Szkoła Letnia AI TECH

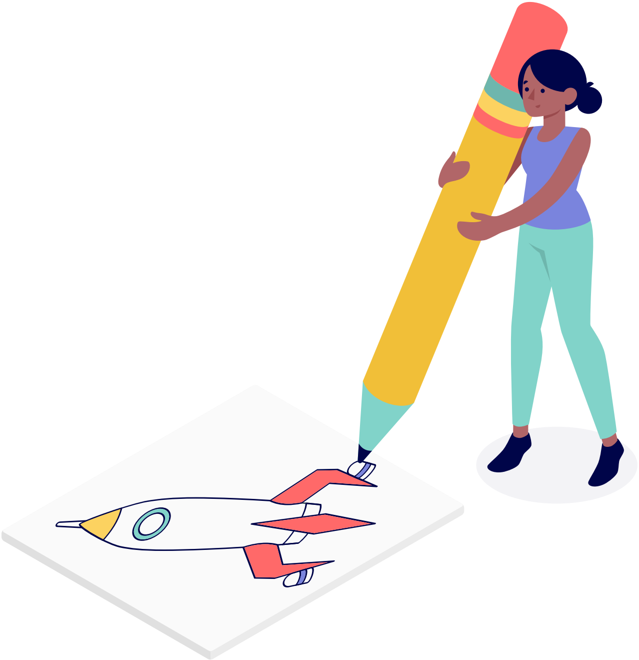
Subjective Problems in NLP

Kamil Kanclerz
Department of Artificial Intelligence
Wroclaw University of Science and Technology, Poland

Fundusze Europejskie
Rzeczpospolita Polska
Ministerstwo Cyfryzacji
Unia Europejska
Europejski Fundusz Rozwoju Regionalnego

AGENDA

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

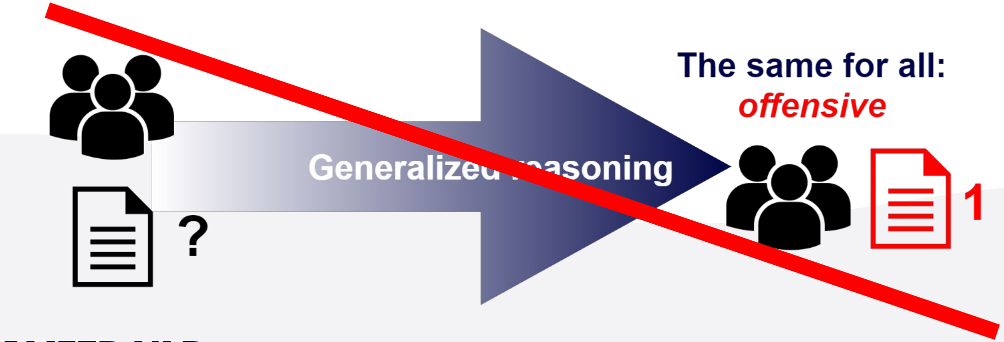
Do you think, it is aggressive or not?

MOTIVATION

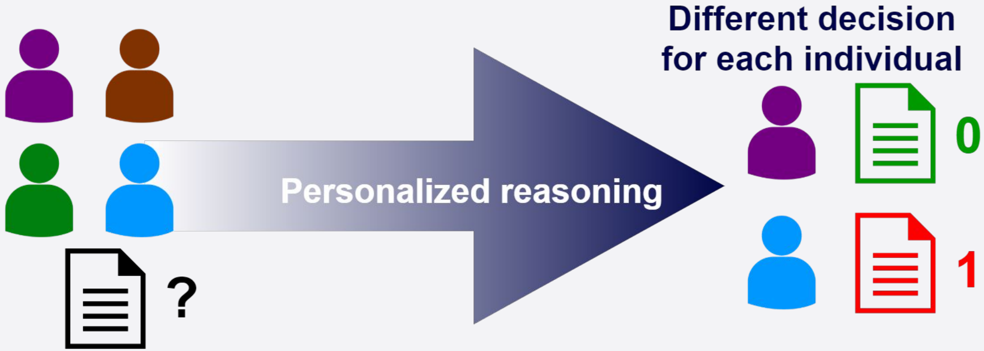
COMMON GENERALIZED NLP

MOTIVATION

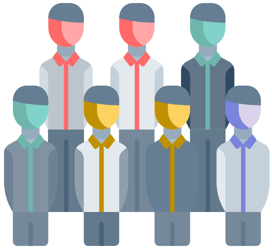
COMMON GENERALIZED NLP



OUR PERSONALIZED NLP



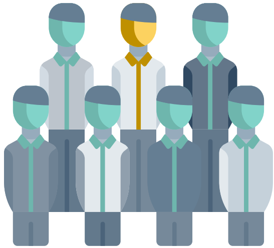
MOTIVATION



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. Persuasiveness
   m. Subjective perception of sentiment polarization

