Convincingness of Emotional Argumentation

Practical Big Data Science

Chair of Database Systems and Data Science

Team Orange: Johanna Reiml, Shaoying Liu, Petros Isaakidis, Severin Schmid, Isabelle Mayerhofer

21.07.2021

Wissenschaftlicher Betreuer:
Michael Fromm, Evgeniy Faerman
Verantwortlicher Professor:
Prof. Dr. Thomas Seidl

DBS

Agenda

- Motivation and research questions
- Methods:
  The datasets and preprocessing
  Manual data annotation
  The emotion detection framework
- Results:
  Training and evaluation
- Analysis and summary

Motivation and research questions

Motivation

Combine **Emotion Detection (ED)** and **Argument Mining (AM)**

ED: Method to detect emotions in sentences [1]

AM: Detect arguments and their quality [2]

Possible areas of application:

- Argument search engine
- Detect emotions in political speeches to persuade/influence audience
- Detect where emotions and facts start to blurr

Motivation

When analyzing arguments, we usually focus on **logos**

In our case, we try to focus on **pathos**

[9]

Our approach:

- Train a model based on emotion detection datasets
- Apply it to argument mining datasets

Approach overview

Training:

Sentences from ED dataset

Pre-trained BERT model [3]

Prediction:

Sentences from AM dataset

BERT trained on Emotion Detection [3]

Output:
Sentence is ...
- Non-Emotional/Emotional
- Neutral/Positive/Negative

Research questions

Hypotheses:

- We can build a neural network that can detect emotions
- Some topics are more emotional than others
- There exists a correlation between stance and emotion
- There exists a positive correlation between argumentativeness and emotion intensity

The datasets and preprocessing

The Emotion Detection datasets

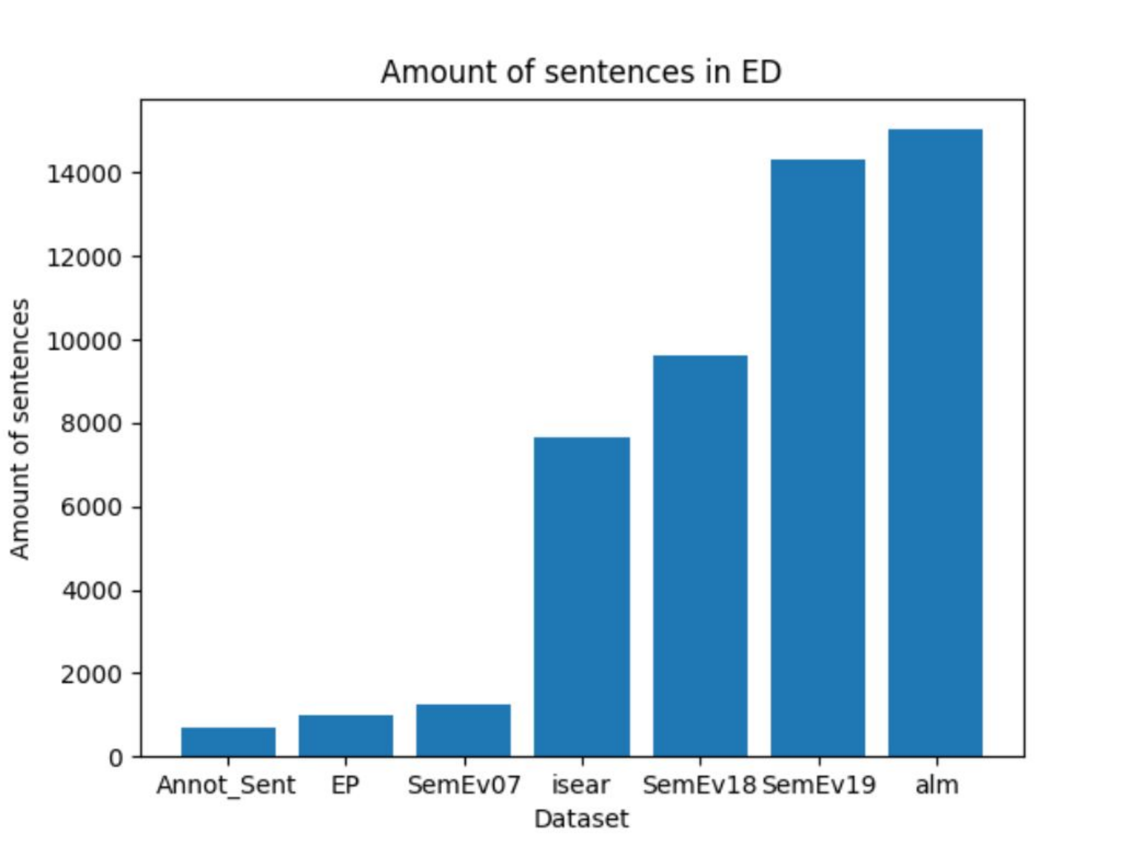
Form: Children's stories, reactions, news headlines, tweets, dialogues

