Personalized NLP

Przemysław Kazienko, Jan Kocoń
Department of Artificial Intelligence
Wroclaw University of Science and Technology, Poland

AGENDA

1. Example and motivation
2. Subjective NLP tasks
3. Measuring diversity
4. Perspectives
5. Research on offensive content
6. Research on emotional dataset
7. Research on multiple tasks
8. Conclusions

1

MOTIVATION

*"Your behaviour is inappropriate and your reaction is exaggerated.
I am not sure if you should have administrator rights."*

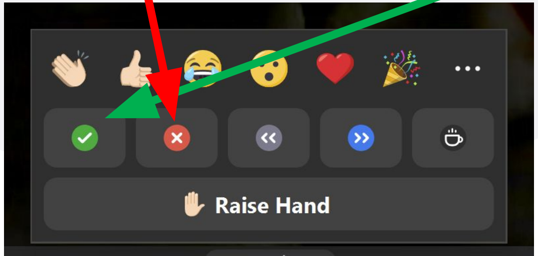
Wikipedia Detox Aggression

Do you think, it is aggressive or not?

"Your behaviour is inappropriate and your reaction is exaggerated. I am not sure if you should have administrator rights."

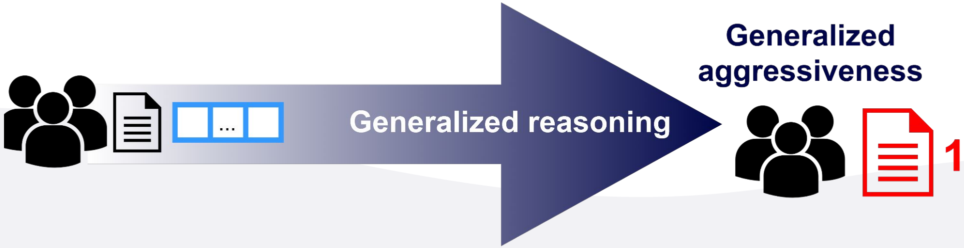
Wikipedia Detox Aggression

Do you think, it is aggressive or not?



MOTIVATION

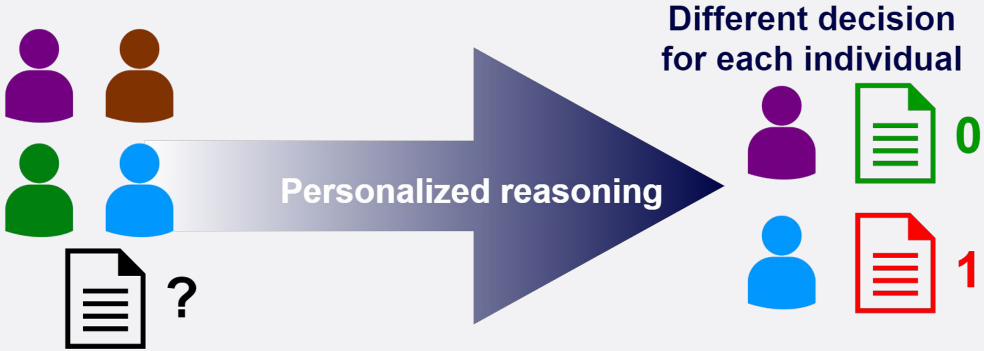
COMMON GENERALIZED NLP



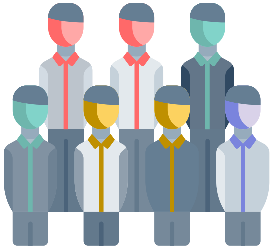
MOTIVATION

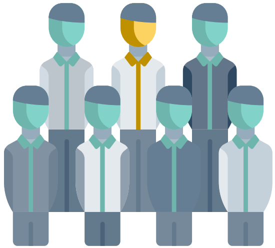
COMMON GENERALIZED NLP

OUR PERSONALIZED NLP



MOTIVATION





Representativeness

Hard to **acquire** data (annotations) from **all** social groups representing all diverse beliefs

“The people like me are not respected by the system”

Fairness

Common generalized solutions are **biased** toward the mainstream

“Since the system does not regard my individual beliefs, I do not trust in it”

2
SUBJECTIVE NLP TASKS

SUBJECTIVE NLP TASKS

1. **Reader** perspective: **perception** prediction
    a. Emotions (many models, multiple dimensions)
    b. Offensive content detection, incl. aggression, toxic, hate speech, cyberbullying, hostile, insulting
    c. Humor, funny
    d. Sarcasm and irony detection
    e. Antagonistic, provocative, trolling speech detection
    f. Counterspeech detection
    g. Hope, supportive speech detection
    h. Obscene language detection
    i. Dismissive, patronising, condescending
    j. Unfair generalisation
    k. Slur usage
    l. Unpalatable questions
    m. Persuasiveness
    n. Inflammatory text
    o. Subjective perception of sentiment polarization

2. **Author** perspective
    a. Sentiment analysis
    b. Content generation (e.g. style-based), summarization, adjustment

3. **Mixed**
    a. Conversations

The tasks often overlap

3

MEASURING DIVERSITY

[Kan21, Mił21, Koc21b]

MEASURING DIVERSITY

Document-oriented

Document **Controversy** (entropy-based) [Kan21]

Human-oriented

Human **Conformity**; general, weighted, class-based [Kan21]

HB-measure – Human Bias [Koc21b]; aggregated Z-score; for emotions: PEB – **Personal Emotional Bias** [Mił21]

Collection-oriented

Krippendorff's alfa [Koc21a]

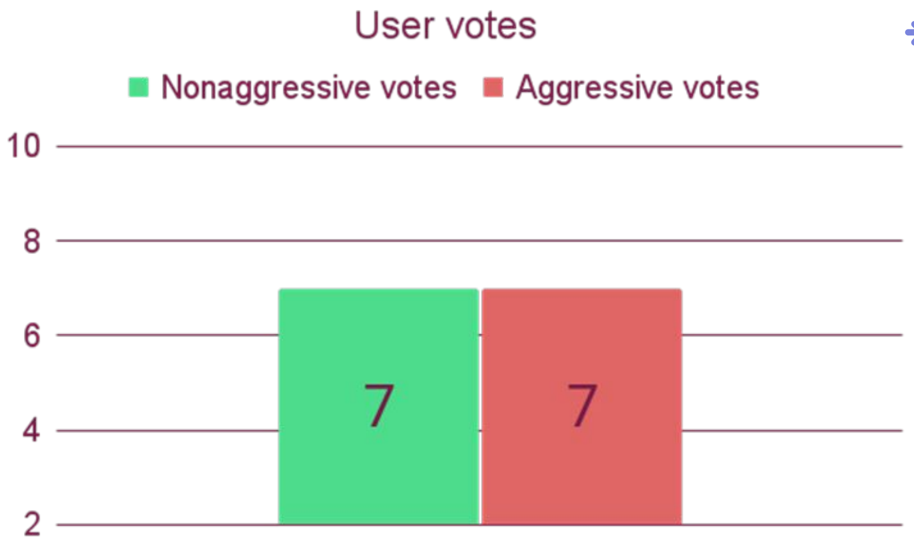
WAVE kappa – Wroclaw Annotators Variability Estimator; Fleiss' kappa aggregated over different no. of users [Koc21a]

CONTROVERSY MEASURE

"Your behaviour is inappropriate and your reaction is exaggerated. I am not sure if you should have administrator rights."

CONTROVERSY = 1.0 (entropy-based)

$$Contr(d) = \begin{cases} 0, \text{ if } n_d^0 = n_d \vee n_d^1 = n_d \\ -\sum_{c=0,1} \frac{n_d^c}{n_d} \log_2 \left(\frac{n_d^c}{n_d} \right) \end{cases}$$



Real data: Wikipedia Detox

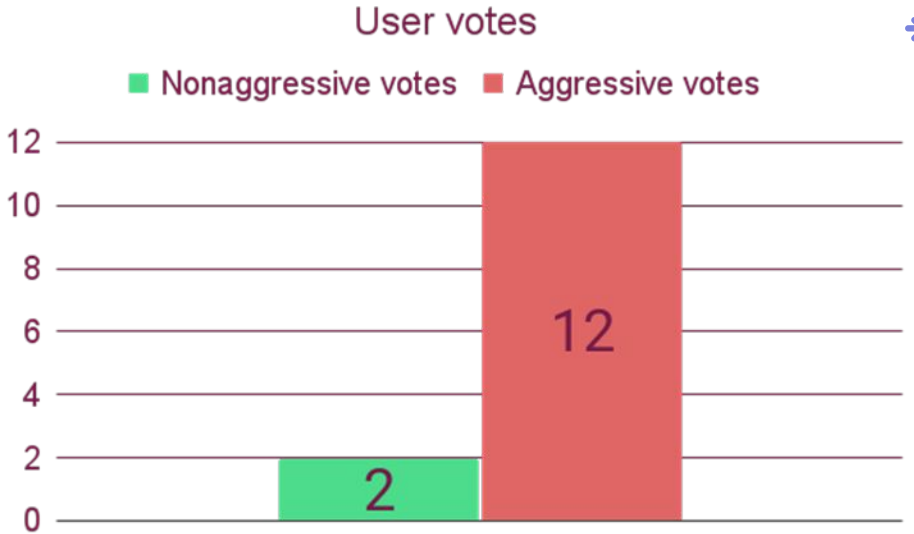
CONTROVERSY MEASURE

~~inappropriate~~

***"Your behaviour is terrible and your reaction is exaggerated.
I am not sure if you should have administrator rights."***

CONTROVERSY = 0.59 ↓
(entropy-based)

$$Contr(d) = \begin{cases} 0, \text{ if } n_d^0 = n_d \vee n_d^1 = n_d \\ -\sum_{c=0,1} \frac{n_d^c}{n_d} \log_2 \left(\frac{n_d^c}{n_d} \right) \end{cases}$$



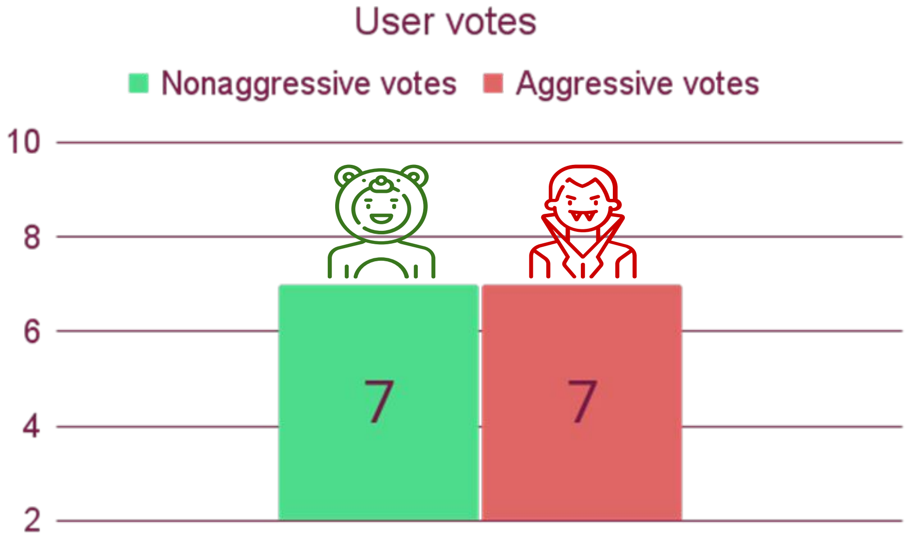
CONFORMITY MEASURE

"Your behaviour is inappropriate and your reaction is exaggerated. I am not sure if you should have administrator rights."

CONFORMITY = 0.50

CONFORMITY = 0.50

$$GConf(a, C) = \frac{\sum_{d \in A_a} \mathbb{1}_{\{l_d \in C \wedge l_d = l_{d,a}\}}}{\sum_{d \in A_a} \mathbb{1}_{\{l_d \in C\}}}$$



Real data: Wikipedia Detox

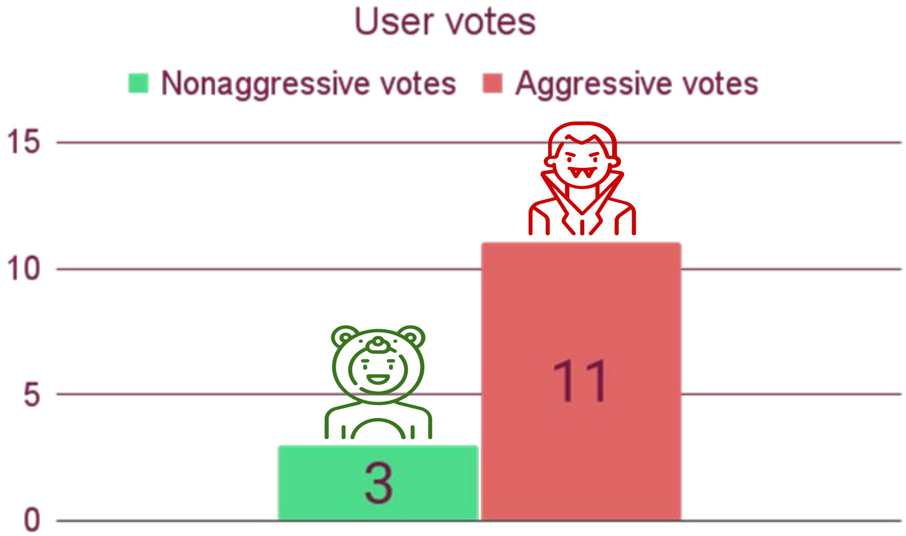
CONFORMITY MEASURE

"Your behaviour is terrible and your reaction is exaggerated. You don't deserve administrator rights."

CONFORMITY = 0.21 = $\frac{3}{14}$

CONFORMITY = 0.79

$$GConf(a,C) = \frac{\sum_{d \in A_a} \mathbb{1}_{\{l_d \in C \wedge l_d = l_{d,a}\}}}{\sum_{d \in A_a} \mathbb{1}_{\{l_d \in C\}}}$$

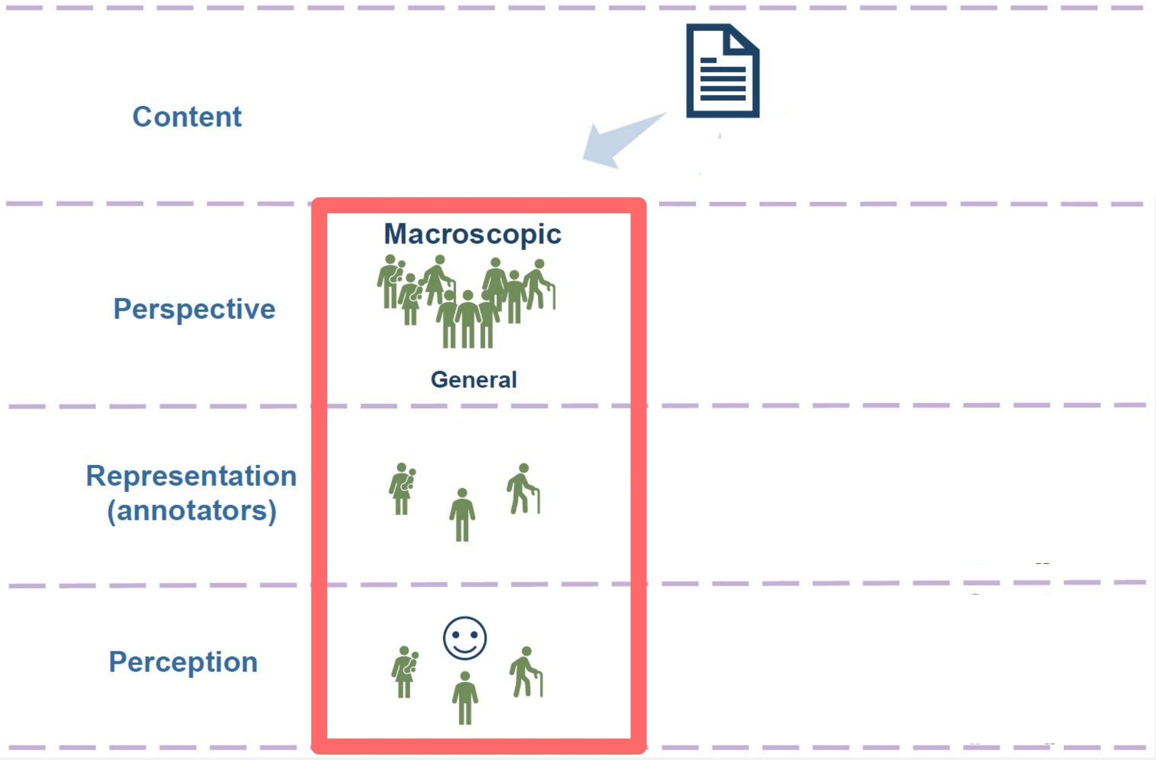


4

PERSPECTIVES

[Koc21a]

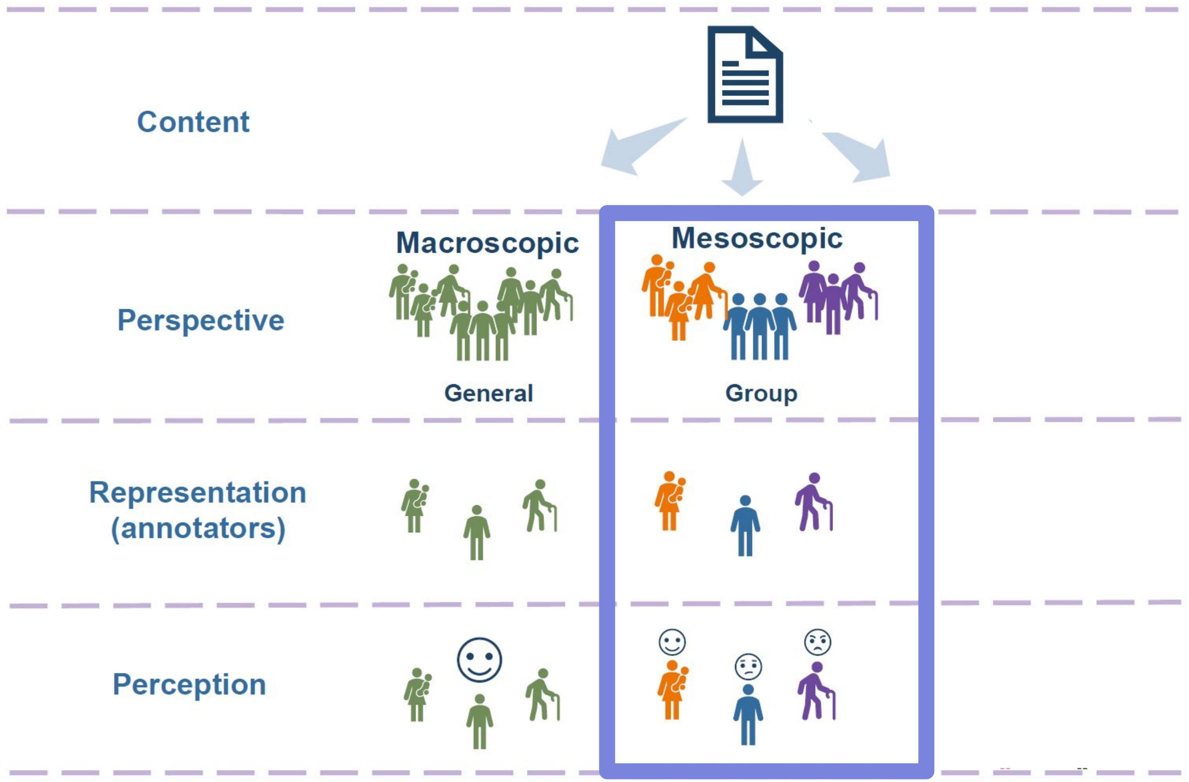
PERSPECTIVES: MACROSCOPIC



PERSPECTIVES: MACROSCOPIC (general)