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

3. **Mixed**
   a. Conversations

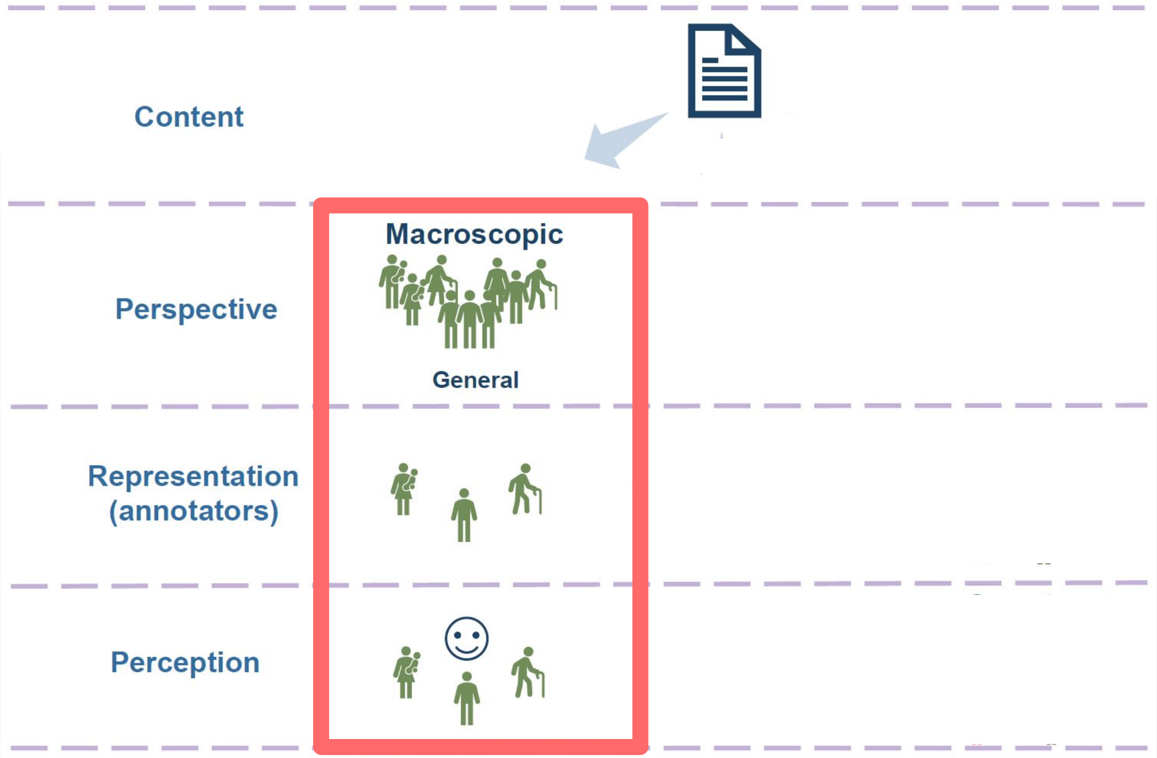
The tasks often overlap

3

PERSPECTIVES

[Koc21a]