Children's stories: "I well know your desire, said the cat"

Headlines: Bombers kill shoppers

Tweets: My heart is so happy I want to explode. ❤️💥, This time in 2 weeks I will be 30... 😥

The Emotion Detection datasets



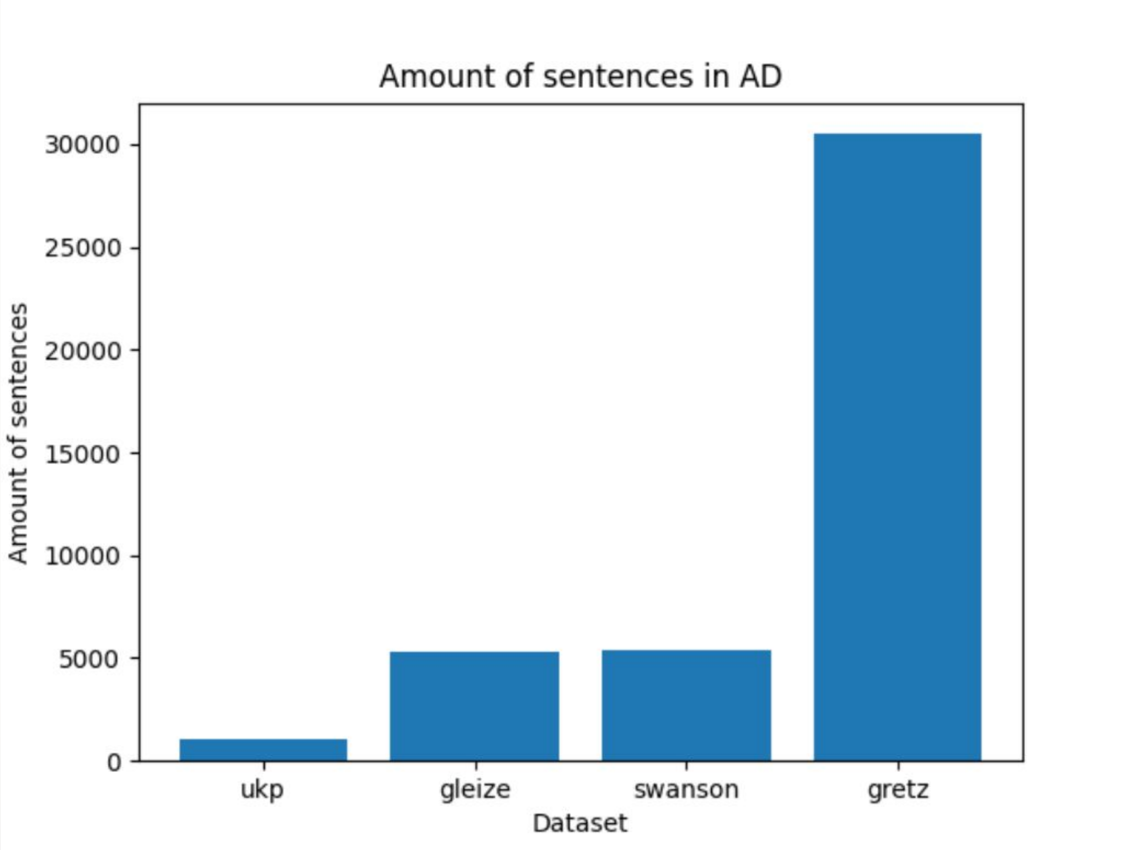
The Argument Mining datasets

Topics covered: evolution, gun control, gay marriage, Firefox vs. Internet Explorer, abortion...