Perspective profile	Statement	Information source	Annotation
Society-based, global, general. Used in most research. Assumes the existence of **common perception** of the content	*"People generally treat some content offensive/funny/sad/..."*	(1) content (2) context of the content, e.g. source	Several **trained/expert annotators** are able to express **common perception** (beliefs)

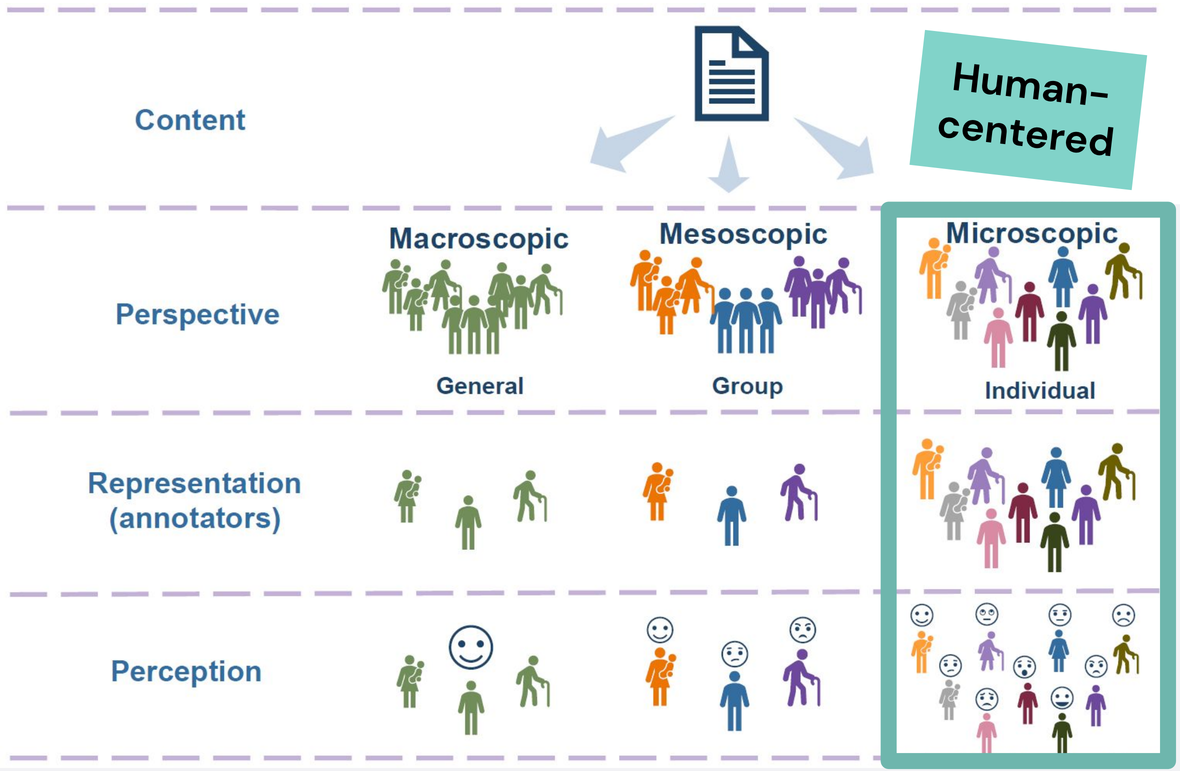
PERSPECTIVES: MESOSCOPIC



PERSPECTIVES: MESOSCOPIC (group-based)

Perspective profile	Statement	Information source	Annotation
Group-based, social or demographic groups. Perception is **shared** in **social groups**	*"There are some groups of people who perceive the content in the same way as offensive/funny/sad/…"*	(1) content (2) context of the content (3) **group demographic profile**, e.g. age (4) **group context**, e.g. culture, shared personality traits, religion	A lot of annotations per document are required. **Annotator profiles** need to be collected (surveys, behaviour)

PERSPECTIVES: MICROSCOPIC



PERSPECTIVES: MICROSCOPIC (personalized)

- Human-centered

Perspective profile	Statement	Information source	Annotation
Individual, fully personalized. Each **individual** may perceive content **differently**.	*"Perception of the content depends on a single human, i.e. on their individual and temporal concext"*	(1) content (2) context of the content (3) individual **behaviour** (4) individual **demographics** (5) individual **social context** (relationships with the author and the social group) (6) temporal **affective state** (mood, emotions)	An **individual** annotator **beliefs** need to be identified using surveys and/or previous annotations

PERSONALIZED NLP: What we need?

Data about human beliefs

Texts earlier annotated by a given individual

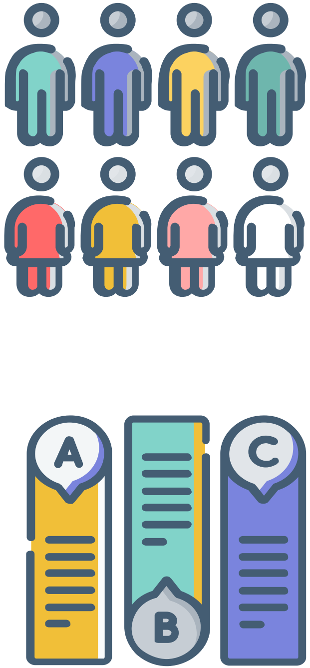
Agreed, generalized labels are useless

Usually obtained by majority voting

5

RESEARCH ON OFFENSIVE CONTENT

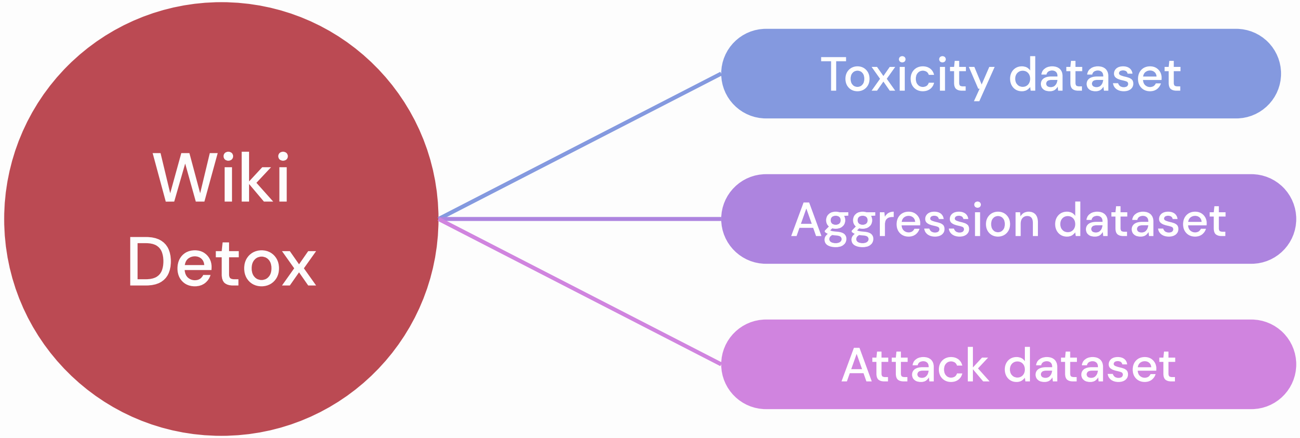
[Koc21a, Kan21, Koc21b]



5a

OFFENSIVE CONTENT: ANNOTATED DATA

WIKI DETOX DATASETS (English)



Publicly available

WIKI: Toxicity

Classes	Texts	People
2	**159,686**	**4,301**

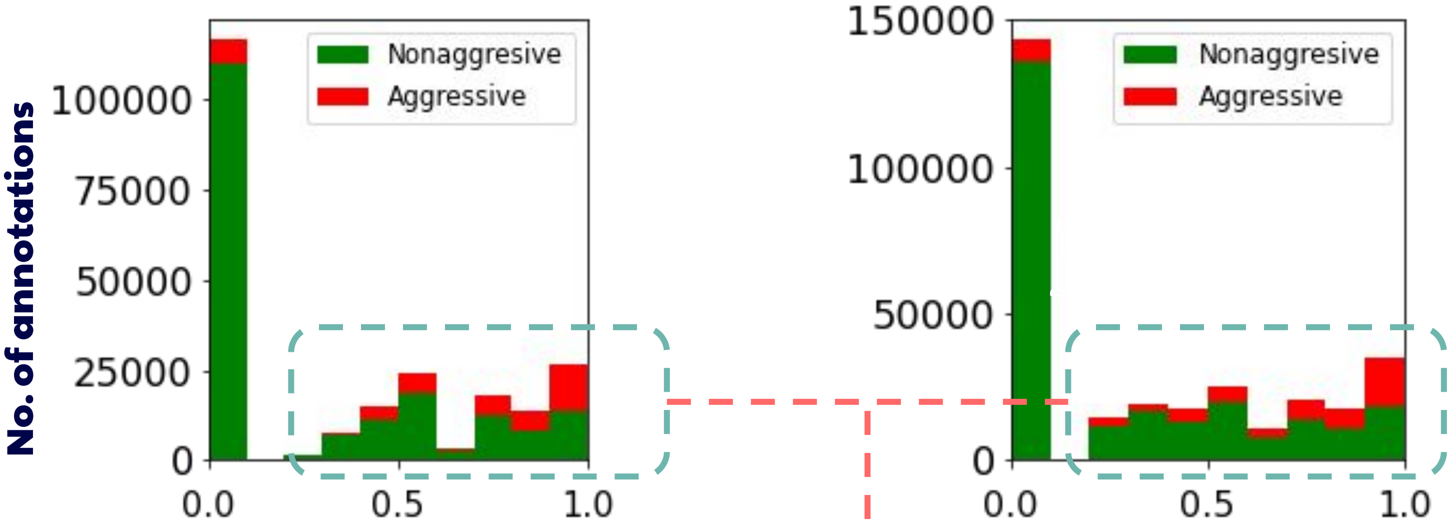
Annotations	Controversial Texts
1,598,289	**40.5 %**

WIKI: Aggression & Attack

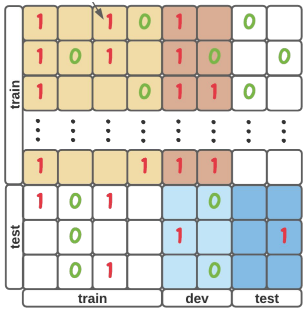
Classes	Texts	People
2	115,864	4,053 2,190