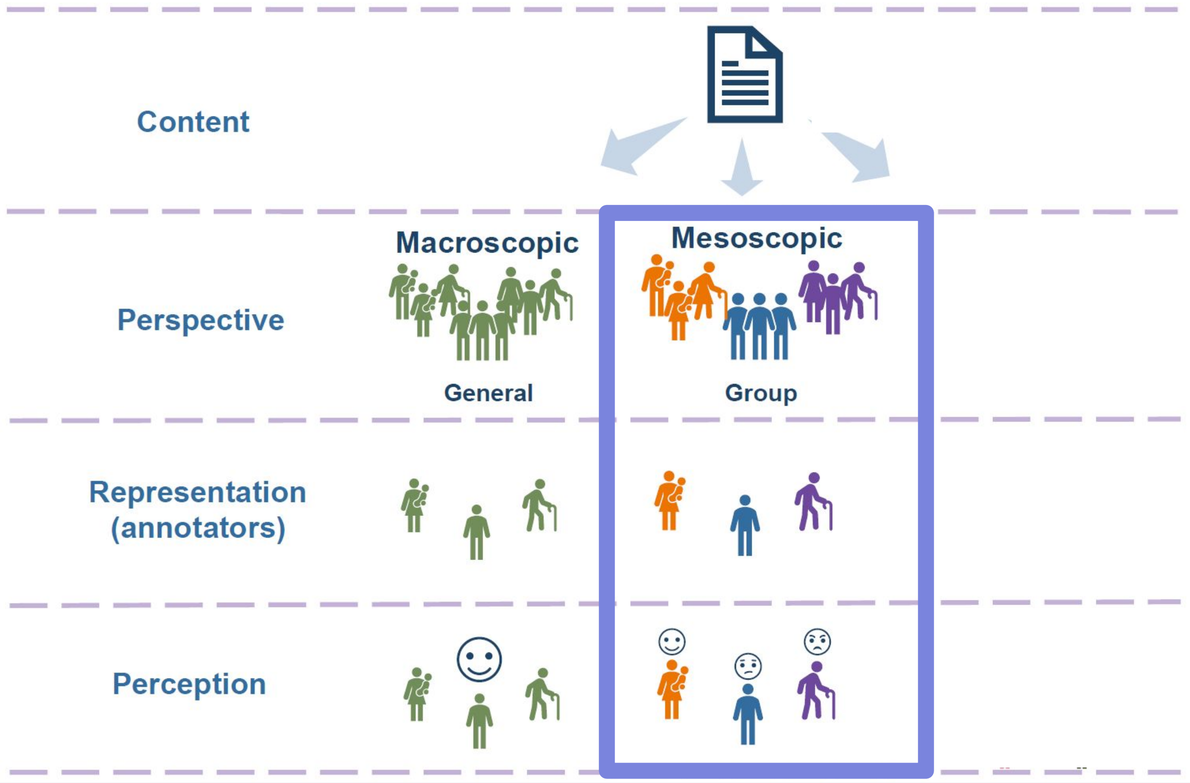
PERSPECTIVES: MACROSCOPIC



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)