Gay marriage: If it walks like a marriage, talks like a marriage, and looks like a marriage, why call it something else?

Firefox vs. Internet Explorer: Firefox has nothing to do with Apple or Steve Jobs.

The Argument Mining datasets



Preprocessing and annotation scheme

Data cleaning & integration: Cleaning of datasets, e.g. deleting emojis (via RegEx), transform different formats (txt, tar etc.) to csv

Annotation scheme:

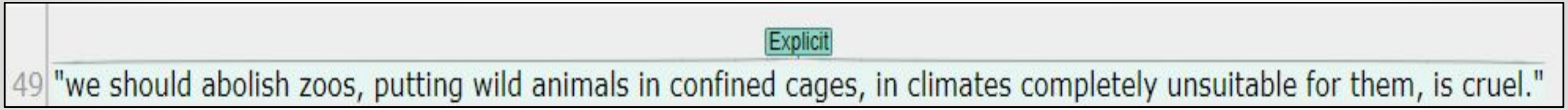
- ED datasets:
  - 2-class setting: emotional vs. non-emotional
  - 3-class setting: positive, neutral, negative
- AM datasets: argument, arg-strength, topic

Labeling sentences

INCEpTION

[3]

Each of us labeled 600 arguments for emotionality



Rules:

- ‘explicit’ emotional arguments express author’s emotionality
- ‘implicit’ emotional arguments target reader’s emotions
- ‘neutral’ arguments are neither explicit nor implicit

Calculating the annotation agreement

Agreement table

document_id	annotation		
	Explicit	Implicit	Neutral
0	3	3	0
1	0	2	4
2	1	1	4
...	...	...	...
599	0	1	5

document_id	annotation	
	Emotional	Non-emotional
0	6	0
1	2	4
2	2	4
...	...	...
599	1	5

Krippendorff's Alpha
0.245 - for implicit/explicit/neutral
0.313 - for emotional/non-emotional

Emotion Detection framework

Key facts about BERT

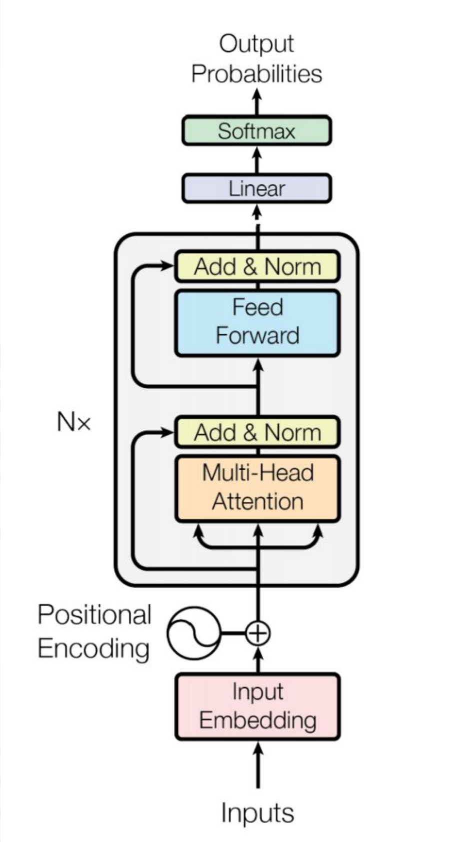
Bidirectional **E**ncoder **R**epresentations from **T**ransformers:

- Bidirectionally contextual model
- Introduces new self-supervised objective(s)
- Completely replaces recurrent architectures by Self-Attention + is also simultaneously able to include bidirectionality

[3]