Annotations	Controversial Texts
1,365,217 855,514	51.3% & 48%

WIKI: Aggressive



Disagreement in ~50% of annotations

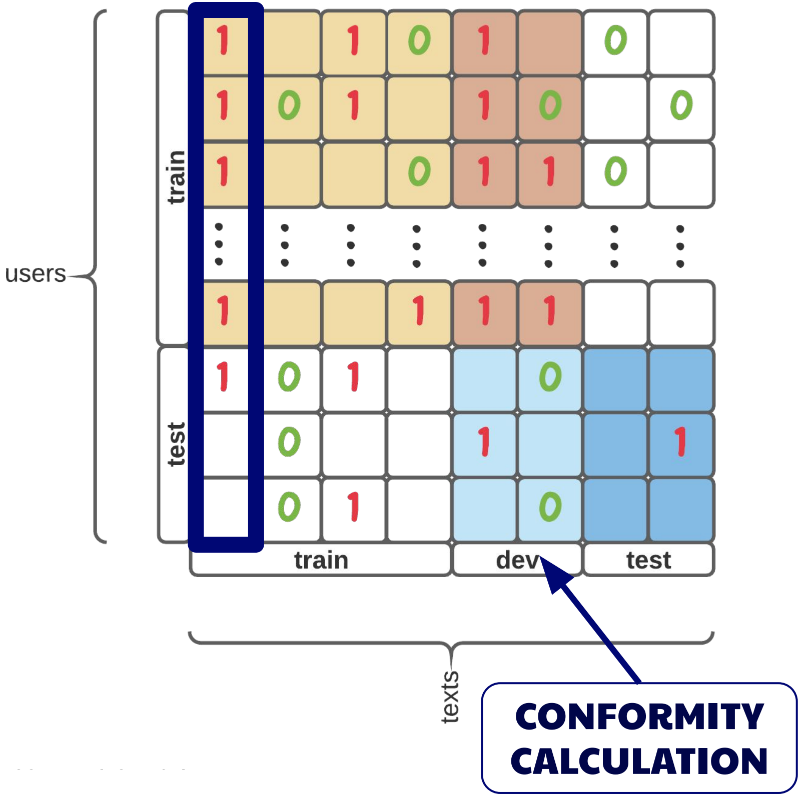


5b

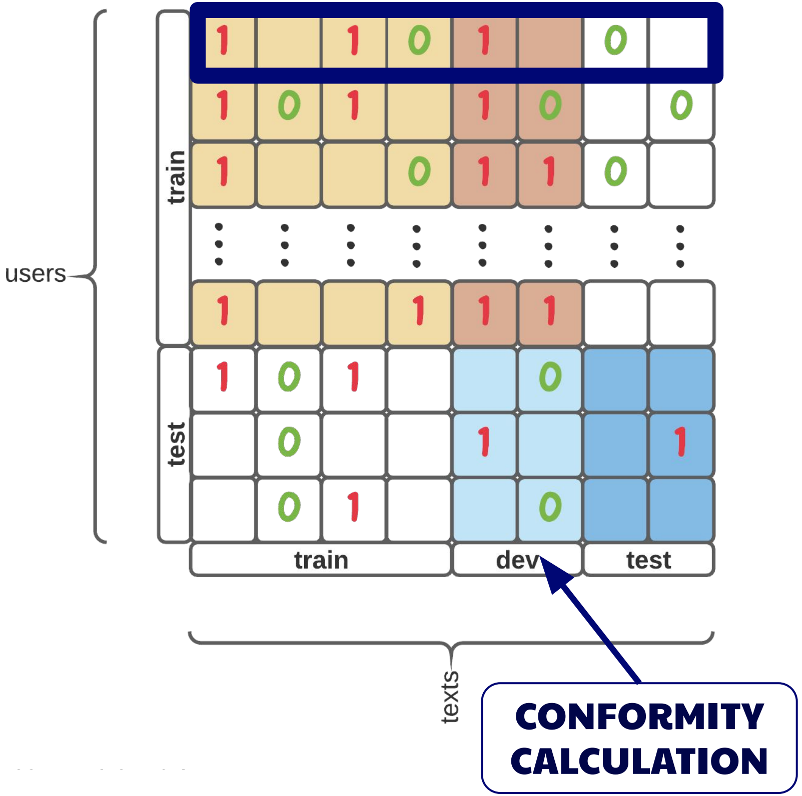
OFFENSIVE CONTENT: DATA SPLIT

Train-dev-test

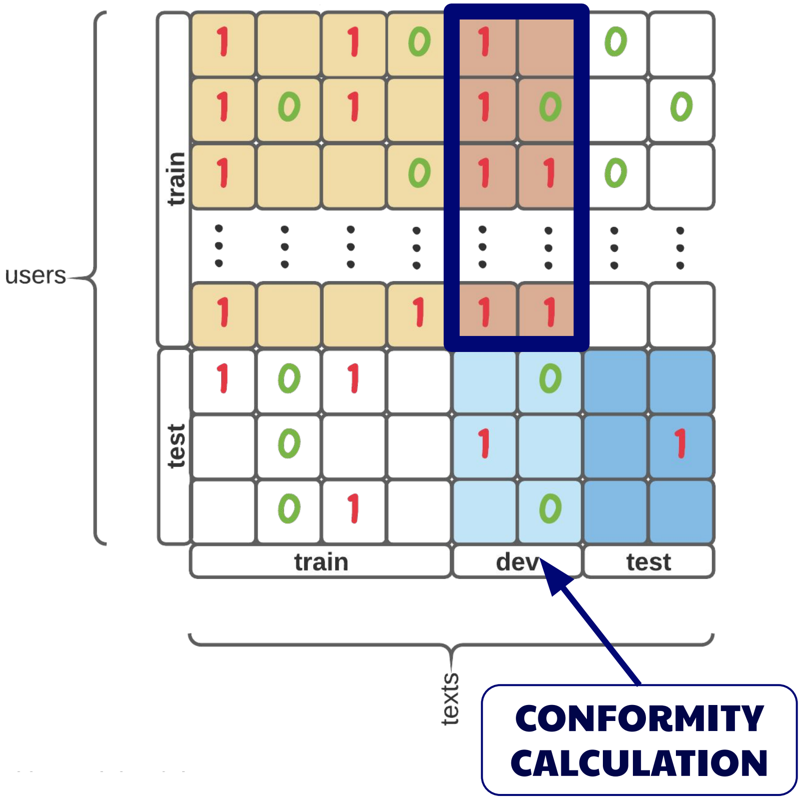
DATASET SPLIT: Wiki



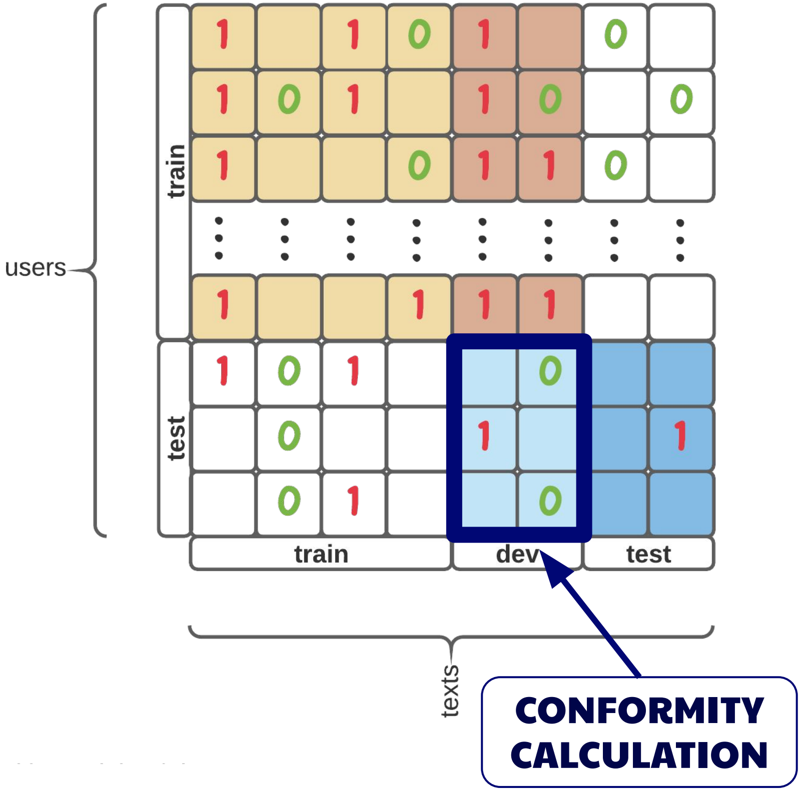
DATASET SPLIT: Wiki



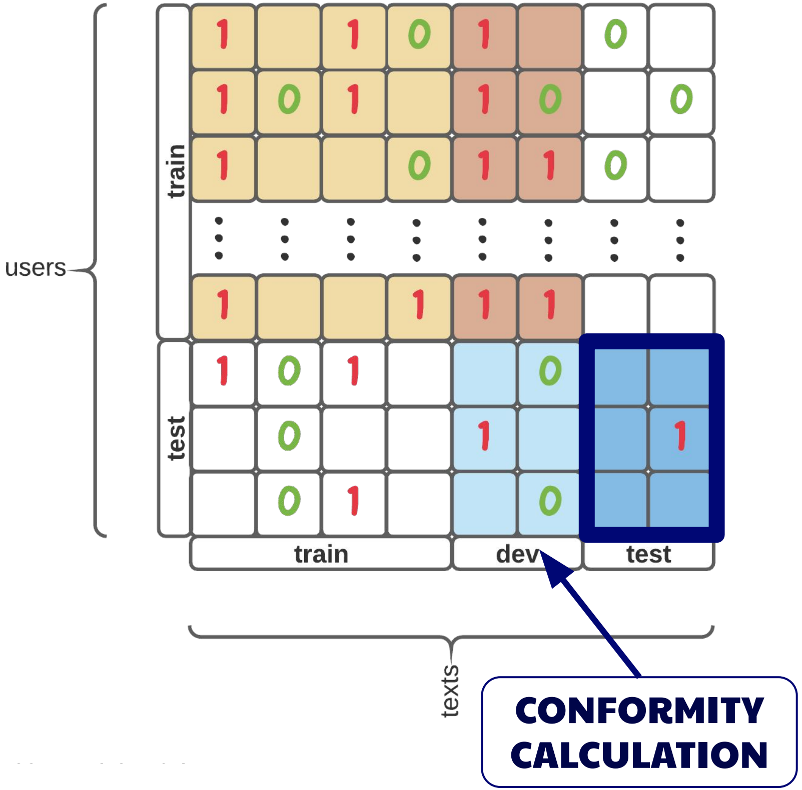
DATASET SPLIT: Wiki



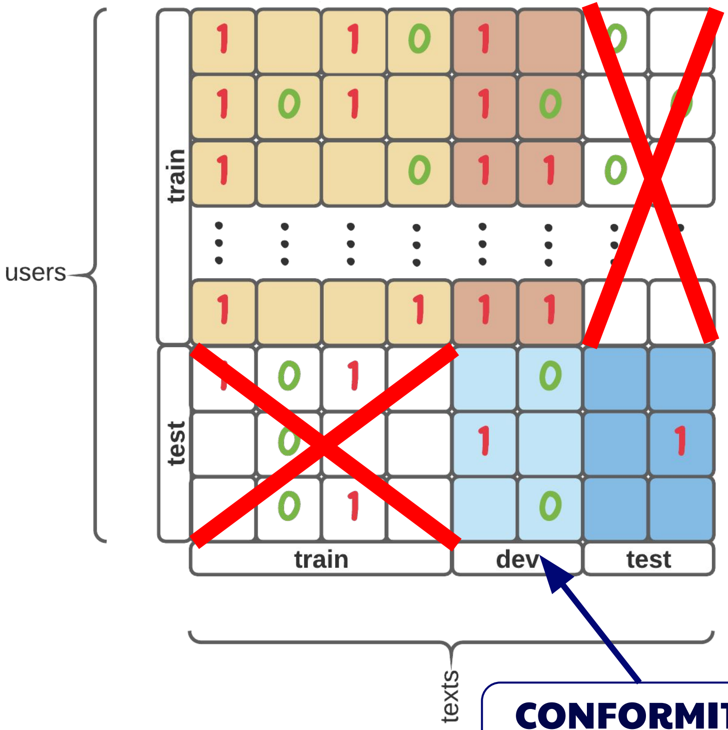
DATASET SPLIT: Wiki



DATASET SPLIT: Wiki



DATASET SPLIT: Wiki

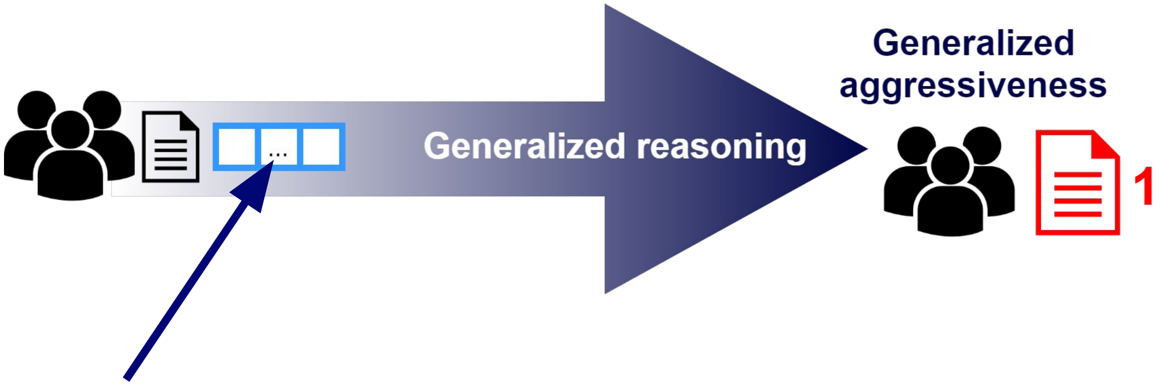


CONFORMITY CALCULATION

5c

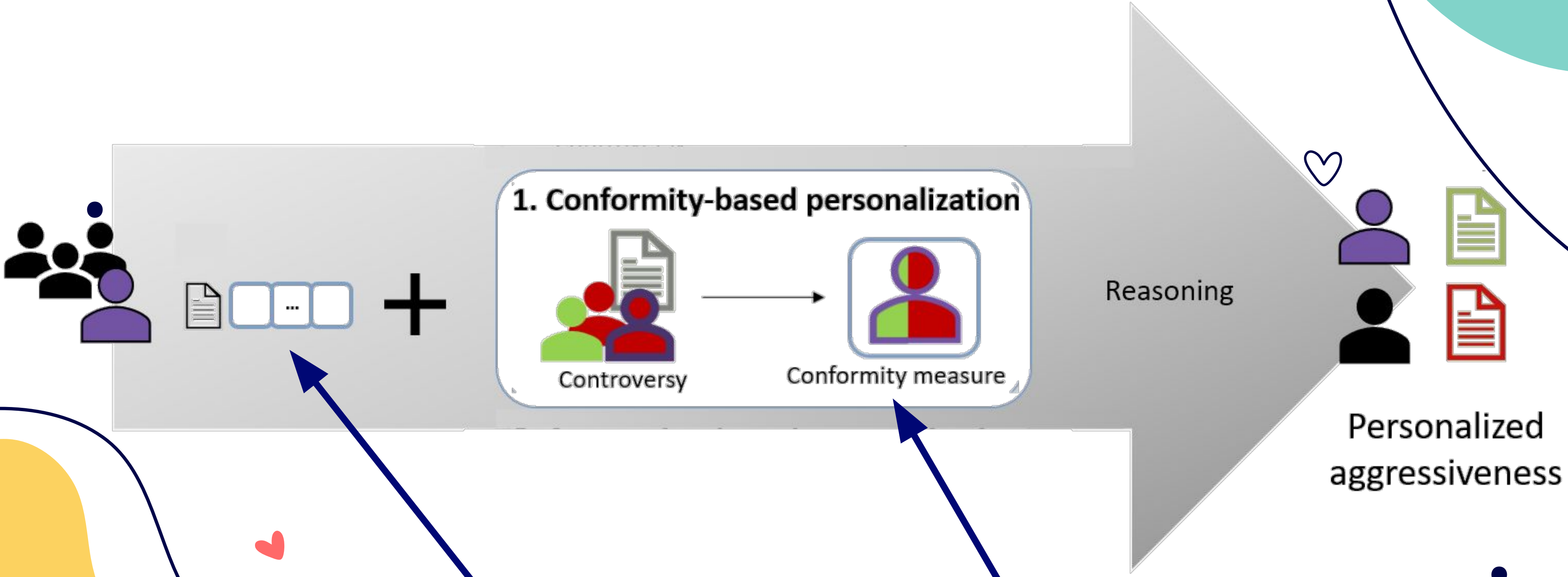
OFFENSIVE CONTENT: METHODS

GENERAL METHOD - BASELINE



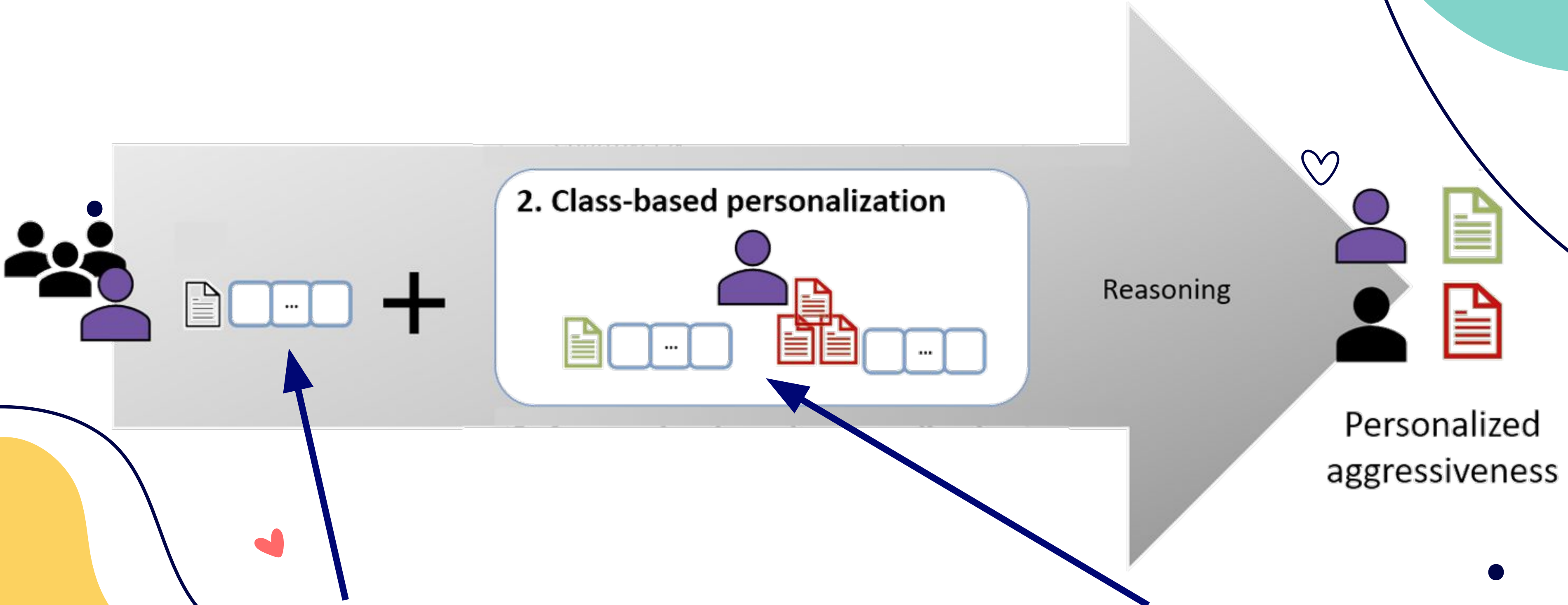
Input: text embedding only

1. CONFORMITY-BASED PERSONALIZATION



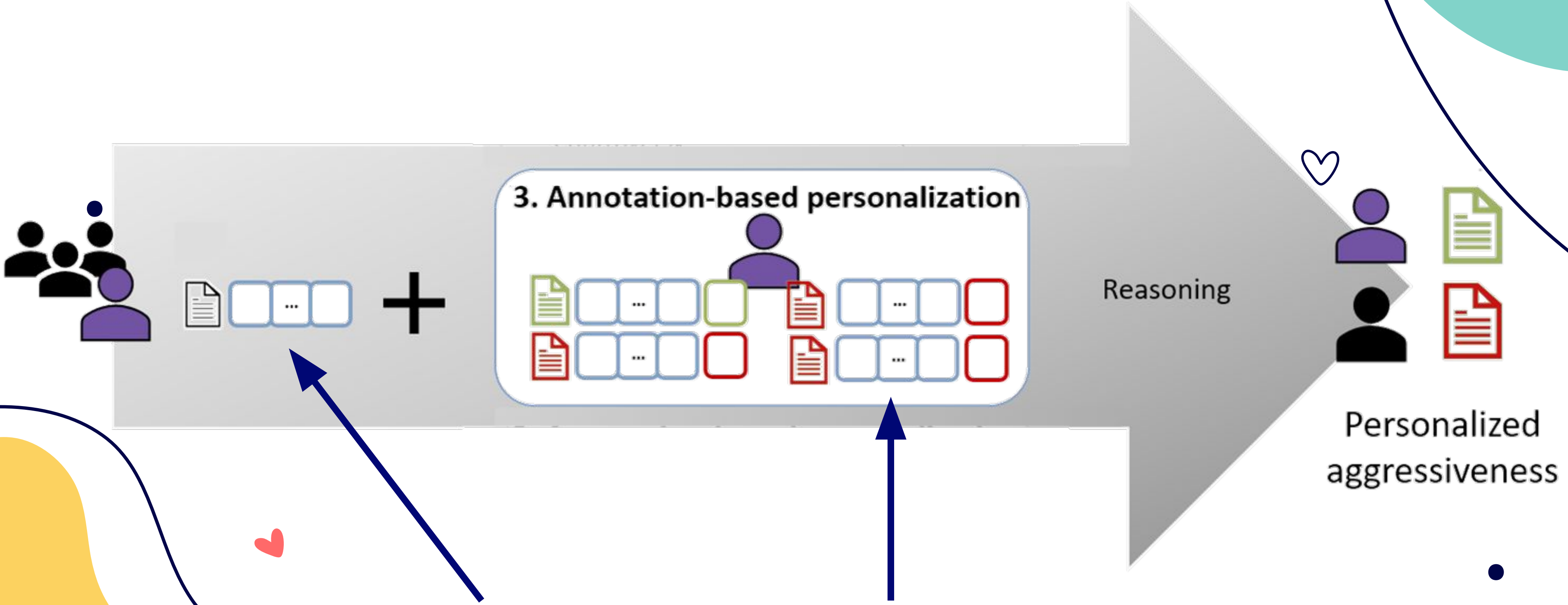
Input: text embedding + user conformity measures (6 features)

2. CLASS-BASED PERSONALIZATION

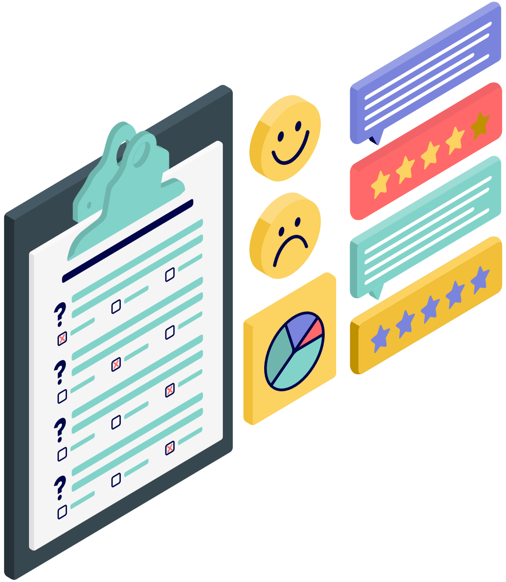


Input: text embedding + texts seen by user as aggressive / non-aggressive (avg. of their embeddings)

3. ANNOTATION-BASED PERSONALIZATION



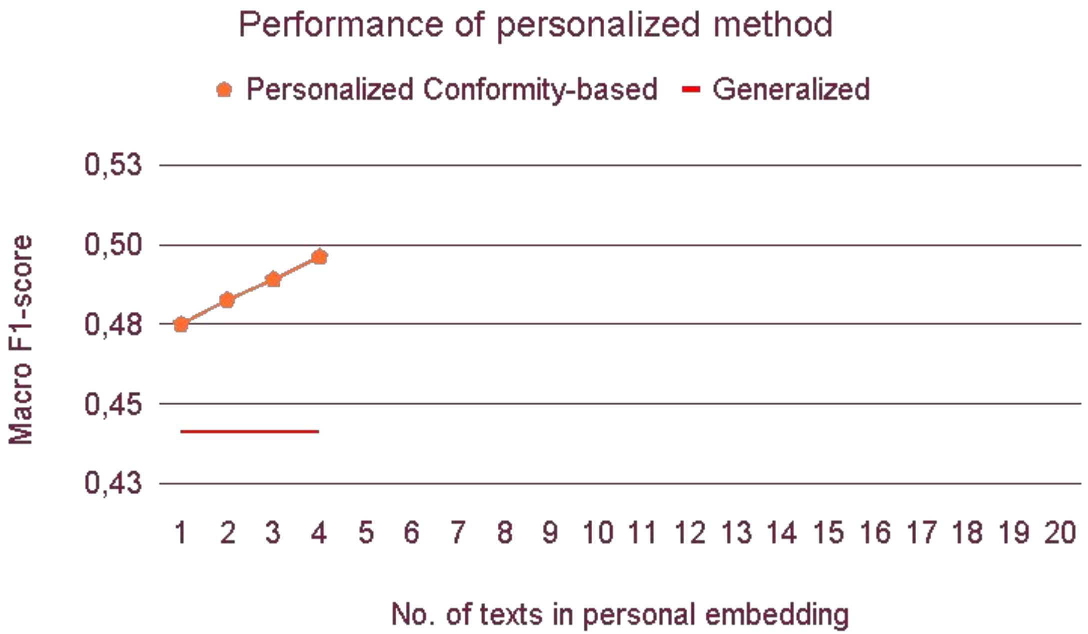
Input: text embedding + all texts prev. seen by the user with their annotations 1 – 0, raw embeddings