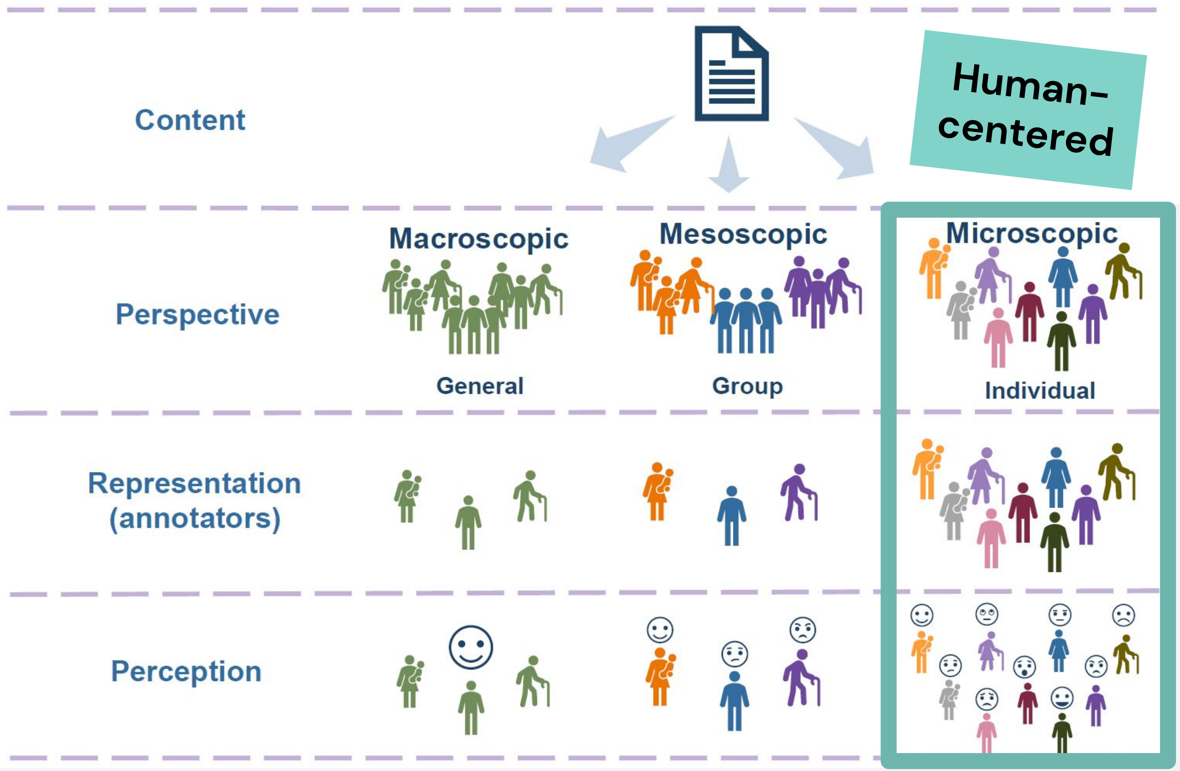
PERSPECTIVES: MESOSCOPIC



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

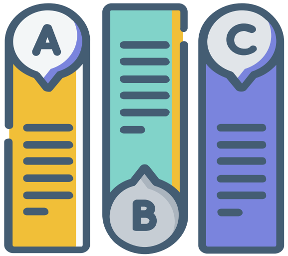
Agreed, generalized labels are useless

Usually obtained by majority voting

4

RESEARCH ON OFFENSIVE CONTENT

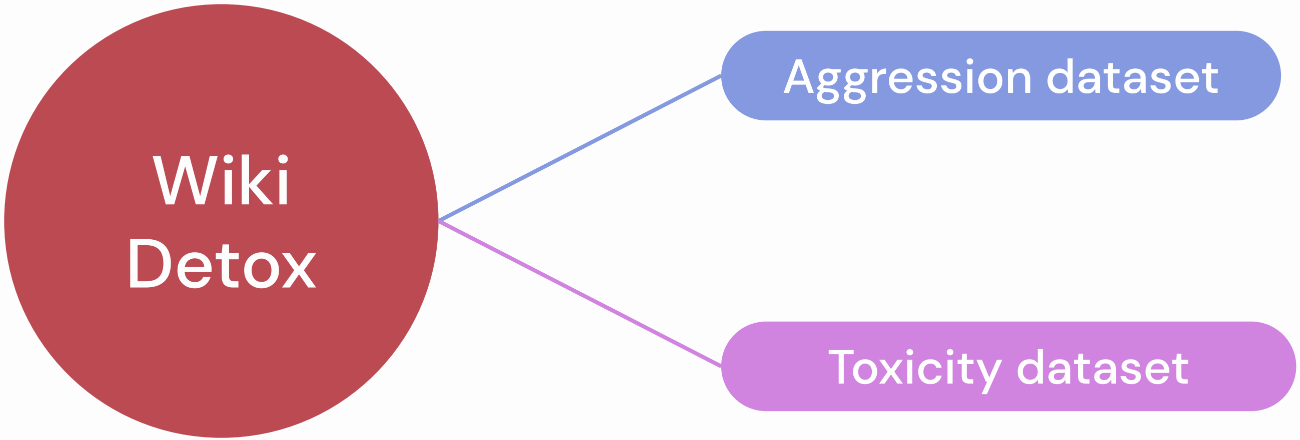
[Koc21a, Kan21, Koc21b]



4a

OFFENSIVE CONTENT: ANNOTATED DATA

WIKI DETOX DATASETS (English)