BERT model architecture

1. Input is converted into multiple embeddings
2. Uses a stack of encoders to create vector representations for each input
3. Vector representation is fed into linear layer
4. Linear layer is fed into softmax, which gives us a probability distribution for the labels



[4]

Emotion Detection framework

- Started with a self-made framework for BERT for training and evaluating the model with pytorch
- Framework mostly worked, but was very basic
- Later, discovered HuggingFace, a state-of-the-start NLP library specifically for transformer models
- Rewrote our framework based on the HuggingFace Trainer

[5]

MLflow

- MLflow tracking server for our evaluation
- Setup was easy with docker
- Uses a S3 storage (minio) for storing the models
- HuggingFace already has MLflow integration built in

[6]

[7]

Training and evaluation

Training our emotion detection model

- Avoid Overfitting:
  - Early Stopping method
  - Label weights in loss function
- F1-Macro of 87.4% for non-emotional/emotional setting

Emotion Detection evaluation results

Results with three-class setting

Dataset	F1-Score [%]	F1-Positive [%]	F1-Negative [%]	F1-Neutral [%]
SemEval_2018	**62,1**	86,2	88,9	11,3
SemEval_2007	**64,6**	68	79,5	46,4
EP	**72,1**	82,1	86	48,4
Annot_sent	**55,4**	82,6	83,5	0
Alm	**61,5**	54,6	54,2	75,7
Combined	**83,6**	**79,5**	**91,2**	**80,2**

Results with two-class setting

Dataset	F1-Score [%]	F1-Emotional [%]	F1-Non Emotional [%]
SemEval_2018	**55,6**	97,3	13,8
SemEval_2007	**64,6**	80,3	48,9
EP	**67,5**	83,1	51,9
Annot_sent	**49,6**	99,3	0
Alm	**67,4**	63,1	71,6
Combined	**87,4**	**94**	**80,8**

Argument Mining evaluation results

Two-class setting:

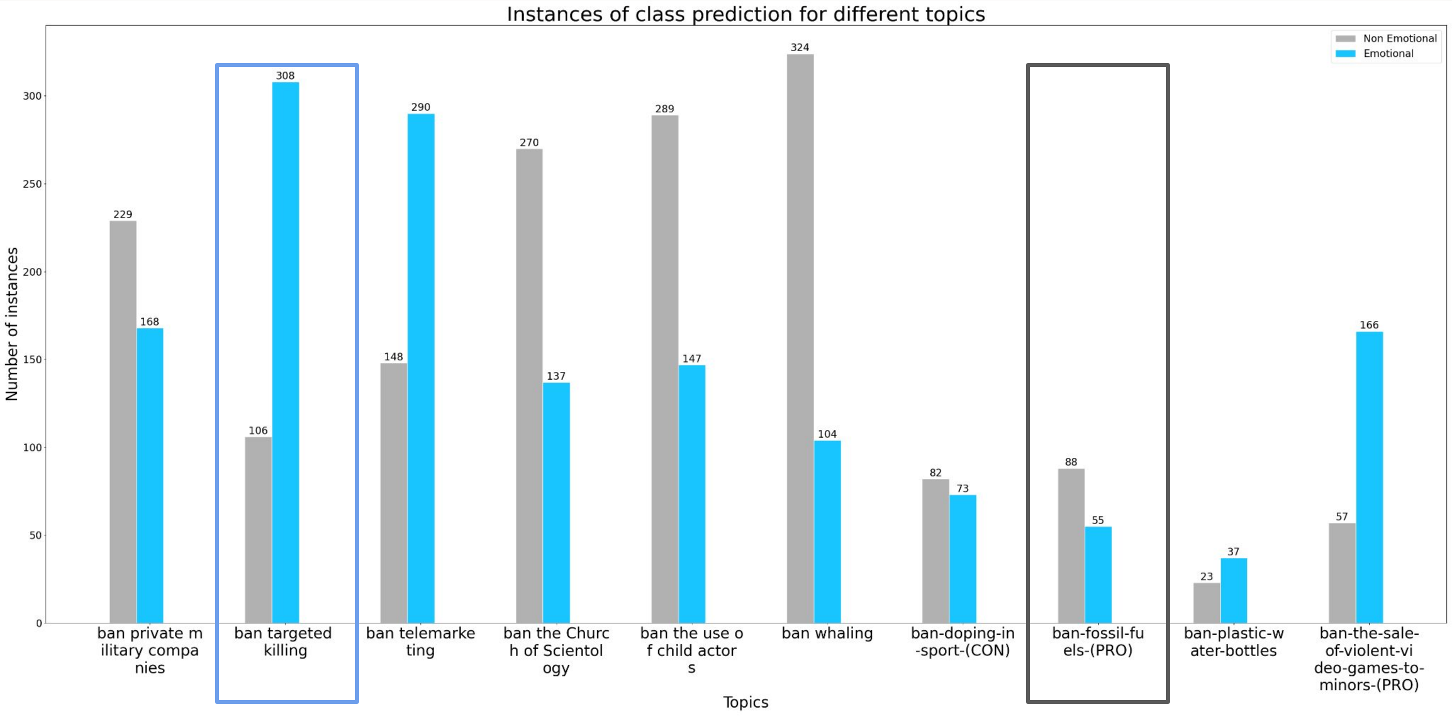
- Annotators agreement ≥ 4 → **f1-score: 70%** (523 sentences)
  (avg. human performance: 80%)
- Annotators agreement ≥ 5 → **f1-score: 76%** (340 sentences)
  (avg. human performance: 90%)
- Annotators agreement = 6 → **f1-score: 80%** (156 sentences)

