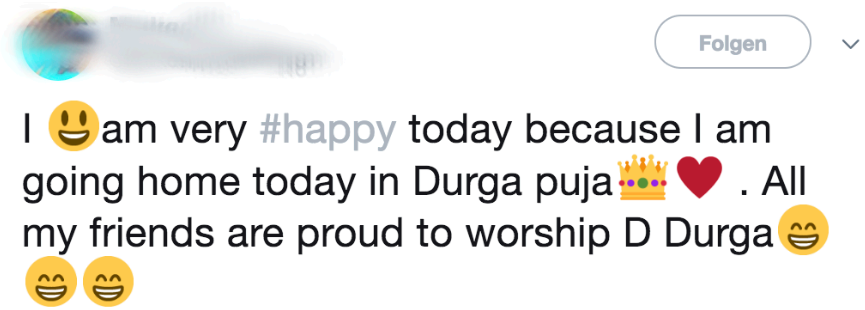
University of Stuttgart
Institute for
Natural Language Processing

IEST Implicit Emotions Shared Task

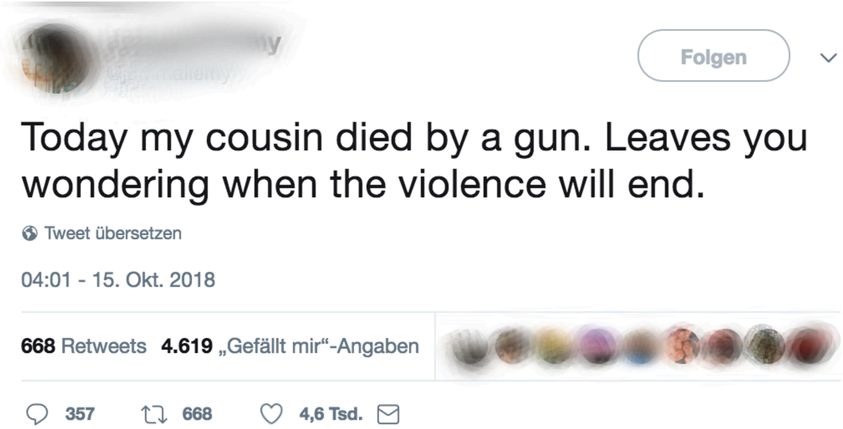
WASSA/EMNLP 2018

October 31, 2018

Roman Klinger, Orphée De Clercq,
Saif M. Mohammad, Alexandra Balahur


[redacted]

I 😃am very #happy today because I am going home today in Durga puja👑❤️ . All my friends are proud to worship D Durga😁 😁😁



[redacted]

Today my cousin died by a gun. Leaves you wondering when the violence will end.

04:01 - 15. Okt. 2018

668 Retweets **4.619** „Gefällt mir“-Angaben

Goal

How well can emotion prediction models work when they are forced to ignore (most of the) explicit emotion cues?

Outline

1. Background
2. Task Definition
3. Results
4. Human Annotation Experiment
5. Conclusion
6. Best System Analysis Award

Outline

1 Background

2 Task Definition

3 Results

4 Human Annotation Experiment

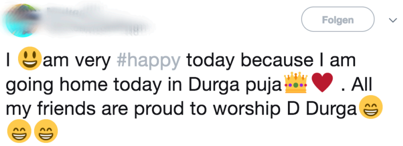
5 Conclusion

6 Best System Analysis Award

Idea

- Emotion prediction in most systems = classification of sentences or documents

$f\Big($



I 😃am very #happy today because I am going home today in Durga puja👑❤️ . All my friends are proud to worship D Durga😁 😁😁

$\Big) \rightarrow$

- We presume: Systems overfit to explicit trigger words
- Issue with generalization: Given an event implicitly associated to an emotion, classification might not work

Background: ISEAR

International Survey On Emotion Antecedents and Reactions

Questionaire

- `Emotion: …`
  `Please describe a situation or event -- in as much detail as possible -- in which you felt the emotion given above.`
- Joy, Fear, Anger, Sadness, Disgust, Shame, Guilt

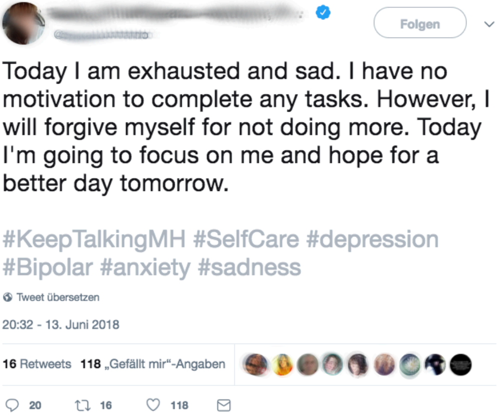
⇒ Focus on events

⇒ Many instances do not contain emotion words

⇒ 7665 instances

Data-Hungry Algorithms

- Classification algorithms today use high numbers of parameters
- Manual annotation is tedious and expensive
- One established approach: Self-labeling by authors with hashtags or emoticon



[redacted]

Folgen

Today I am exhausted and sad. I have no motivation to complete any tasks. However, I will forgive myself for not doing more. Today I'm going to focus on me and hope for a better day tomorrow.