Publicly available

WIKI: Aggression

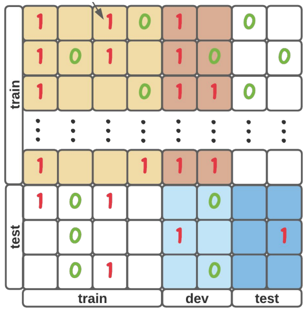
Classes	Texts	People
2	115,864	4,053

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

WIKI: Toxicity

Classes	Texts	People
2	159,686	4,301

Annotations	Controversial Texts
1,598,289	40.5 %

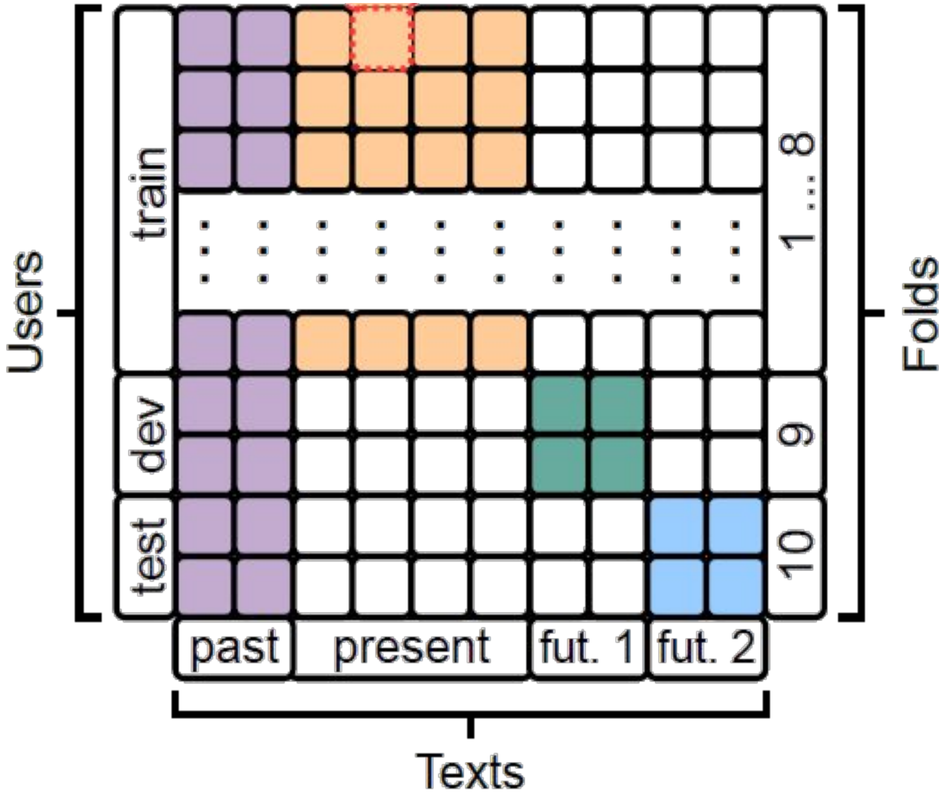


4b

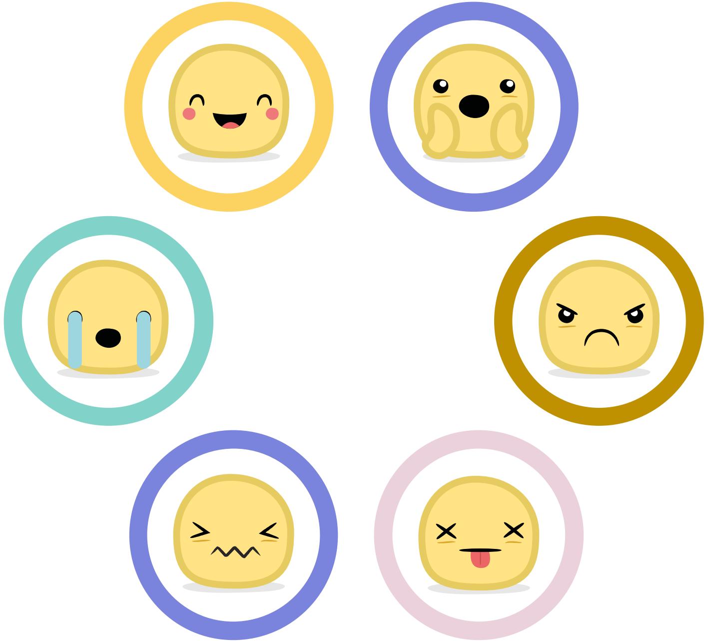
OFFENSIVE CONTENT: DATA SPLIT

Train-dev-test

DATA SPLIT



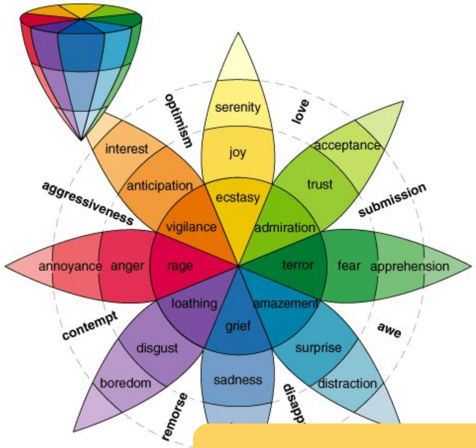
Users
train
dev
test
1 ... 8
9
10
Folds
past
present
fut. 1
fut. 2
Texts