5d

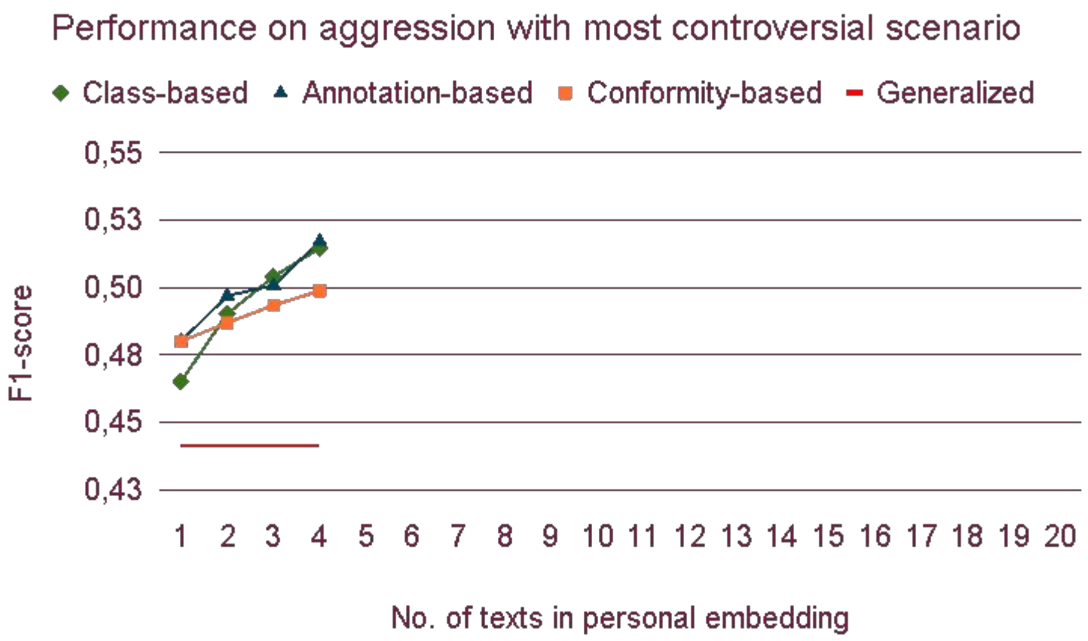
OFFENSIVE CONTENT: RESULTS

EVALUATION RESULTS

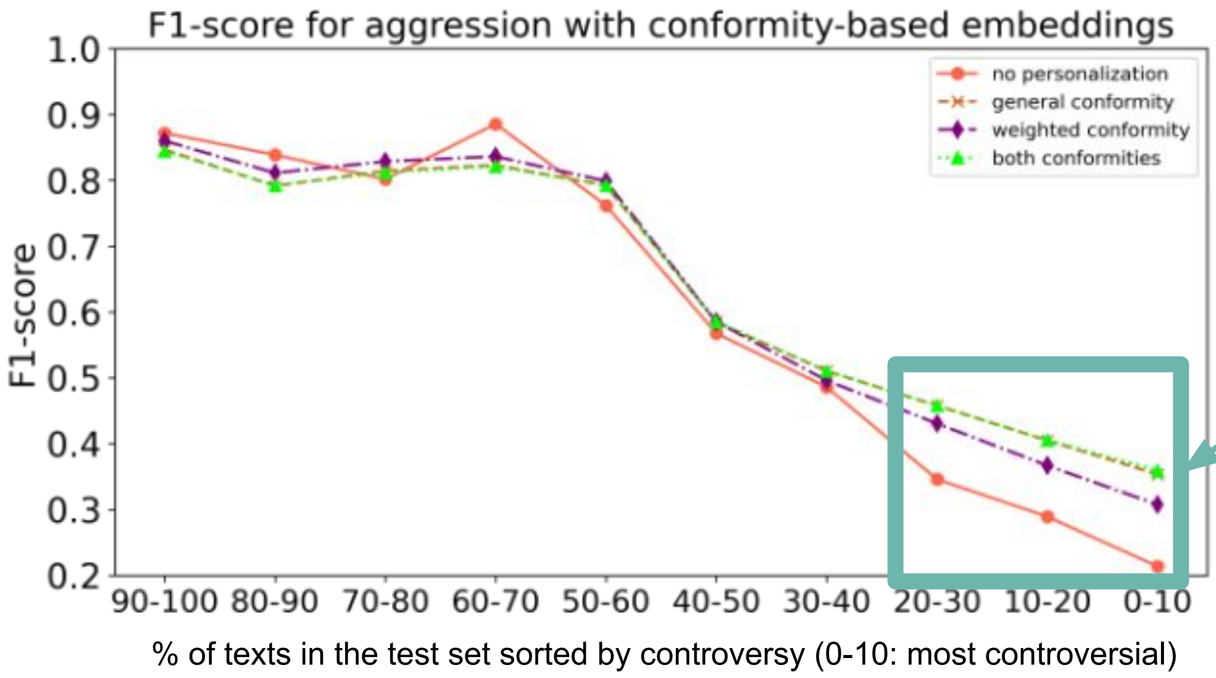


F1 for the ***aggression*** class only

EVALUATION RESULTS

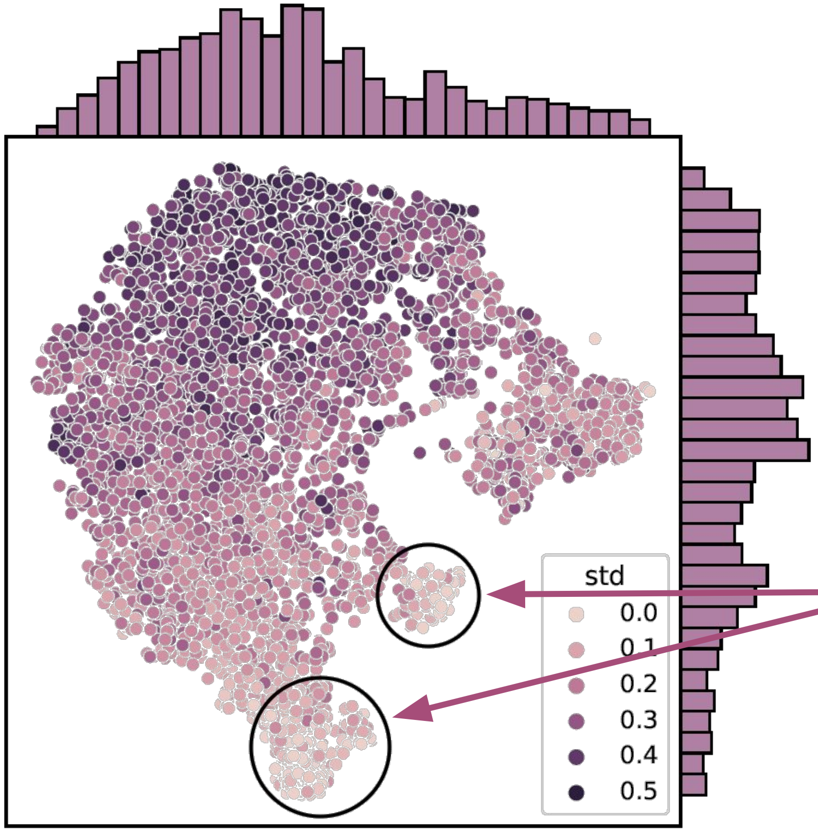


Where PNLP gains?



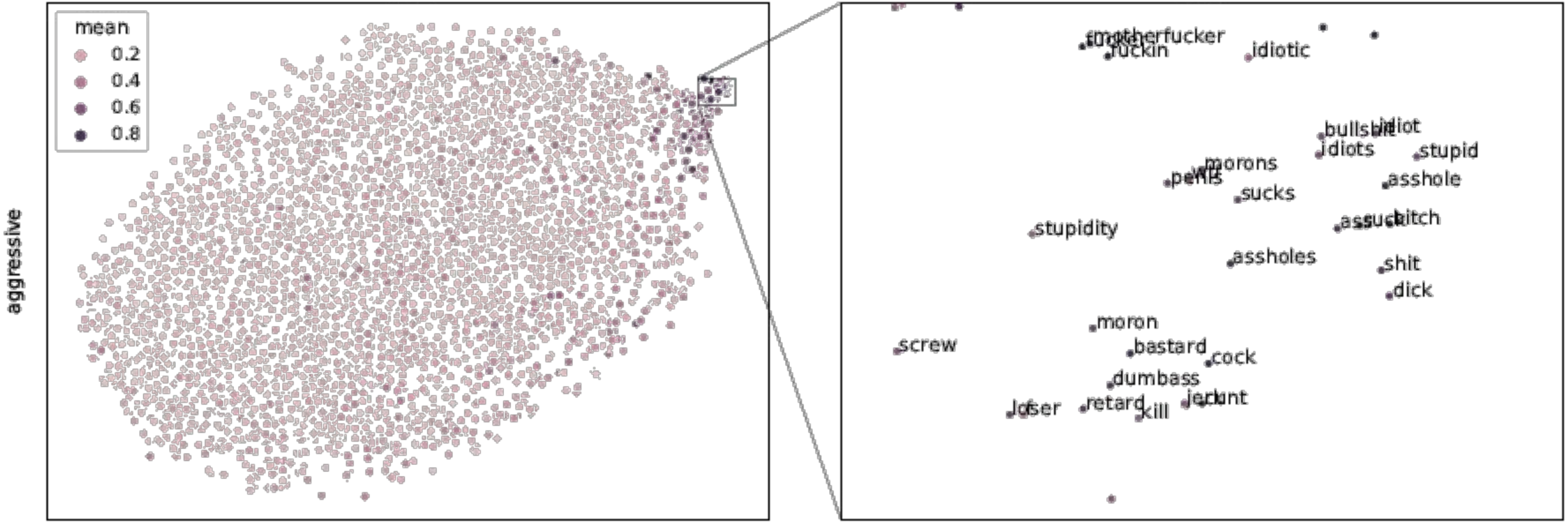
Crucial gain for most controversial texts

HUMAN EMBEDDINGS: Wiki Aggression

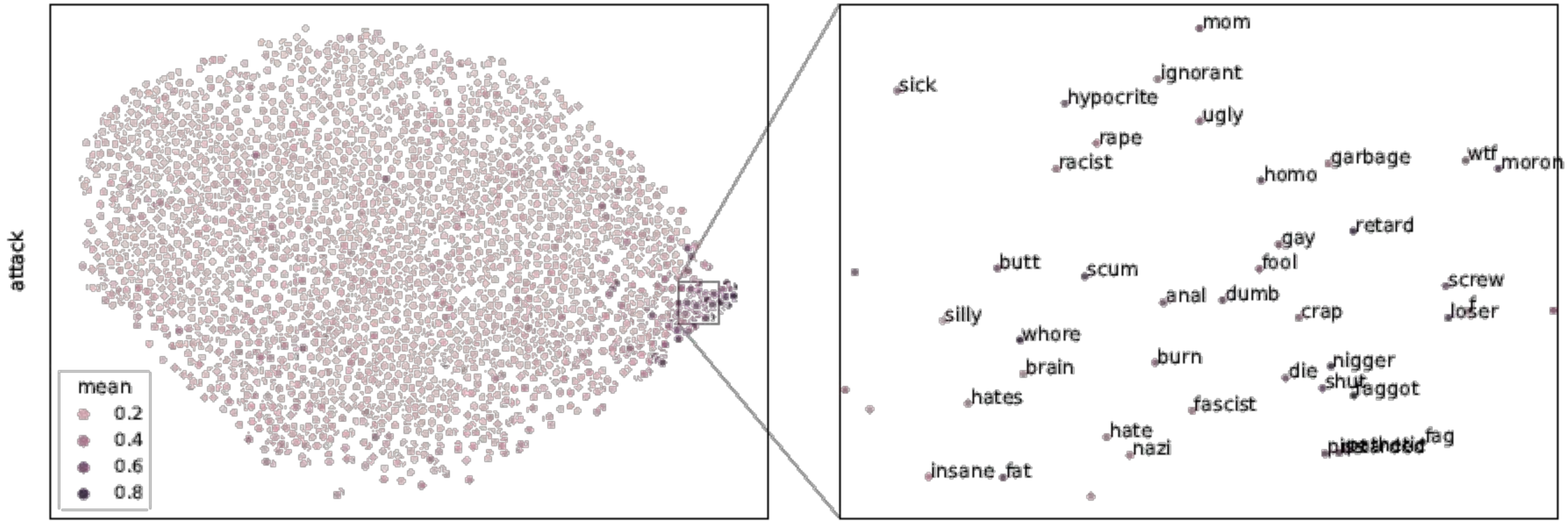


Low std. dev.
for some annotators
⇒ not credible ones?

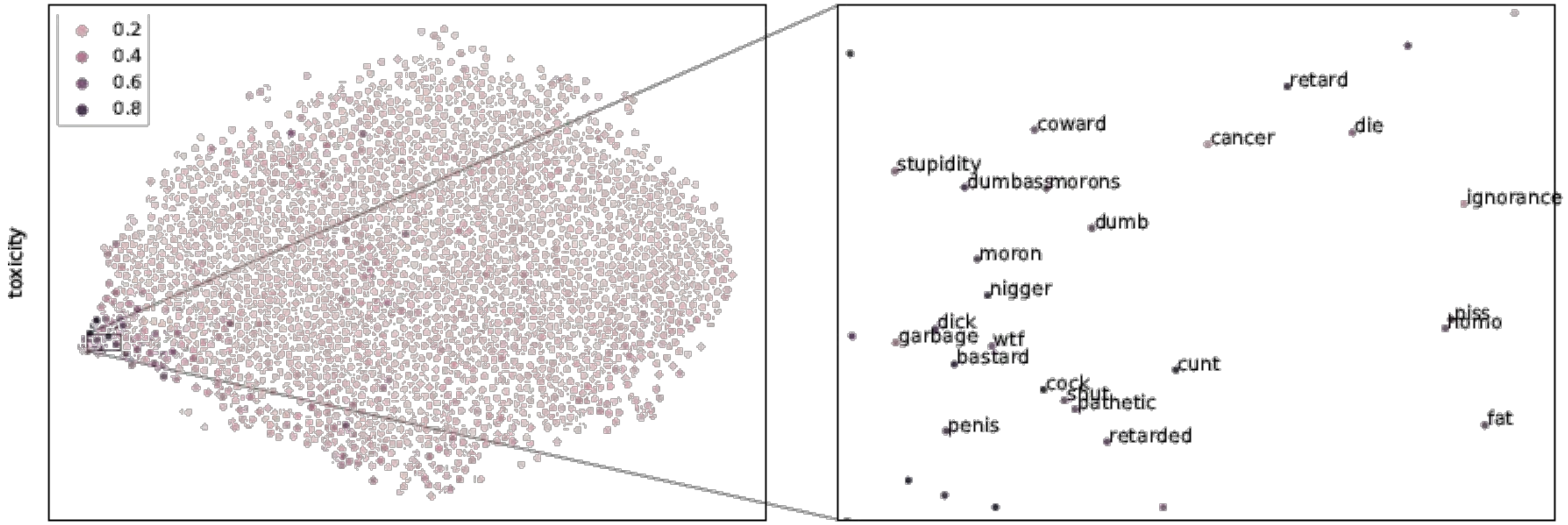
WORD EMBEDDINGS: Wiki Aggression

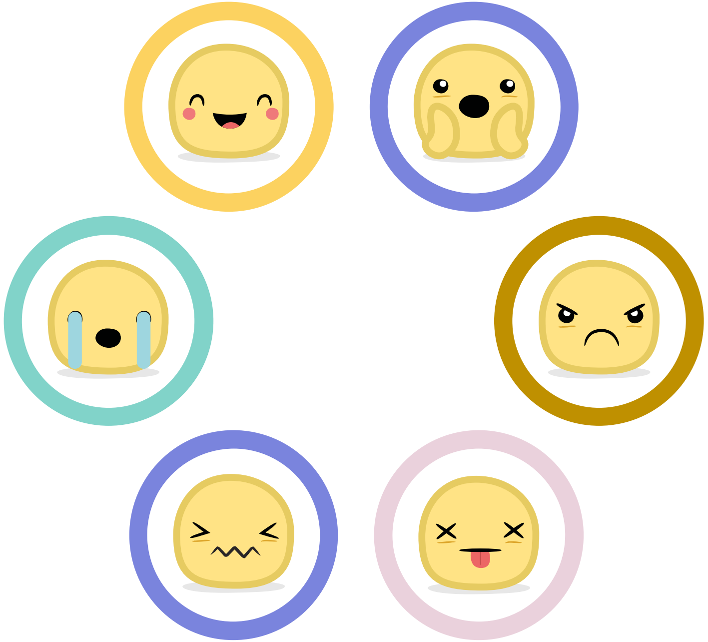


WORD EMBEDDINGS: Wiki Attack



WORD EMBEDDINGS: Wiki Toxicity

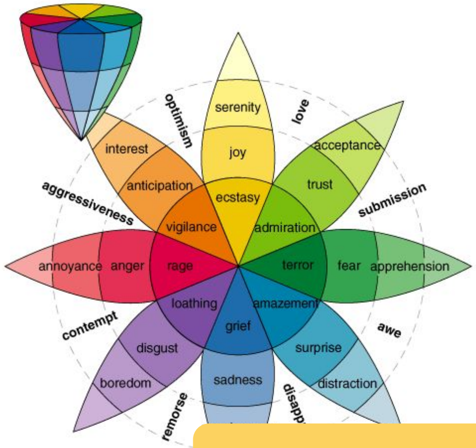




6

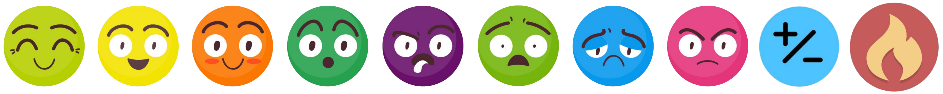
RESEARCH ON EMOTIONAL CONTENT PERCEPTION

ACL2021 – [Mił21]
ICDM2021 – [Koc21b]



EMOTIONAL DATA (in Polish)

Sentimenti



Emotions	Texts	People
10 values	**7,004**	**8,853**

Annotations	Controversial Texts
3,774,338	**100 %**

NOT publicly available

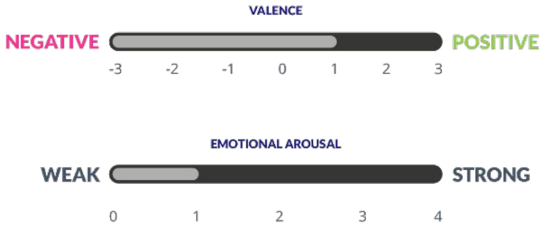
EMOTIONAL TEXTS: example

Sentimenti

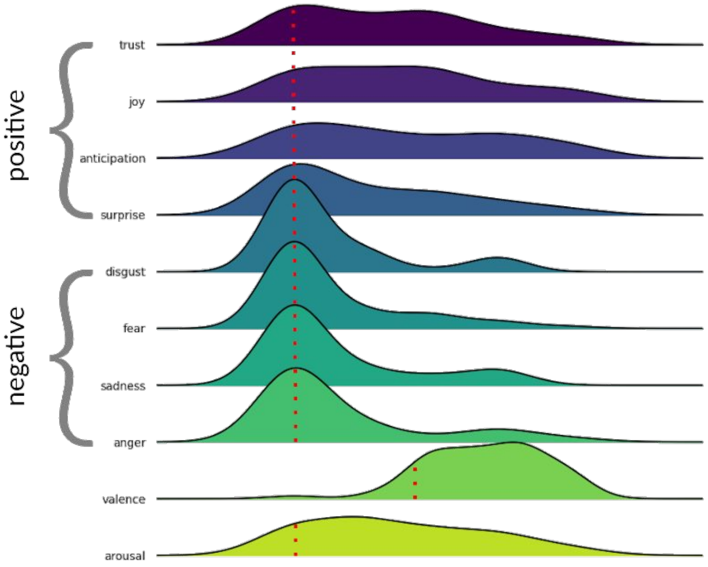
Example opinion

A modern, clean, well-maintained closed housing estate. Tastefully furnished apartments with full equipment. Great swimming pools, playground for children, exercise room - two treadmills and some other equipment, sauna. In fact, the car park is constantly full, we parked in front of the estate's gate. I do not recommend parking in prohibited places, because the security first sticker on the glass sticker, which is said to be hard to take off and then call the police. 10 minutes walk to the sea. Nearby a few places with home-made lunches, a little further on a grocery store. To the promenade on foot about half an hour.