#KeepTalkingMH #SelfCare #depression #Bipolar #anxiety #sadness

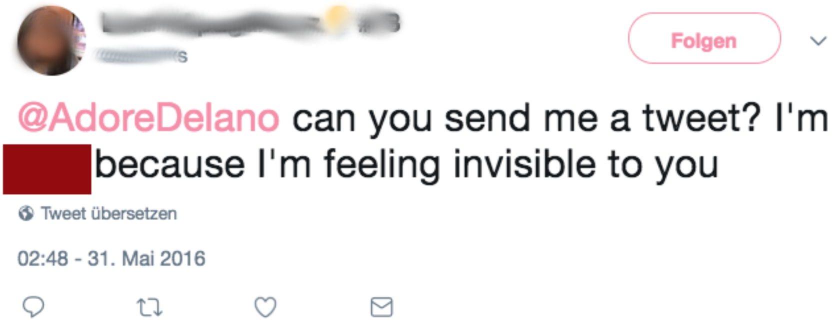
Tweet übersetzen

20:32 - 13. Juni 2018

16 Retweets 118 „Gefällt mir"-Angaben

20 16 118

Idea: Distant Labeling with Event Focus



[redacted]
[redacted]

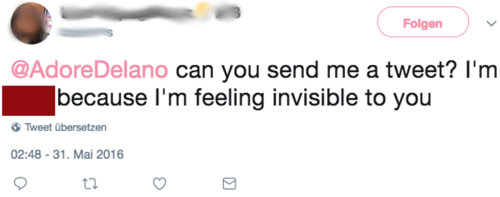
@AdoreDelano can you send me a tweet? I'm [redacted] because I'm feeling invisible to you

02:48 - 31. Mai 2016

Outline

1. Background
2. Task Definition
3. Results
4. Human Annotation Experiment
5. Conclusion
6. Best System Analysis Award

Task Definition



[redacted] [redacted]

@AdoreDelano can you send me a tweet? I'm [redacted] because I'm feeling invisible to you

02:48 - 31. Mai 2016

- Input:
  Tweet with emotion synonym replaced by unique string
- Output:
  Emotion for which the removed work is a synonym

Example

`sadness [USERNAME] can you send me a tweet? I'm [#TRIGGERWORD#] because I'm feeling invisible to you`

Data and Task Setting

- Query API for `EMOTIONWORD (that|when|because)`
- Emotion words:
  - Anger: angry, furious
  - Fear: afraid, frightened, scared, fearful
  - Disgust: disgusted, disgusting
  - Joy: cheerful, happy, joyful
  - Sadness: sad, depressed, sorrowful
  - Surprise: surprising, surprised, astonished, shocked, startled, astounded, stunned
- Stratified sampling, no tweets with >1 emotion words
- Train: 153383, Trial: 9591, Test: 28757 instances
- Evaluation: Macro F_1
- MaxEnt Bag-of-Words Baseline