5

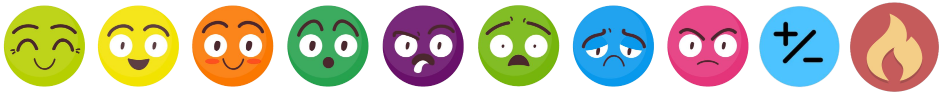
RESEARCH ON EMOTIONAL CONTENT PERCEPTION

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

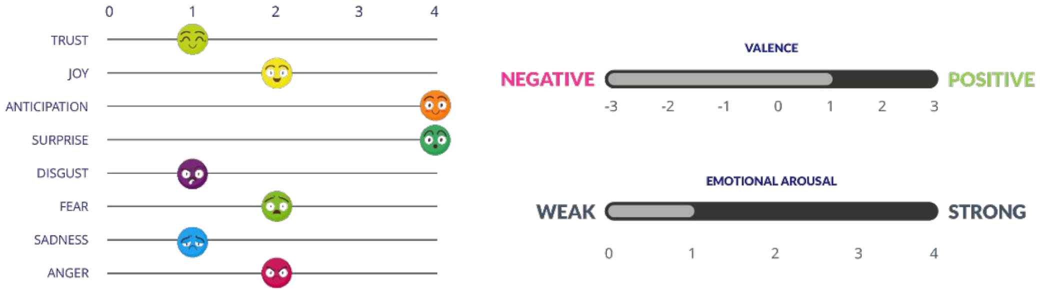
EMOTIONAL TEXTS: example

Sentimenti

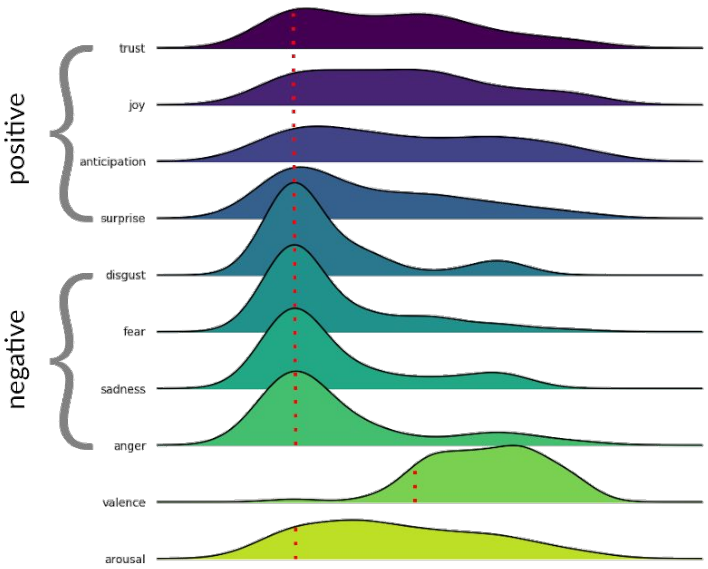
Example opinion

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

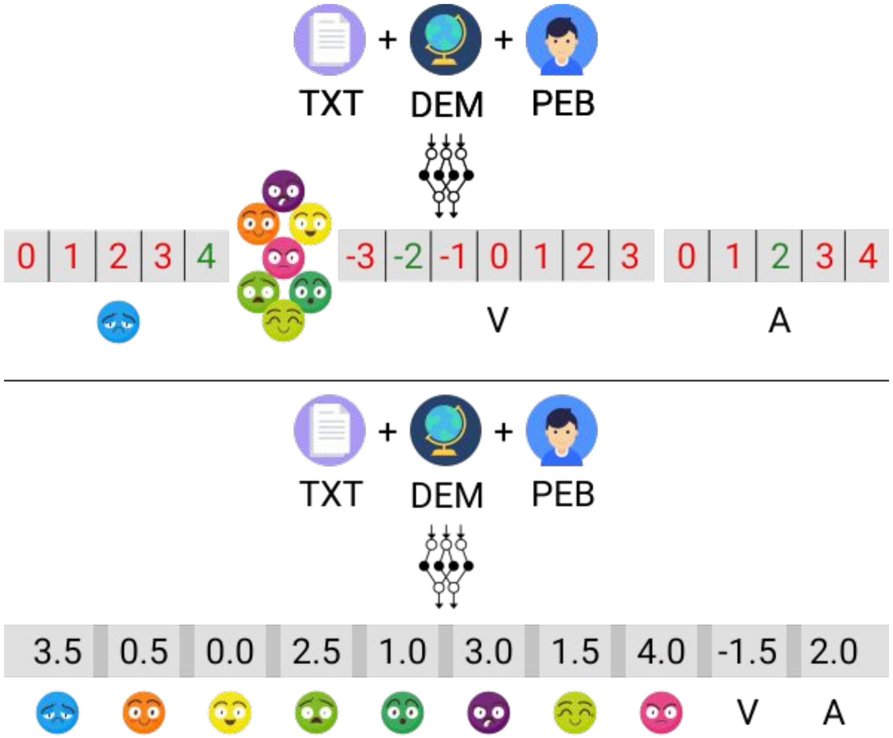


EMOTIONAL EXPERIMENTS

Sentimenti

(1) Multi-task classification

(2) Multivariate regression

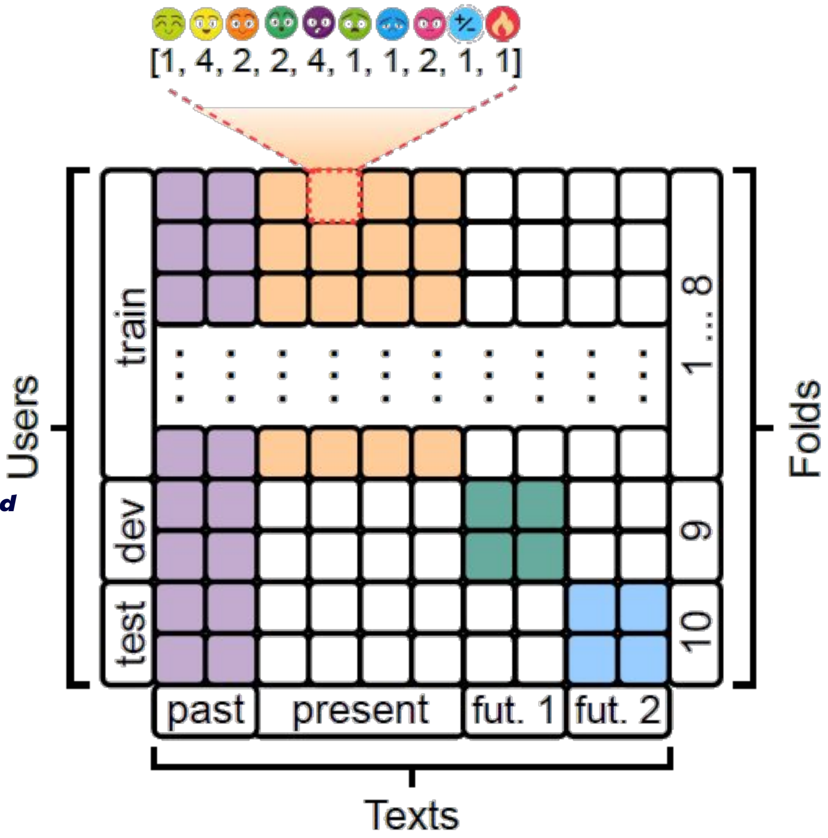