Example anotation

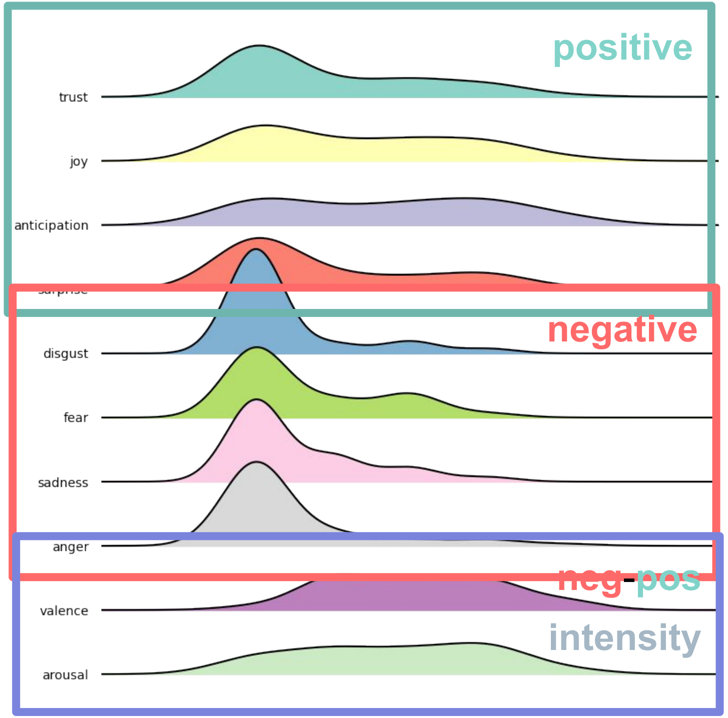


All anotations



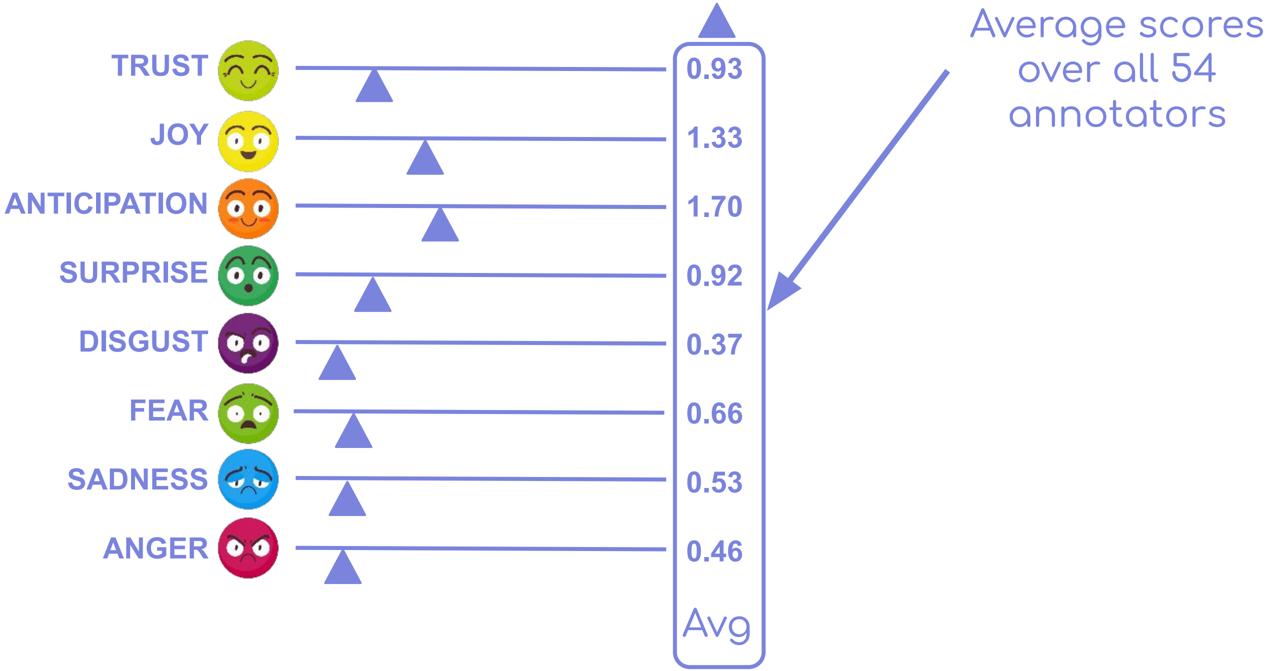
Example opinion

"She closed an unsuccessful chapter in her life and decided to start all over again."



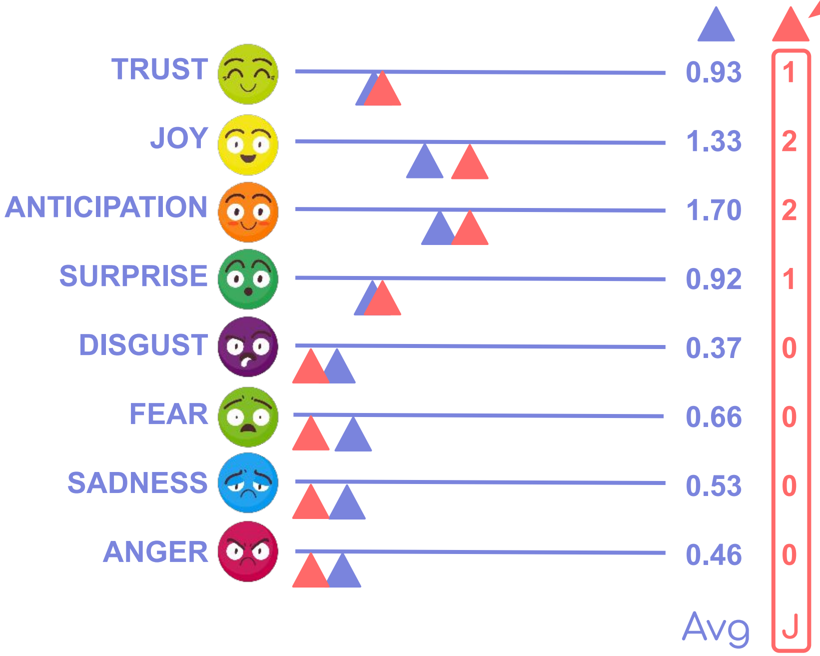
Different answers

"She closed an unsuccessful chapter in her life and decided to start all over again."



Different answers

“She closed an unsuccessful chapter in her life and decided to start all over again.”



John

scores fitting majority

low Personal Emotional Bias (PEB)

PEB: Z-score

$$PEB(u,c) = \frac{\sum_{d \in D_u^{past}} \frac{v_{c,d,u} - \mu_{c,d}}{\sigma_{c,d}}}{|D_u^{past}|}$$

Different answers

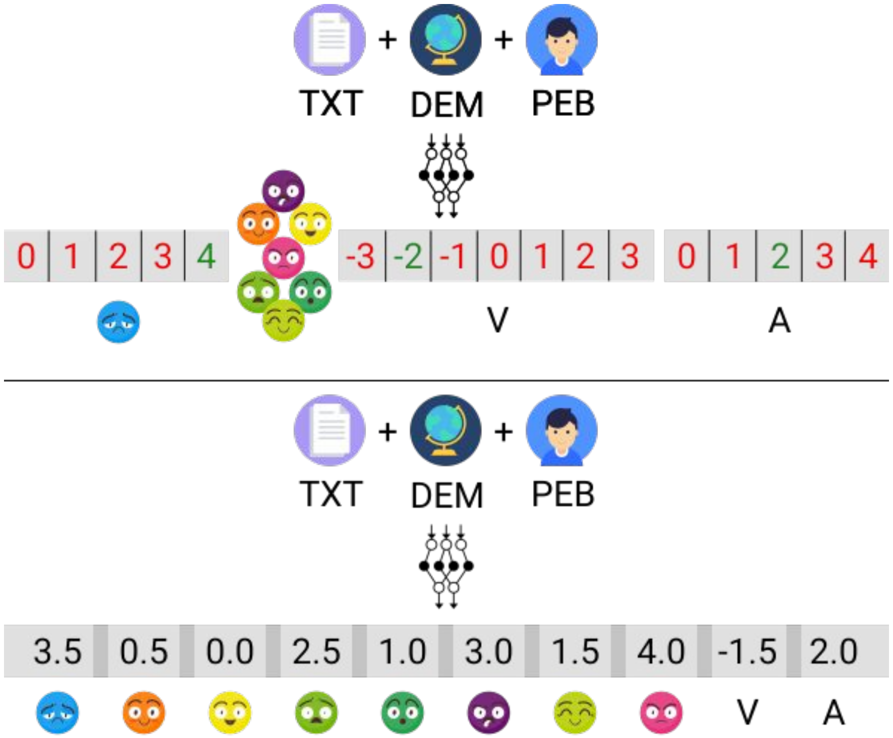
"She closed an unsuccessful chapter in her life and decided to start all over again."

EMOTIONAL EXPERIMENTS

Sentimenti

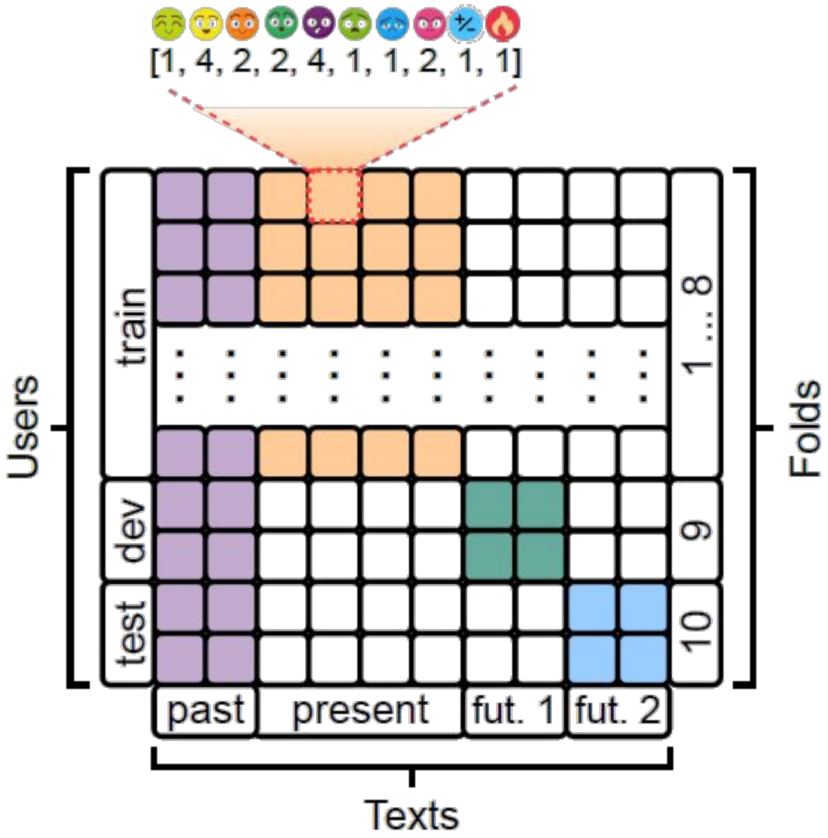
(1) Multi-task classification

(2) Multivariate regression



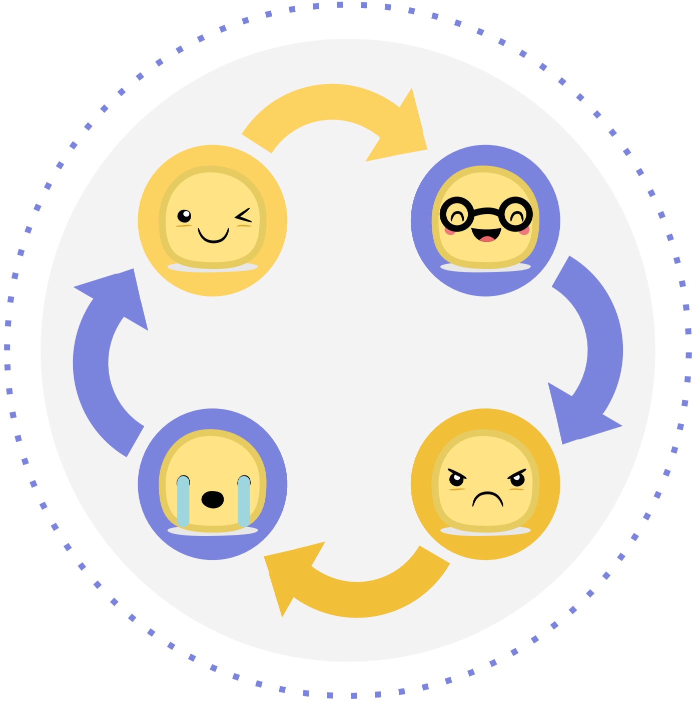
EMOTIONAL DATA SPLIT

Similar to offensive data but **with 10 folds**



PEB: Z-score

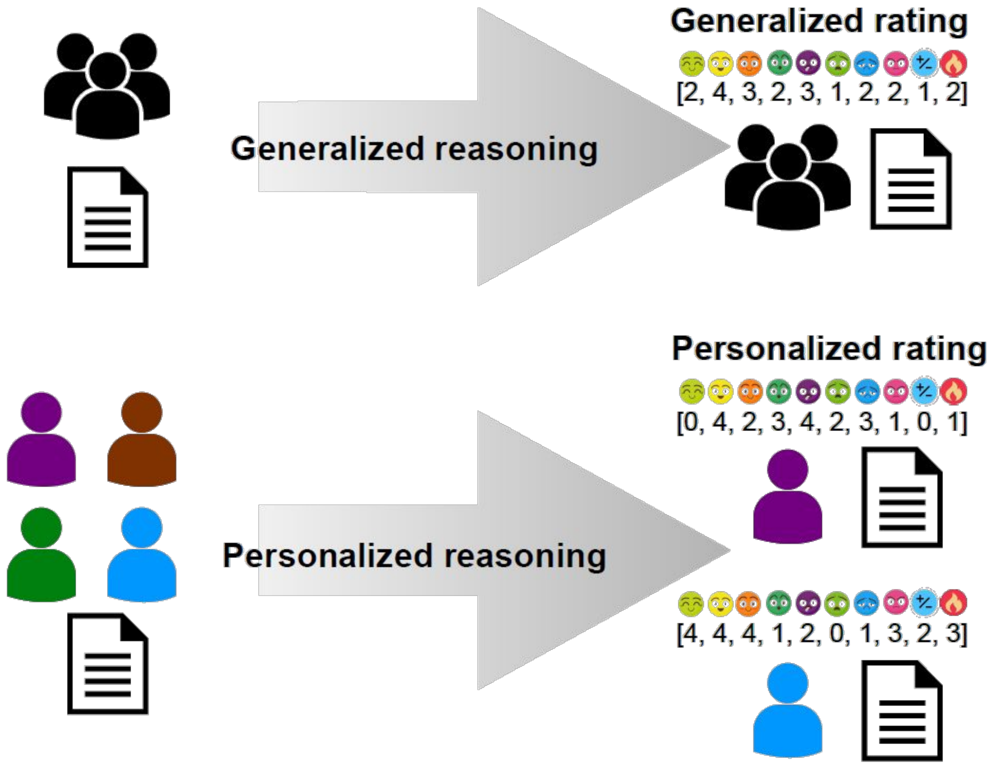
$$PEB(u,c) = \frac{\sum_{d \in D_u^{past}} \frac{v_{c,d,u} - \mu_{c,d}}{\sigma_{c,d}}}{|D_u^{past}|}$$



6a

RESEARCH ON EMOTIONS: METHODS

GENERALIZED vs. PERSONALIZED NLP



FOUR METHODS

1

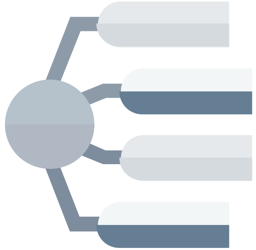
Text

State-of-the art text embeddings (baseline)

2

Demographics

Text + demographic features describing an individual



3

Personal Emotional Bias

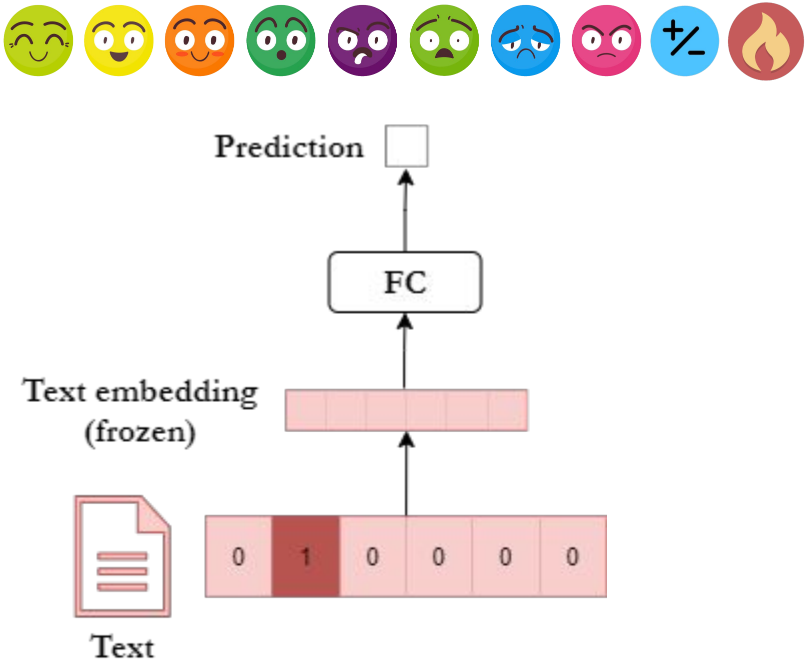
Text + one pre-computed personal feature (human bias)