- Annotators agreement ≥ 4: Baseline evaluation with "always neutral"
  → f1-score: 57%

⇒ **Our emotion model works with arguments!**

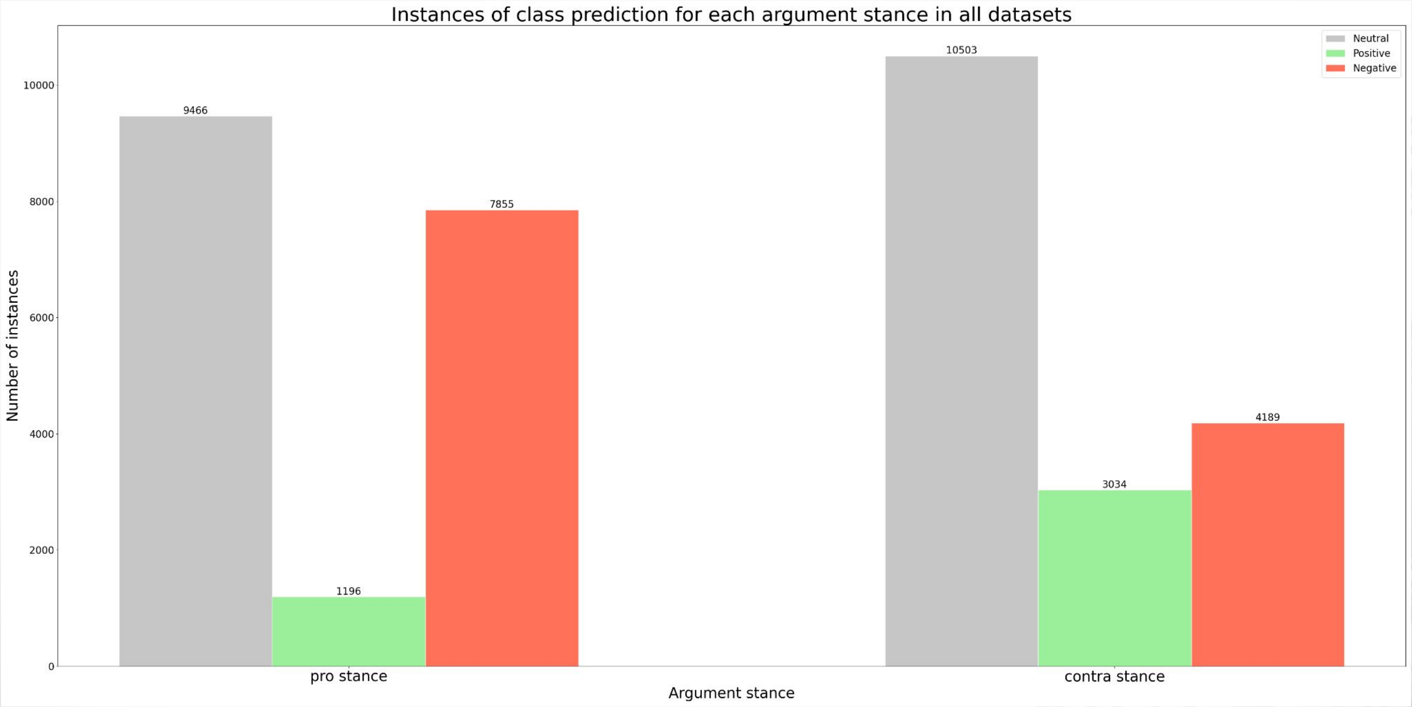
Hypothesis 1 - Some topics are more emotional than others