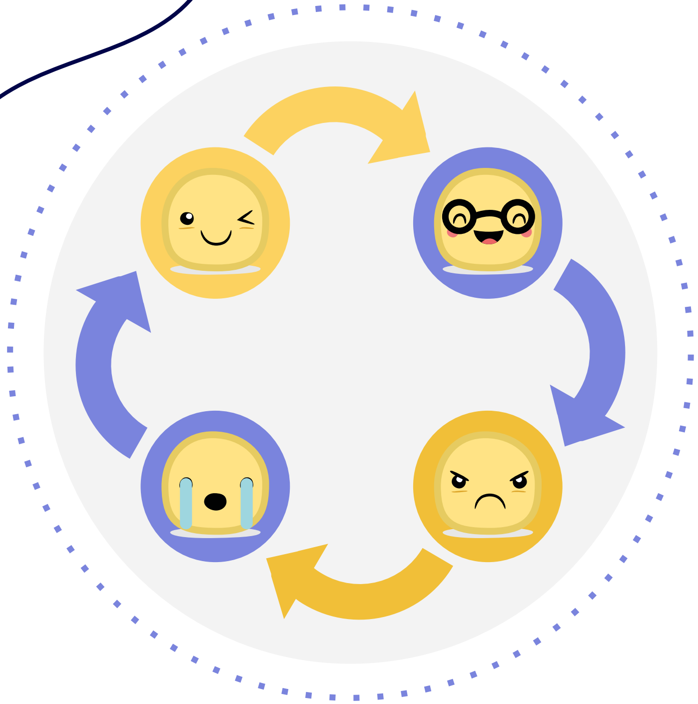
EMOTIONAL DATA SPLIT

PEB: Z-score

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

u - user
t - task / dimension
d - document
D_u^{past} - set of documents annotated by user ***u*** from the ***past*** fold
$v_{t,d,u}$ - value assigned to task ***t*** for document ***d*** by user ***u***
$\mu_{t,d}$ - mean value assigned to task ***t*** for document ***d***
$\sigma_{t,d}$ - standard deviation of values assigned to task ***t*** for document ***d***