4

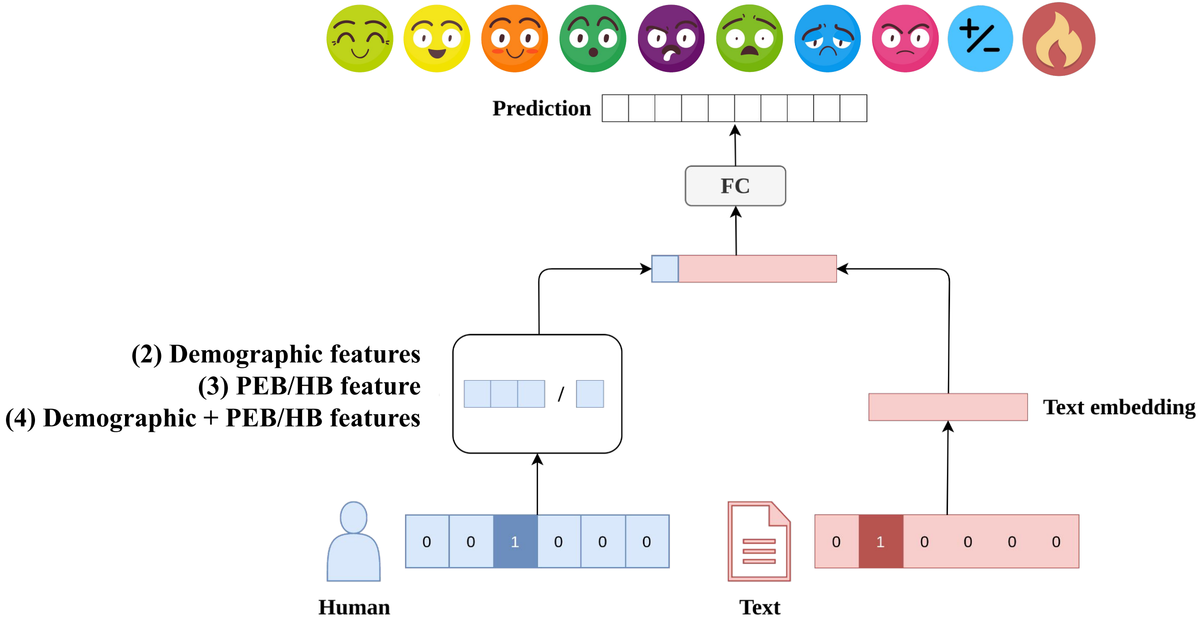
Combined

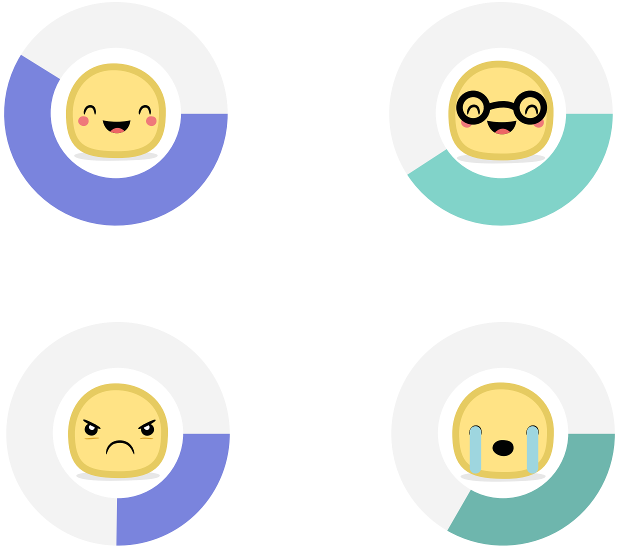
Checking how it performs together

(1) TEXT ONLY: BASELINE



(2) DEMOGRAPHICS &
(3) PERSONAL EMOTIONAL BIAS (PEB/HB)
(4) ALL: demogr. + PEB feature





6b

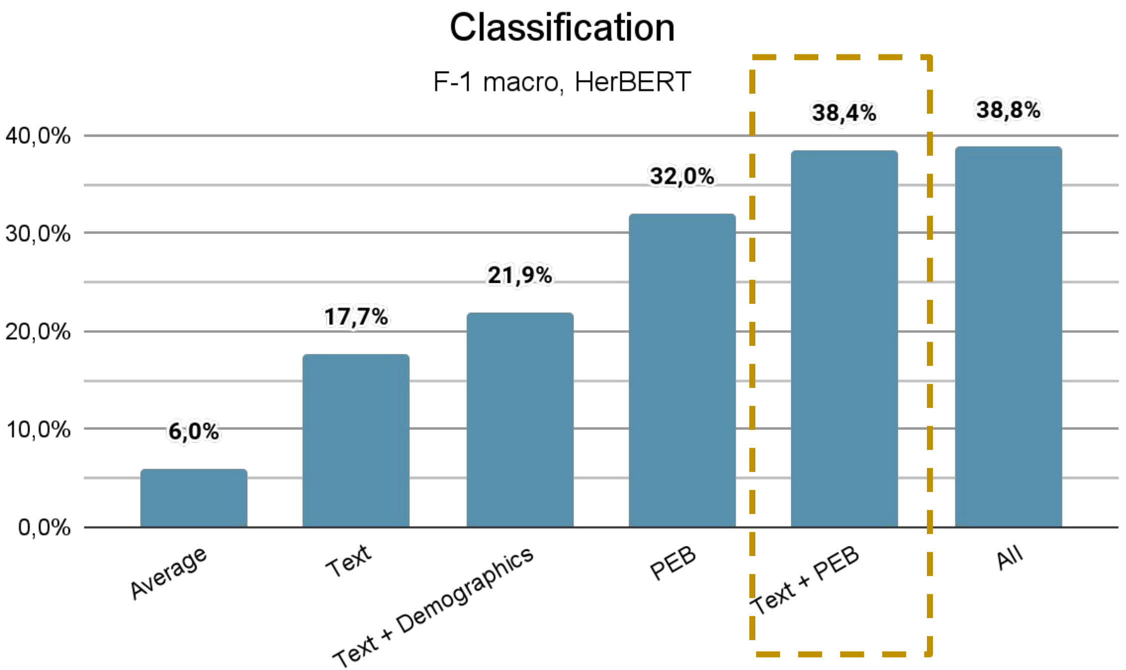
RESEARCH ON EMOTIONS: RESULTS

CLASSIFICATION: all emotions aggregated

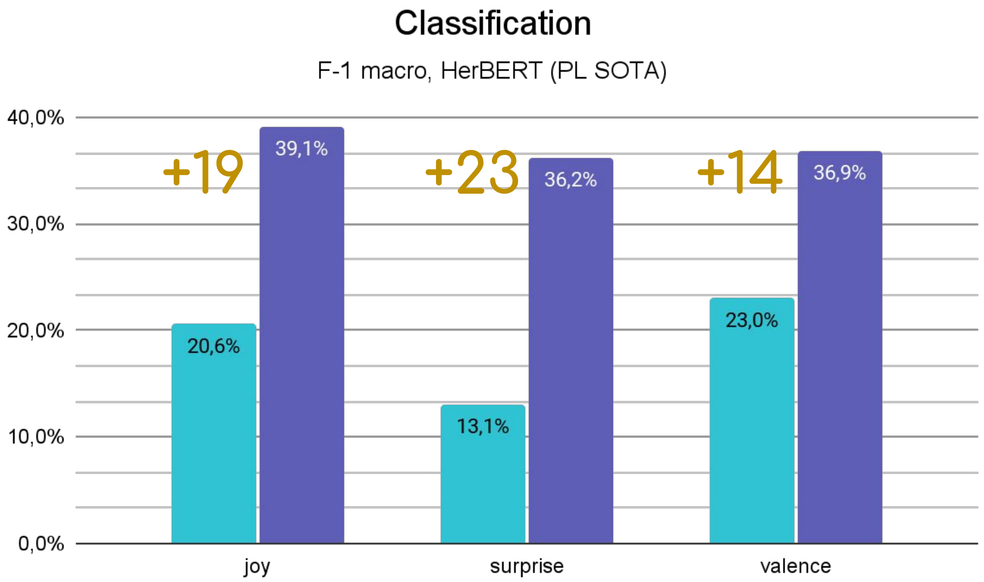
Other language models:

- XLM-RoBERTa
- fastText + LSTM
- Polish RoBERTa

Worse by <1.5 p.p.



CLASSIFICATION: three emotional dimensions



(1) Text only

Model based only on text embeddings

(3) Text and PEB

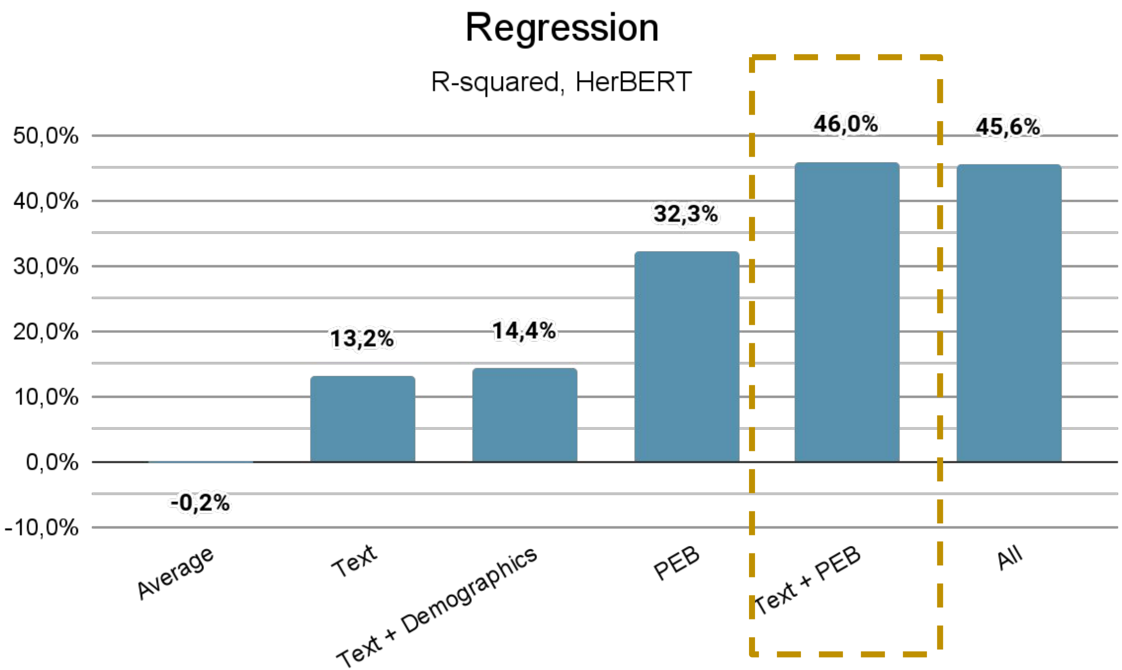
Model prepared on text embeddings and Personal Emotional Bias

REGRESSION: all emotions aggregated

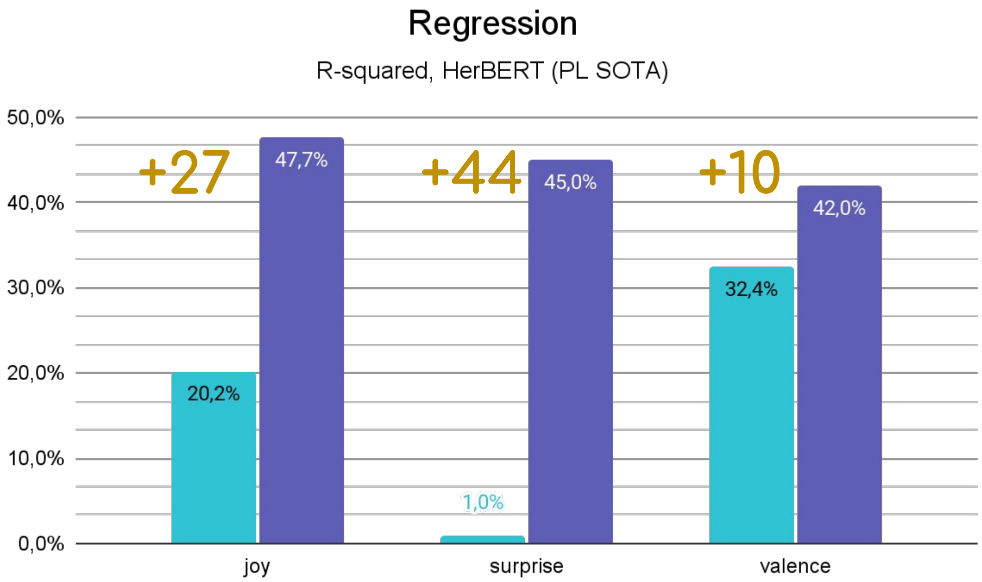
Other language models:

- XLM-RoBERTa
- fastText + LSTM
- Polish RoBERTa

Worse by 3 p.p.



REGRESSION: three emotions



(1) Text only

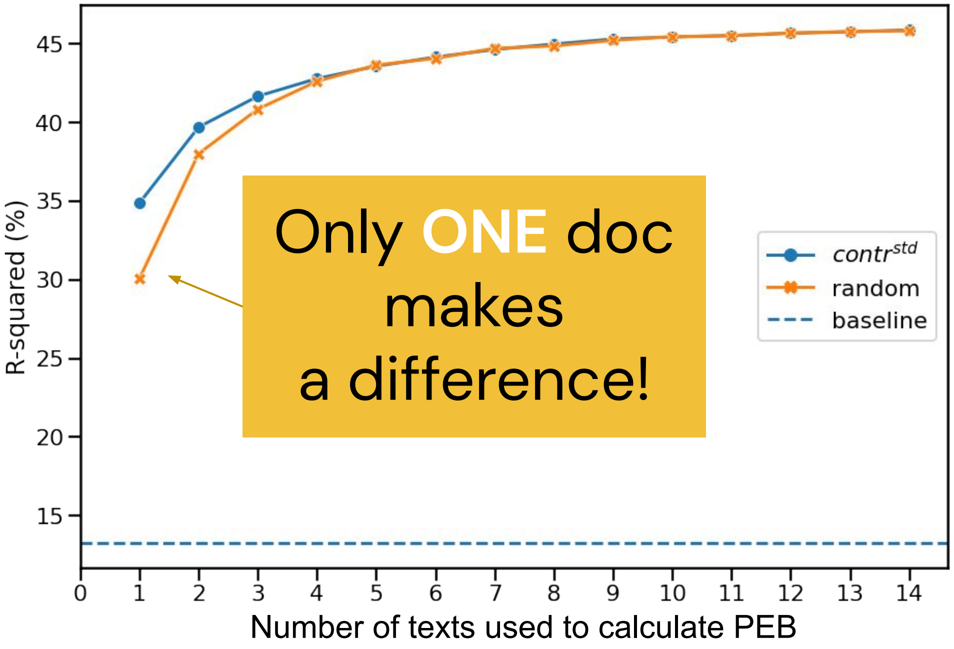
Model based only on text embeddings

(3) Text and PEB

Model prepared on text embeddings and Personal Emotional Bias

How many texts are needed for PEB?

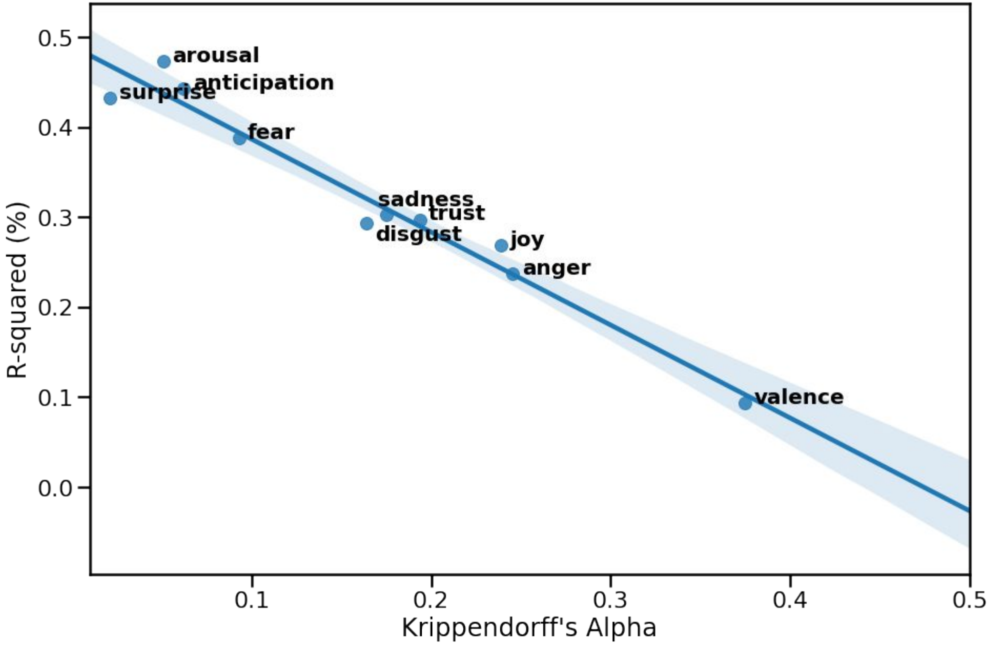
(1) TXT – baseline
(3) TXT+PEB:
- random texts for PEB
- most controversial texts for PEB