Hypothesis accepted ✅



Hypothesis 2 - There is a correlation between argument stance and emotion

Hypothesis rejected ❌



Hypothesis 3 - There is a positive correlation between argumentativeness and emotion intensity

Hypothesis rejected ❌

- no clear pattern indicating a positive correlation
  - positive arguments ⇒ likely less convincing
  - non-emotional vs emotional ⇒ no clear correlation

Swanson	negative	positive	neutral
arg_strength_mean	0,53	**0,42**	0,54
pearson	-0,05	**-0,20**	-0,02
spearman	-0,04	**-0,19**	-0,05

Swanson	emotional	non-emotional
arg_strength_mean	0,51	0,54
pearson	-0,03	-0,06
spearman	-0,05	-0,05

Analysis

Evaluation of the annotated data

The F1 score of our model given our annotated arguments with min. 5 agreements

Dataset	F1 score
UKP	0.778
Gleize	0.752
Gretz	0.798
Swanson	0.706

Evaluation of the annotated data

The F1 score of our model given our annotated arguments with min. 5 agreements

Topic	F1 score
All topics	0.765
Death Penalty	0.635
Evolution vs Creation	**0.817**
Gun Control	0.798
William Farquhar ought to be honoured as the rightful founder of Singapore	**0.467**
Gay Marriage	0.707
Evolution	0.487

Error analysis

⇒ The model seemingly tends to predict sentences as emotional, if they contain emotional keywords such as “like”

⇒Example: “Like it or not Social Media is part of living in 2018. Sure it has problems, yes it has advantages and my 2 cents is everything in moderation.”

- Ground truth: Neutral
- Model prediction: Emotional

Error analysis

⇒ The model seemingly tends to predict that sentences containing “I” and “you” as emotional, but “they” and “we” are neutral

⇒Example: “You see, what you fail to realize is that your Amendment...IS DEAD!”

- Ground truth: Emotional
- Model prediction: Emotional

Summary and outlook

Research questions - Conclusion

- ✅ We can build a neural network that can detect emotions
- ✅ Some topics are more emotional than others
- ❌ There exists a correlation between stance and emotion
- ❌ There exists a positive correlation between argumentativeness and emotion intensity

Future work

- Multiple architectures (LSTM) or other transformer approaches
- Fine-tuning on the emotion annotated arguments
- Including of the topic in the training process
- Analysis with more detailed emotion labels

Sources

[1] Alswaidan, Nourah, and Mohamed El Bachir Menai. "A survey of state-of-the-art approaches for emotion recognition in text." Knowledge & Information Systems 62.8 (2020).

[2] Trautmann, Dietrich, et al. "Fine-grained argument unit recognition and classification." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 05. 2020.

[3] http://jalammar.github.io/illustrated-bert/

[4] Deep Learning for Natural Language Processing (WS20/21) - BERT & Co.

[5] https://huggingface.co

[6] https://www.mlflow.org

[7] https://www.docker.com

[8] https://inception-project.github.io

[9] https://www.storyboardthat.com/de/articles/e/ethos-pathos-logos

Thank you for your attention