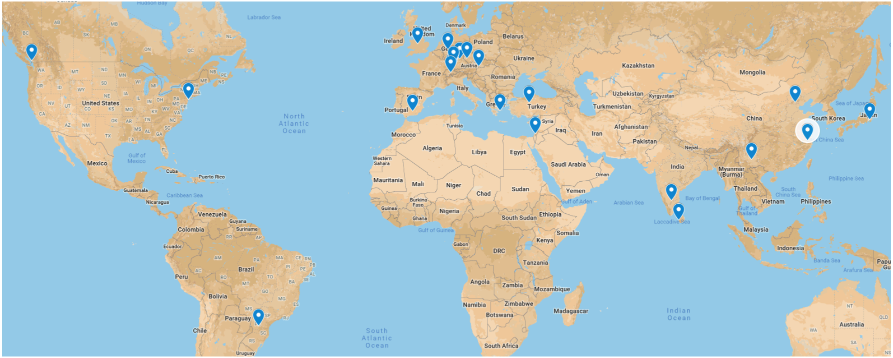
Outline

1. Background
2. Task Definition
3. Results
4. Human Annotation Experiment
5. Conclusion
6. Best System Analysis Award

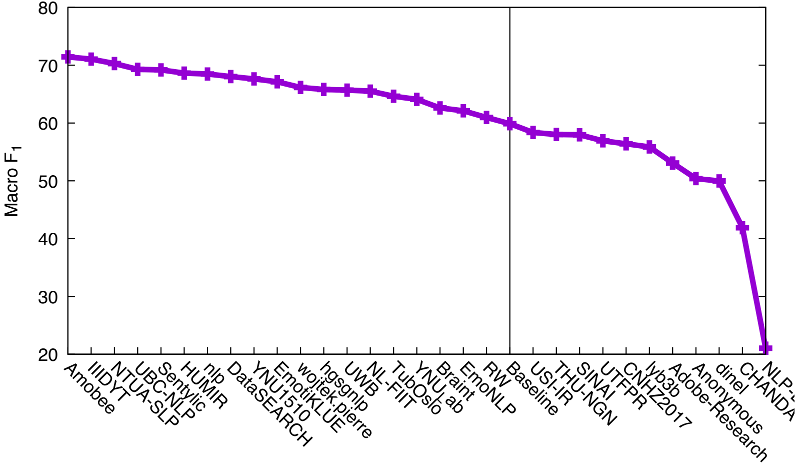
Participants

- 107 expressions of interest
- 30 valid submissions
- 26 short system descriptions
- 21 paper submissions
- 19 paper acceptances

Participants



Results


Macro F_1
80
70
60
50
40
30
20
Amobee
IIIDYT
NTUA-SLP
UBC-NLP
Sentylic
HUMIR
nlp
DataSEARCH
YNU1510
EmotiKLUE
wojtek.pierre
hgsgnlp
UWB
NL-FIIT
TubOslo
YNU Lab
Braint
EmoNLP
RW
Baseline
USI-IR
THU-NGN
SINAI
UTFPR
CNHZ2017
lyb3b
Adobe-Research
Anonymous
dinel
CHANDA
NLP-

Tools

- Deep learning:
  - Keras, Tensorflow
  - PyTorch of medium popularity
  - Theano only once
- Data processing, general ML:
  - NLTK, Pandas, ScikitLearn
  - Weka and SpaCy of lower popularity
- Embeddings/Similarity measures:
  - GloVe, GenSim, FastText
  - ElMo less popular

Methods

- Nearly everybody used embeddings
- Nearly everybody used recurrent neural networks (LSTM/GRU/RNN)
- Most top teams used ensembles (8/9)
- CNNs distributed ≈ equally across ranks
- Attention mechanisms 5/9 top, not by lower ranked teams
- Language models used by 3/4 top teams

Error Analysis

Anger, all teams correct

Anyone have the first fast and TRIGGER that I can borrow?

Anger, nobody correct

I’m kinda TRIGGER that I have to work on Father’s Day

Error Analysis

Disgust, all teams correct

nyc smells TRIGGER when it's wet.

Disgust, nobody correct

I wanted a cup of coffee for the train ride. Got ignored twice. I left TRIGGER because I can't afford to miss my train. #needcoffee :(

Error Analysis

Joy, all teams correct

maybe im so unTRIGGER because i never see the sunlight?

Joy, nobody correct

I am actually TRIGGER when not invited to certain things. I don't have the time and patience to pretend

Outline

1. Background
2. Task Definition
3. Results
4. Human Annotation Experiment
5. Conclusion
6. Best System Analysis Award

Human Annotation Experiment: Setting

- 900 instances:
  - 50 tweets for each of 6 emotions
  - 18 pair-wise combinations with because, that, when
- Questionaire
  - Figure-Eight (previously known as CrowdFlower)
  - Question 1: Best guess for emotion
  - Question 2: Other guesses for emotion
- 3619 judgements
- 3 annotators at least for each instance

Human Annotation Results

	Human	Baseline
Human Q1	47	54
Human Q2	57	
“because”	51	50
“when”	49	53
“that”	41	60
Anger	46	41
Disgust	21	51
Fear	51	58
Joy	58	60
Sadness	52	58
Surprise	34	58

Humans confuse:

- Disgust and Fear
- Fear and Sadness
- Surprise and Anger/Joy

Outline

1. Background
2. Task Definition
3. Results
4. Human Annotation Experiment
5. Conclusion
6. Best System Analysis Award

Conclusion

- Shared task with substantial participation
- Team results well distributed across performance spectrum
- Best teams: Ensembles, Deep Learning, Fine-tuning to tasks

Criticism and Future Work

- Data retrieval partially pretty noise
    - “Fast and Furious”, “unhappy”

  ⇒ Improve retrieval
- Results better than human performance

  ⇒ Manual annotation of data sets
- Assumption still unproven
    - Do these models generalize better to implicit statements?
    - Could this data be used for adversarial optimization of models on other data sets?

Winners

Rank of Submissions

- Rank 1: Amobee at IEST 2018: Transfer Learning from Language Models (71.45)
- Rank 2: IIIDYT at IEST 2018: Implicit Emotion Classification With Deep Contextualized Word Representations (71.05)
- Rank 3: NTUA-SLP at IEST 2018: Ensemble of Neural Transfer Methods for Implicit Emotion Classification (70.29)

Best System Analysis

IIIDYT at IEST 2018: Implicit Emotion Classification With Deep Contextualized Word Representations

IEST Implicit Emotions Shared Task

😀😩😡😮😨😖🤔

Thank you!