feeling safe in my own house in my own land and can't do anything for it such a shame to peoples who says they are leaders of our country #CovidIndia #IndiaCovidCrisis“

HS-EMO corpus (1,000 instances)

Emotions in HOF vs. NONE:

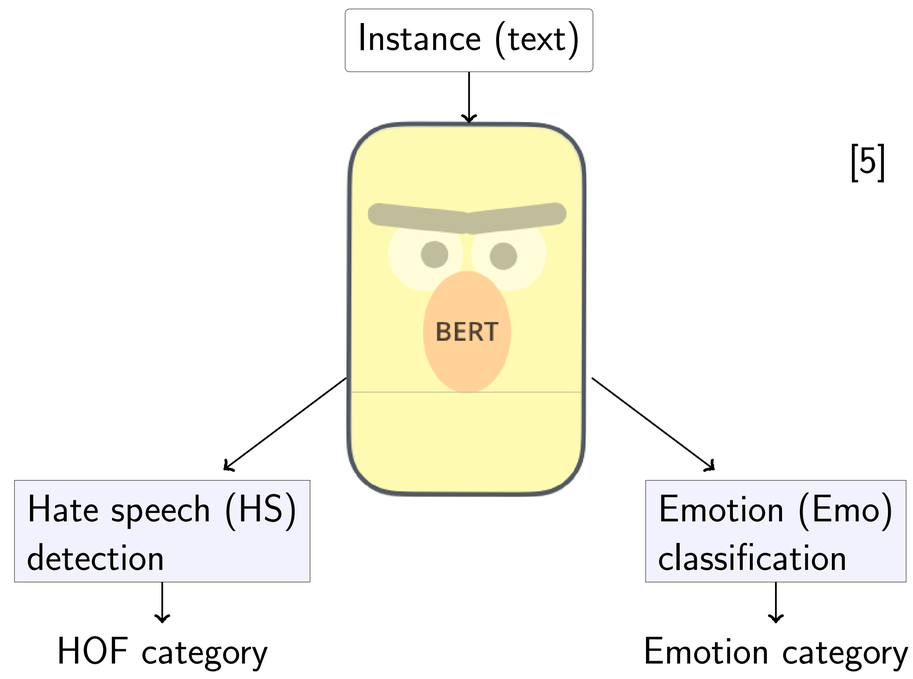
	Anger	Disgust	Sadness	Joy	Fear	Surprise
HOF	278 (43%)	139 (21%)	47 (7%)	78 (12%)	30 (5%)	40 (6%)
NONE	74 (21%)	33 (9%)	111 (32%)	35 (10%)	49 (14%)	22 (6%)

Emotions in HOF categories:

	Anger	Disgust	Sadness	Joy	Fear	Surprise
PRFN	136 (44%)	33 (11%)	14 (5%)	65 (21%)	8 (3%)	26 (8%)
OFFN	75 (46%)	48 (29%)	9 (5%)	11 (7%)	9 (5%)	7 (4%)
HATE	67 (38%)	58 (33%)	24 (14%)	2 (1%)	13 (7%)	7 (4%)

Joint classification experiment

- 3 datasets: HASOC (HOF), TEC [4] (Emo), HS-EMO (HOF+Emo)
- Motivation: learning common features



Training settings:

1. HASOC (Baseline 1)
2. HASOC and TEC alternately (Baseline 2)
3. HASOC, **HS-EMO**
4. HASOC and TEC alternately, **HS-EMO**

Results:

Training Data	F1 *NONE*	F1 *PRFN*	F1 *OFFN*	F1 *HATE*	macro-avg F1
1. HASOC (Baseline 1)	.7018	**.7827**	**.5121**	.5438	**.6351**
2. HASOC, TEC (Baseline 2)	.7100	.7143	.4054	.5436	.5933
3. HASOC, HS-EMO	.7154	.7315	.4509	**.5655**	.6158
4. HASOC, TEC, HS-EMO	**.7169**	.7522	.4823	.5620	.6283

Conclusion

Analysis of emotions in hate speech:

- **Correlation** of hate speech with some negative emotions (especially Anger and Disgust)
- For some other negative emotions (such as Sadness): **no correlation** with offensive language/hate speech;
  two cases of hate speech with Joy emotion

Joint learning experiment:

- Some promising results
  (improvement over classical multi-task learning with separate datasets)
- Failed to show that emotion annotation can help learning hate speech detection

Future work:

- Annotation of full dataset, preliminary results:
  similar distribution of emotion categories in full corpus
- Multiple annotations (evaluation of annotation quality), preliminary results (4 annotators, 2 per instance): Cohen's $\kappa \approx .33$ and .38

This work: Johannes Schäfer and Elina Kistner. (2023). HS-EMO: Analyzing Emotions in Hate Speech. In: KONVENS 2023.
Data and code: https://github.com/Johannes-Schaefer/HS-EMO.

References

[1] Alorainy, W. et al. (2018). Suspended accounts: A source of tweets with disgust and anger emotions for augmenting hate speech data sample.
[2] Mandl, T. et al. (2021). Overview of the HASOC Subtrack at FIRE 2021: HateSpeech and Offensive Content Identification in English and Indo-Aryan Languages.
[3] Ekman, P. (1988). Gesichtsausdruck und Gefühl: 20 Jahre Forschung von Paul Ekman.
[4] Mohammad, S. (2012). #Emotional Tweets. http://saifmohammad.com/WebPages/SentimentEmotionLabeledData.html.
[5] Alammar, J. (2018). The Illustrated Transformer. Retrieved from https://jalammar.github.io/illustrated-bert/.