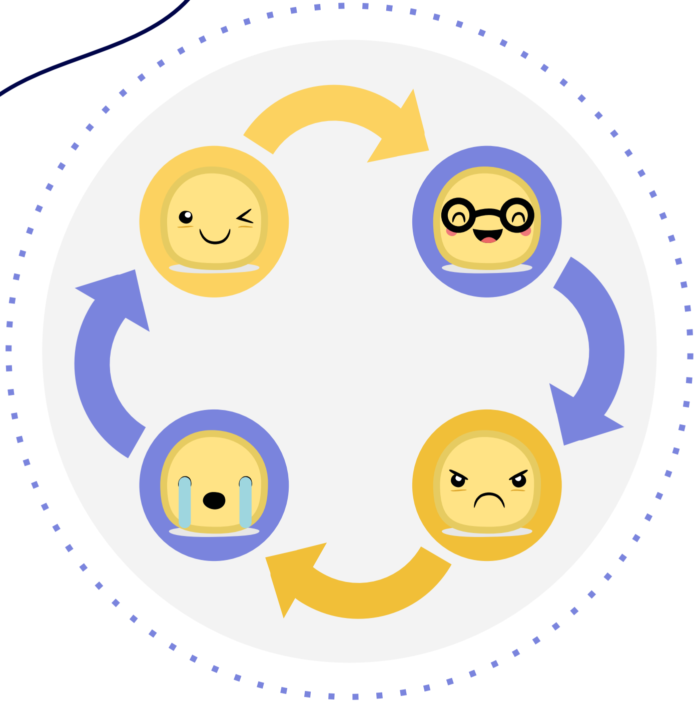
All emotions, HerBERT

AGREEMENT LEVEL (controversy) vs. performance

PEB-only model Performance



Controversy in the collection



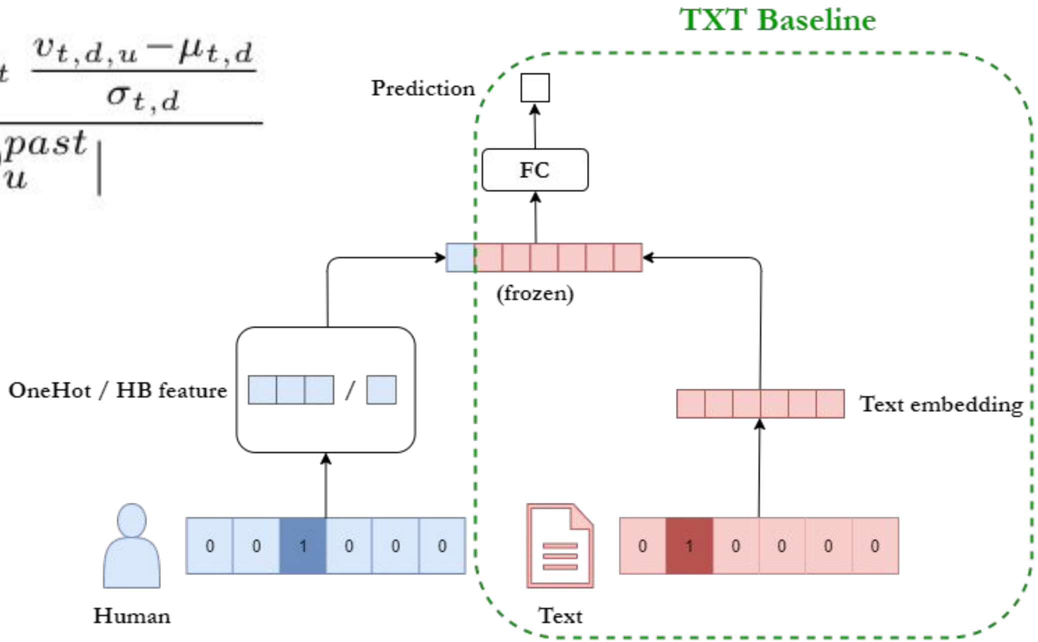
7

RESEARCH ON MULTIPLE TASKS AND MODELS

Wiki Detox: Attack,
Aggression, Toxicity
+ Emotions
ICDM2021: [Koc21b]

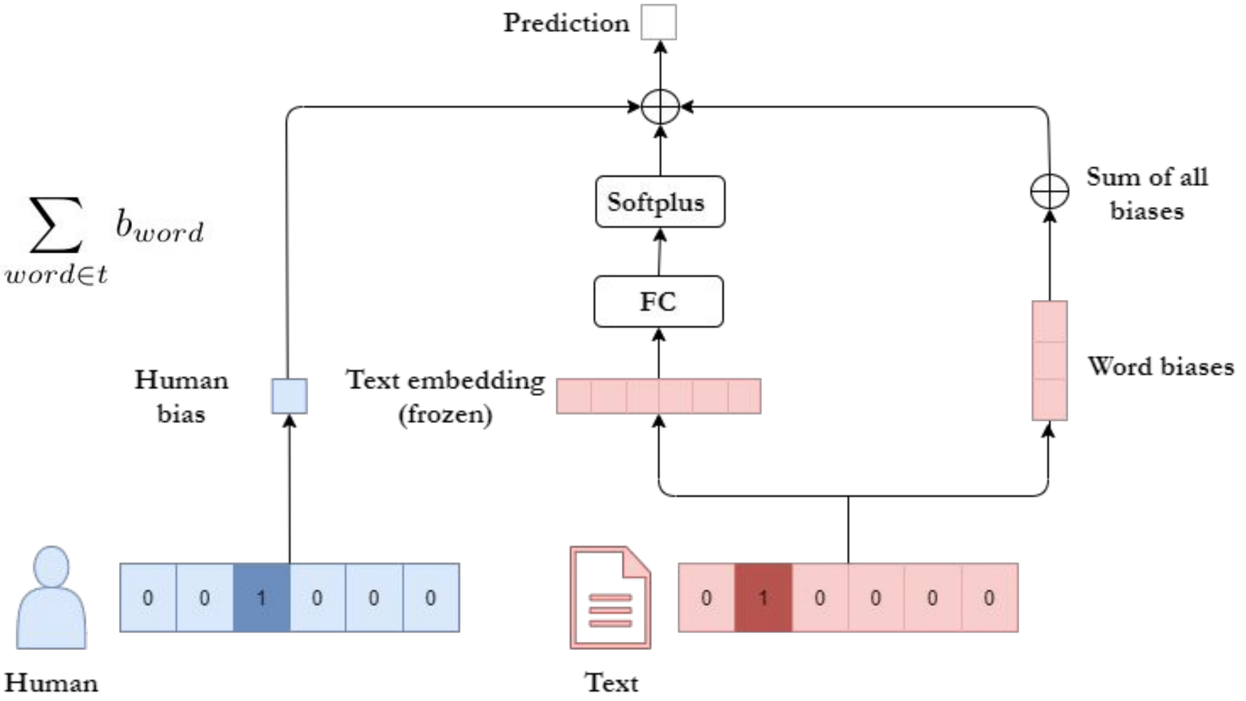
MODELS:
Baseline (TXT) & OneHot ID & HuBi-Formula

$$HB(u,t) = \frac{\sum_{d \in D_u^{past}} \frac{v_{t,d,u} - \mu_{t,d}}{\sigma_{t,d}}}{|D_u^{past}|}$$



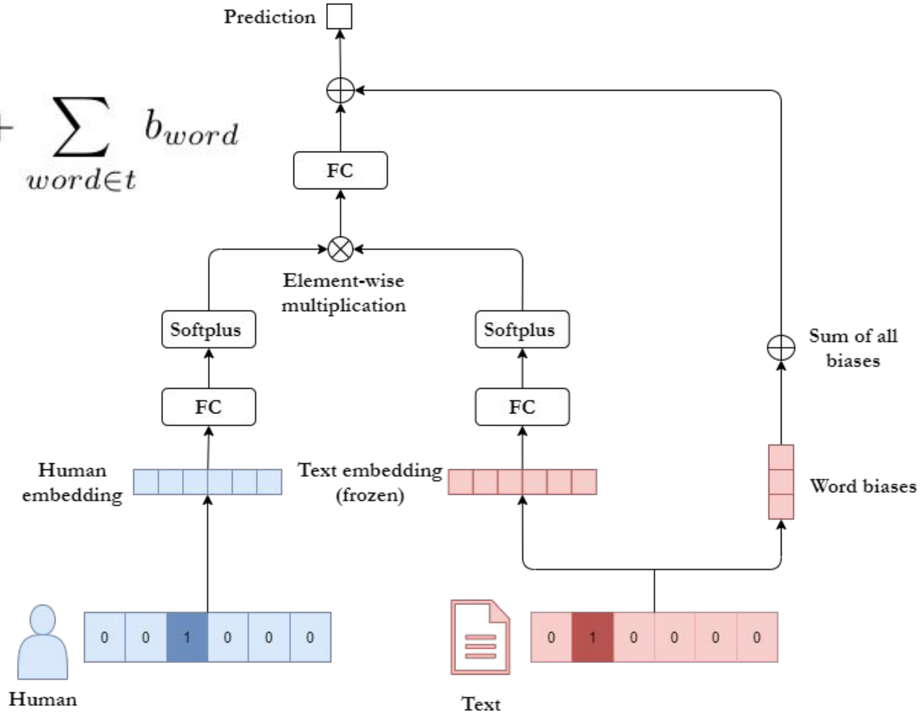
MODELS:
HuBi-Simple: learned human bias

$$y(t,u) = a(W_T x_t) + b_u + \sum_{word \in t} b_{word}$$



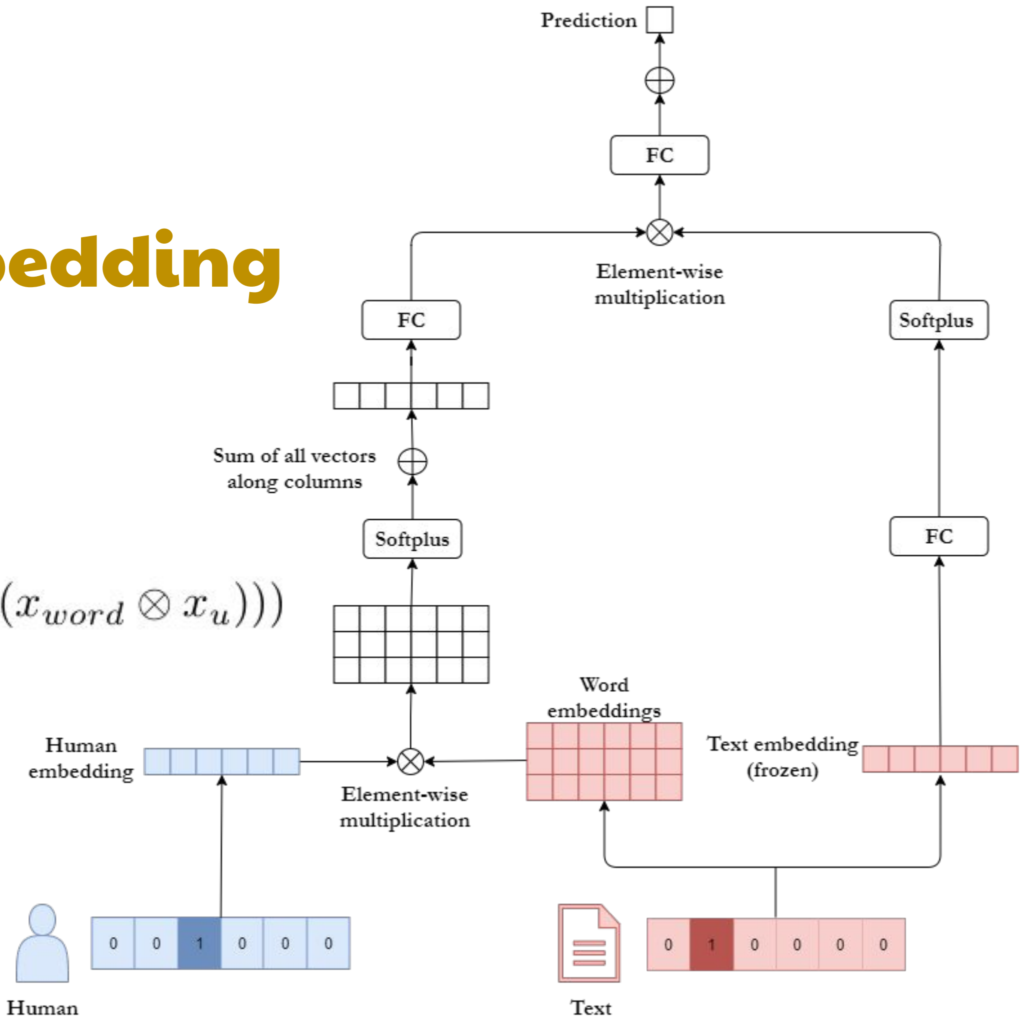
MODELS:
HuBi-Medium: learned human embedding

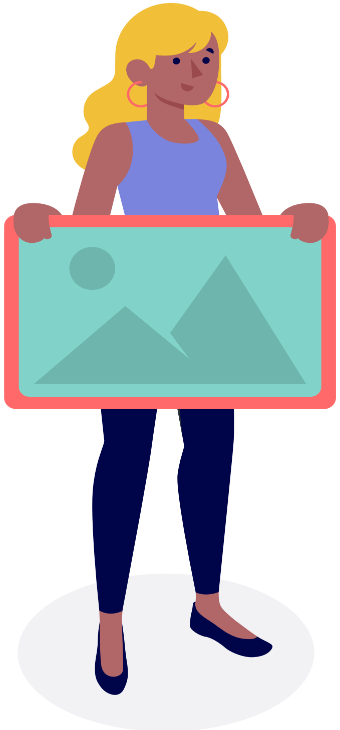
$$y(t,a) = W_{TU}(a(W_T x_t) \otimes a(W_U x_u)) + \sum_{word \in t} b_{word}$$



MODELS: HuBi-Complex: human-word embedding

$$y(t,a) = W(a(W_T x_t) \otimes W_{WU}(\sum_{word \in t} a(x_{word} \otimes x_u)))$$



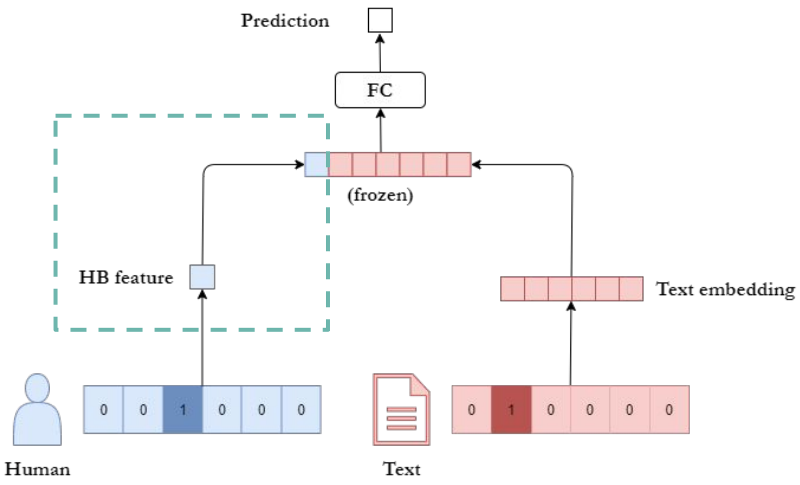


7a

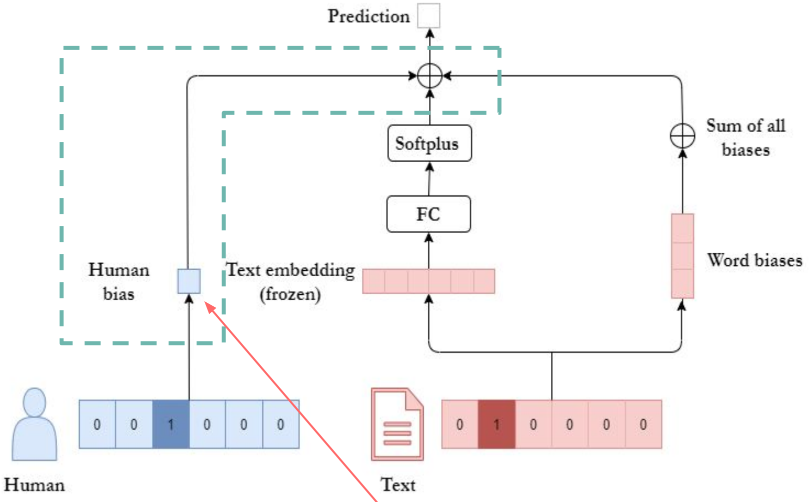
MULTIPLE TASKS: RESULTS

Wiki Detox + Emotions

FORMULA vs. LEARNED BIAS
HB feature vs. HuBi-Simple (learned bias)



vs.

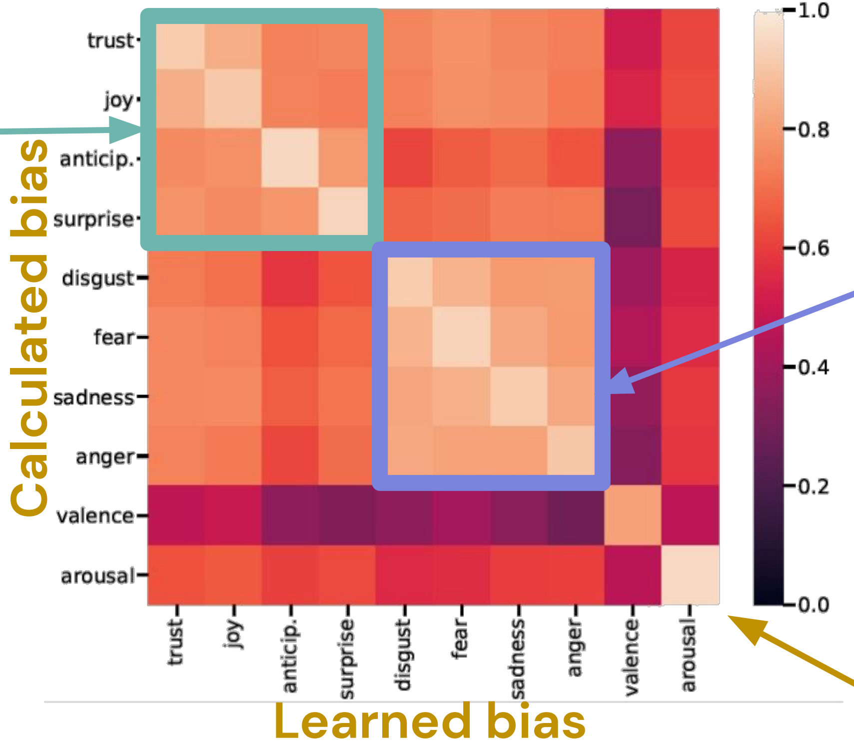


HB calculated feature (**formula**)

HuBi-Simple: **learned** human bias

FORMULA vs. LEARNED BIAS
Correlation between biases

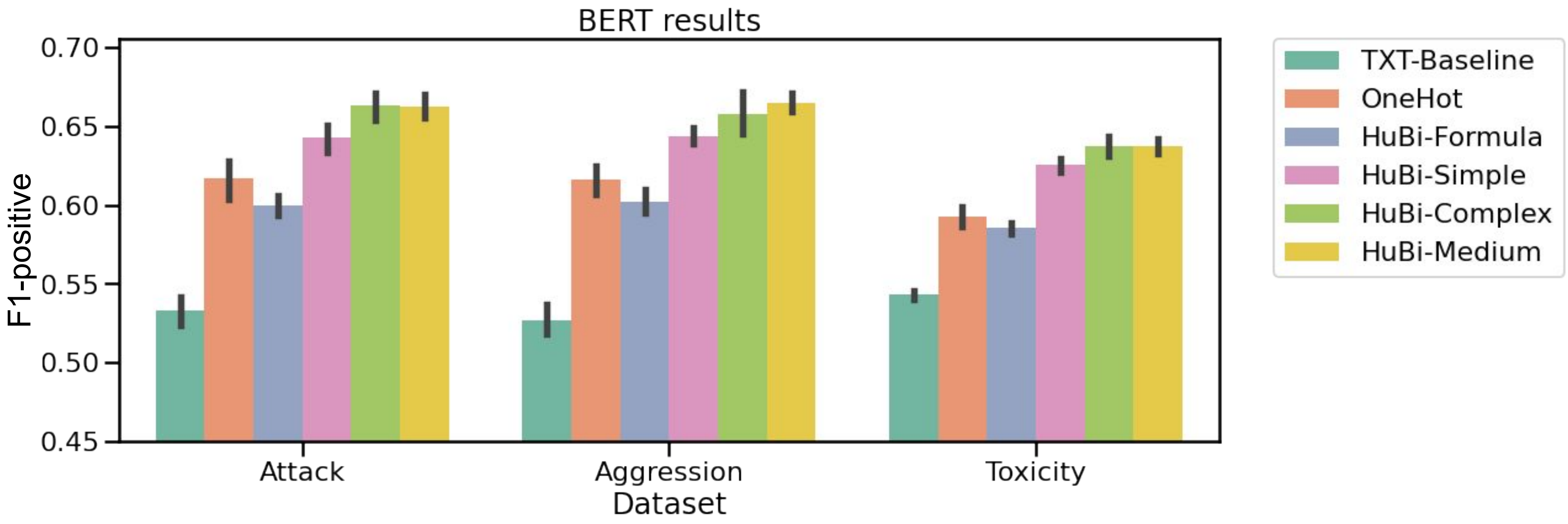
Positive emotions
are highly correlated
73% and more



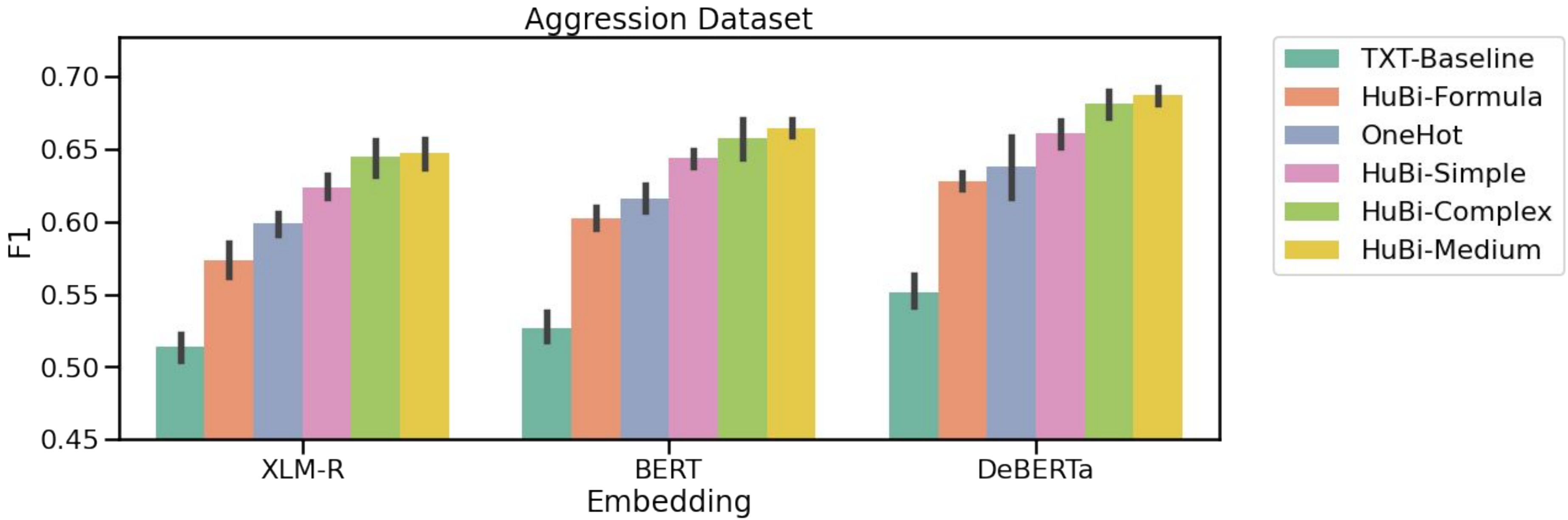
Negative emotions
are highly correlated
80% and more

Biases are
very highly correlated
90% and more
(diagonal)

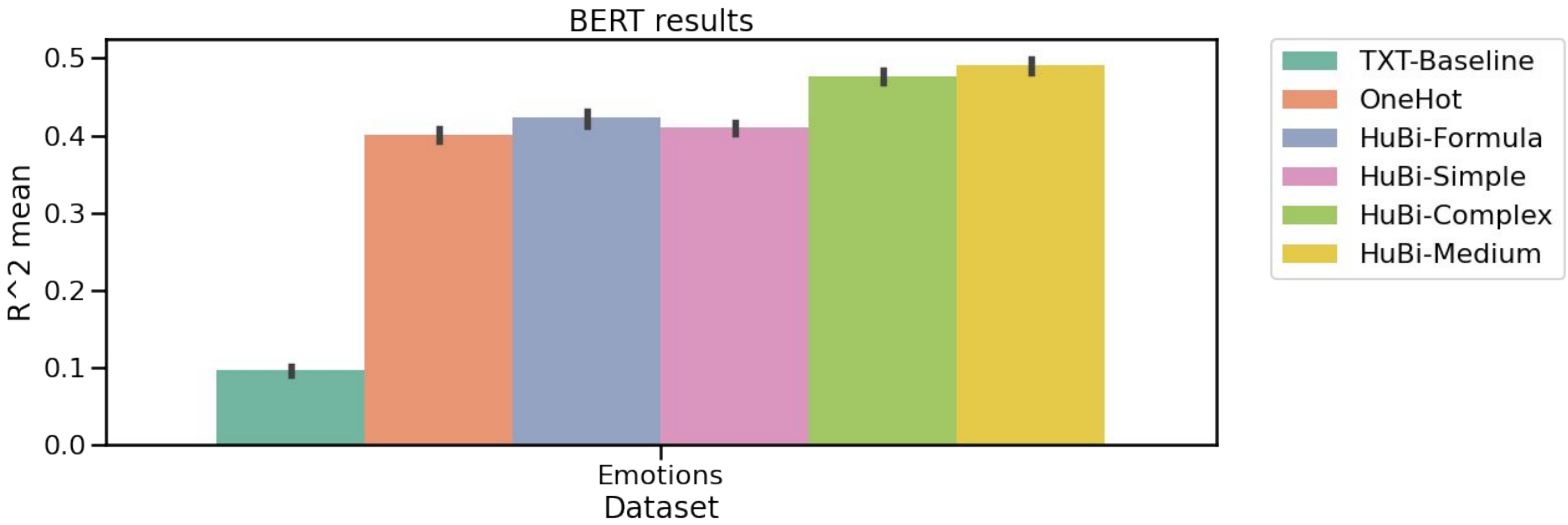
WIKI: results on three datasets



WIKI: Results on Aggression Data

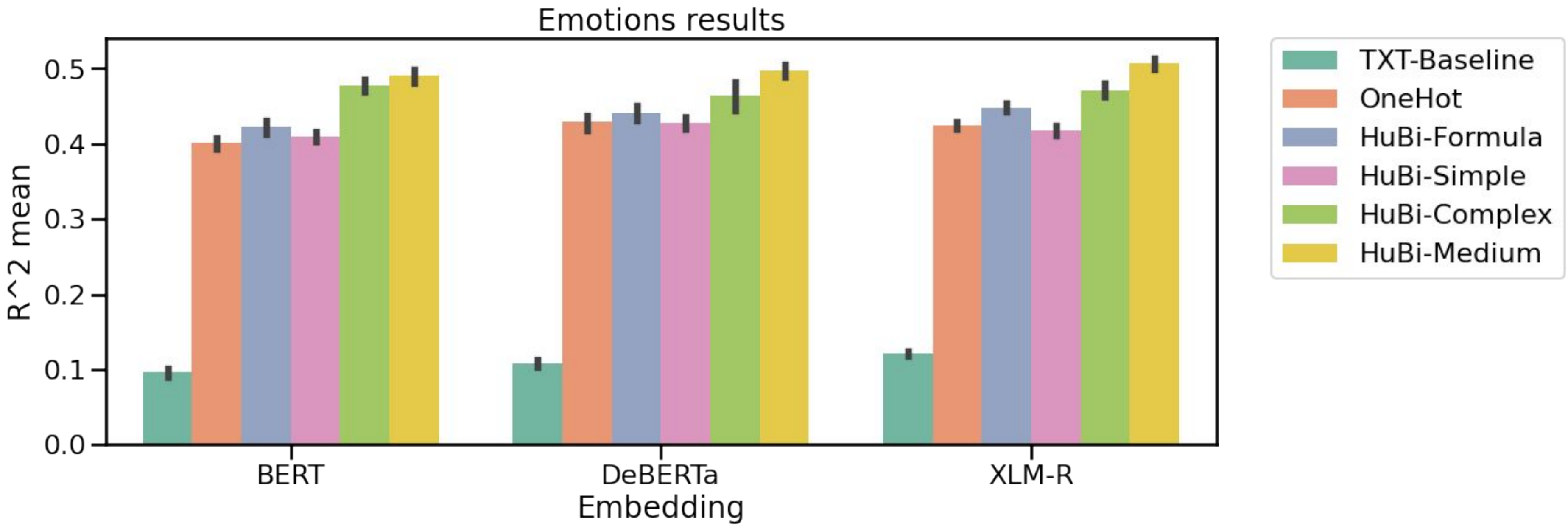


EMOTIONS: Results



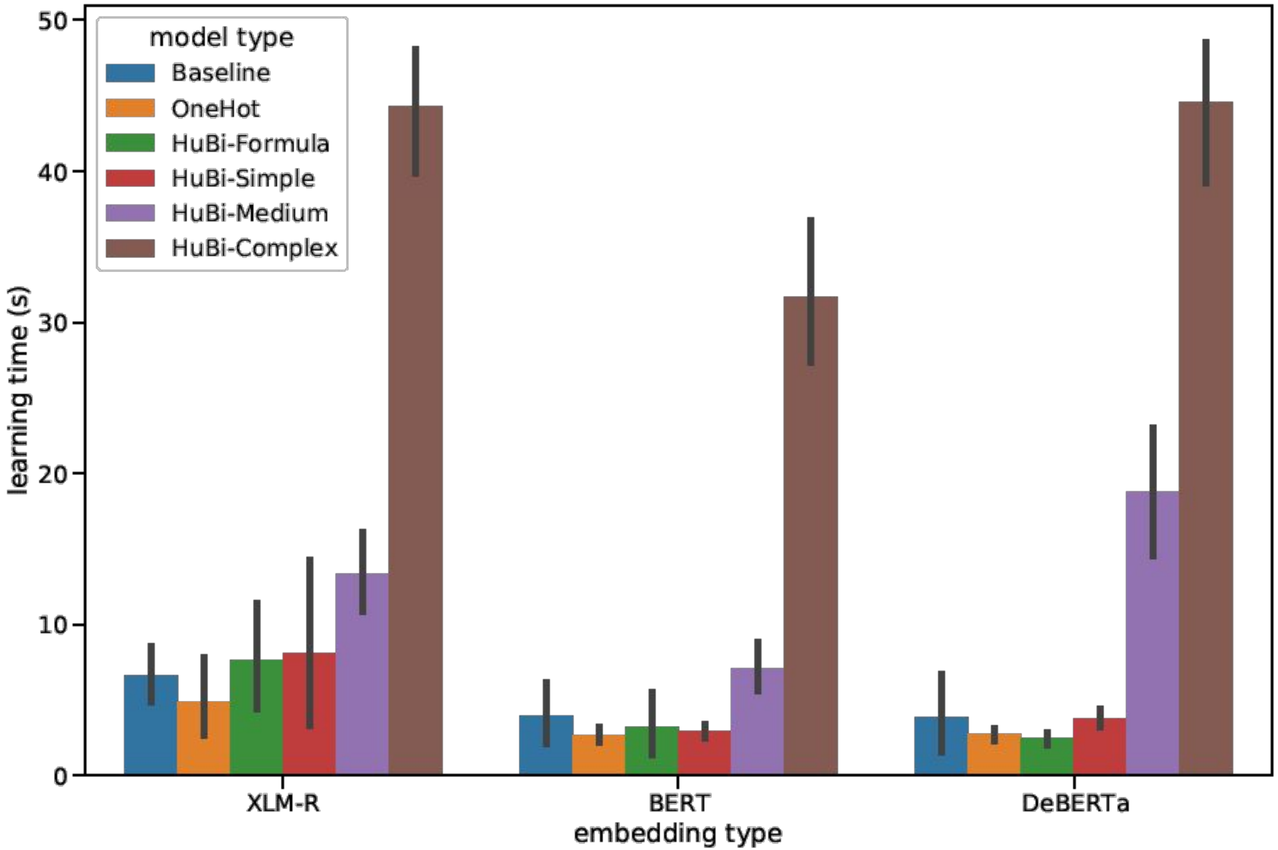
Multivariate regression

EMOTIONS: Results

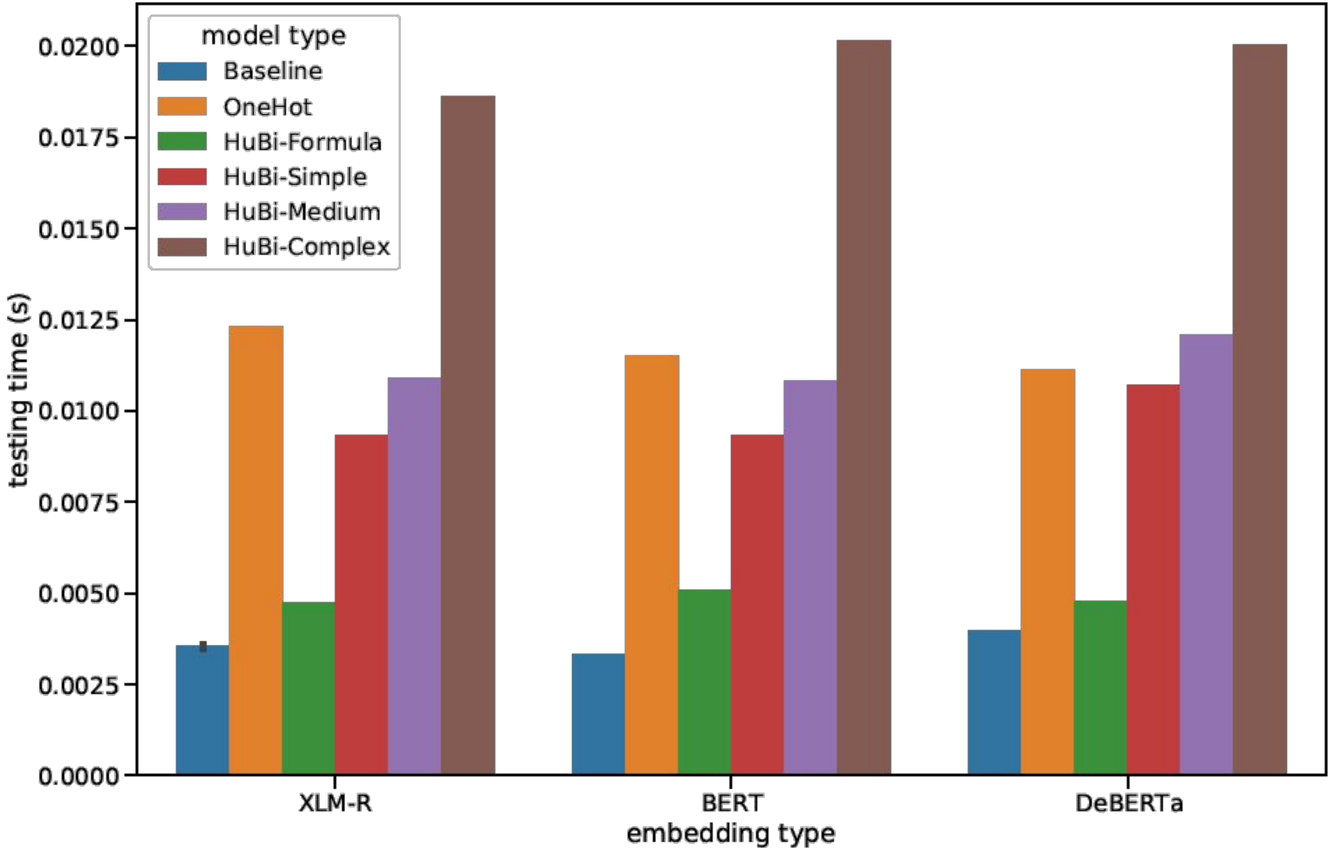


Multivariate regression

TRAINING TIME: emotions



TESTING TIME: emotions



8

CONCLUSIONS

CONCLUSIONS #1

PNLP vs. GNLP

Personalized methods **ALWAYS** perform better than the generalized ones

PNLP vs. language

Each PNLP method gains **much more than** language models

Diversity

Conformity, **Controversy** and **Human Bias** deliver vital information about the user

Few docs is enough

Even four docs provide user information that improves reasoning (5-6 docs for emotional texts)

CONCLUSIONS #2

Validation

Train/dev/test split should be based on **users** instead of texts

Application

Our PNLP methods can be applied to **any** subjective task

Demographics

Demographic data only slightly improves reasoning

Data

Human-centered annotations are crucial for personalised NLP

TEAM

Przemysław Kazienko

Jan Kocoń

Kamil Kanclerz

Julita Bielaniewicz

Marcin Gruza

Piotr Miłkowski

BIBLIOGRAPHY

[Koc21a] Kocoń J., Figas A., Gruza M., Puchalska D., Kajdanowicz T., Kazienko P.: *Offensive, aggressive, and hate speech analysis: from data-centric to human-centred approach*. **Information Processing and Management**, 58(5) 2021, art. 102643.

[Kan21] Kanclerz K., Figas A., Gruza M., Kajdanowicz T., Kocoń J., Puchalska D., Kazienko P.: *Controversy and Conformity: from Generalized to Personalized Aggressiveness Detection*. **ACL 2021**, 5915–5926.

[Mił21] Miłkowski P., Gruza M., Kanclerz K., Kazienko P., Grimling D., Kocoń J.: *Personal Bias in Prediction of Emotions Elicited by Textual Opinions*. **ACL 2021**, Student Research Workshop, 248–259.

[Koc21b] Kocoń J., Gruza M., Bielaniewicz J., Grimling D., Kanclerz K., Miłkowski P., Kazienko P.: *Learning Personal Human Biases and Representations for Subjective Tasks in Natural Language Processing*, IEEE **ICDM 2021**, Dec. 2021.

Take-home message

Personalized NLP is much better than generalized for all subjective tasks

Thank you for your attention!

Q&A

THE END