6

RESEARCH ON MULTIPLE TASKS AND MODELS

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

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

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

u - **user**

t - **task / dimension**

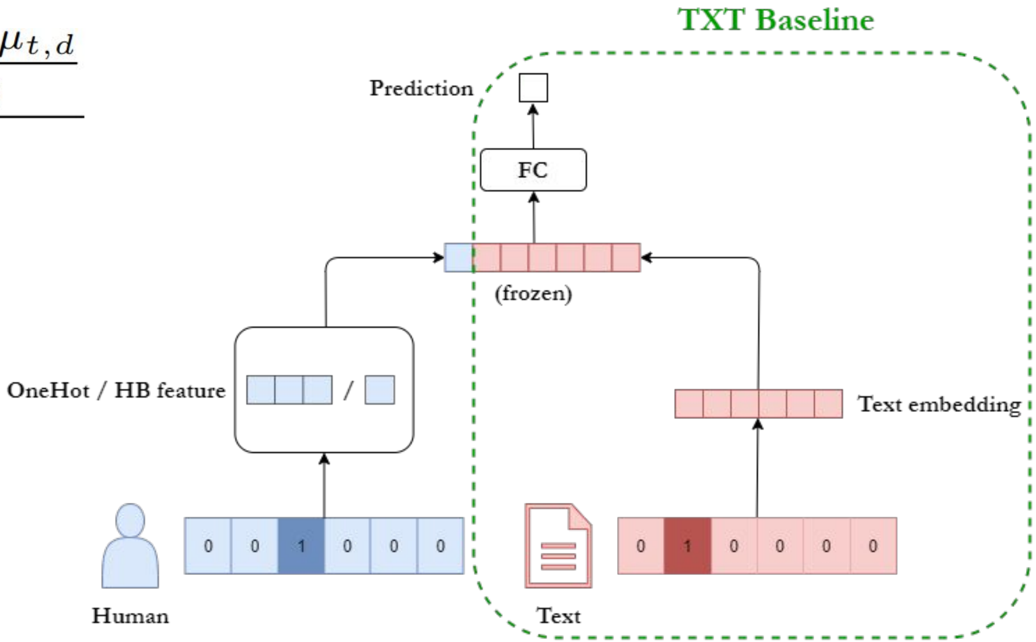
d - **document**

D_u^{past} - **set of documents annotated by user *u* from the *past* fold**

$v_{t,d,u}$ - **value assigned to task *t* for document *d* by user *u***

$\mu_{t,d}$ - **mean value assigned to task *t* for document *d***

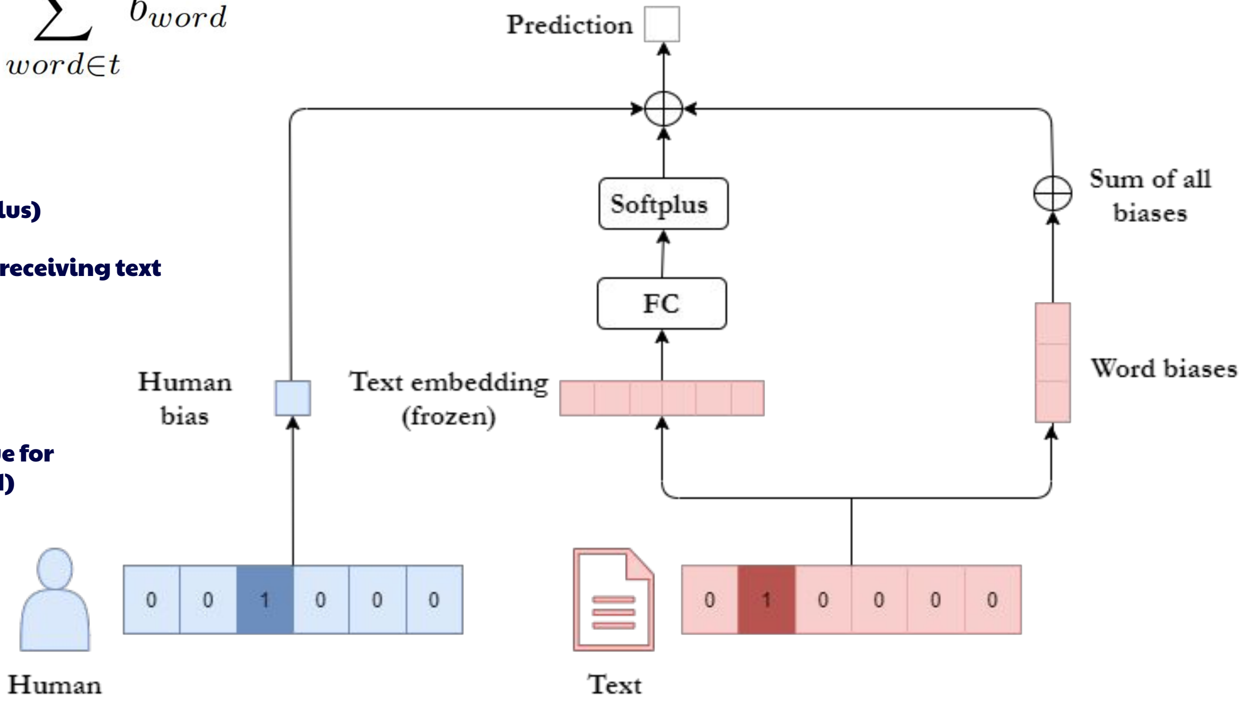
$\sigma_{t,d}$ - **standard deviation of values assigned to task *t* for document *d***



MODELS:
HuBi-Simple: learned human bias

$$y(u,d) = a(W_D x_d) + b_u + \sum_{word \in t} b_{word}$$

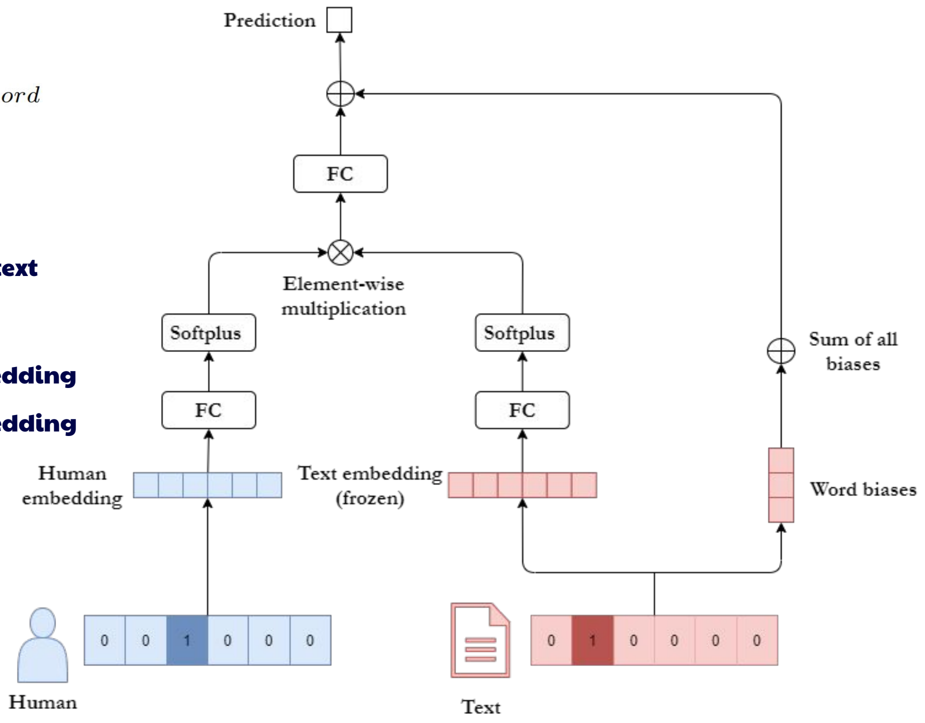
u	**- user**
d	**- text**
a	**- activation function (softplus)**
W_D	**- weights vector of FC layer receiving text embedding**
x_d	**- text embedding**
b_u	**- learned human bias**
b_{word}	**- word bias (avg. label value for train texts containing word)**



MODELS:
HuBi-Medium: learned human embedding

$$y(u,d) = W_{DU}(a(W_D x_d) \otimes a(W_U x_u)) + \sum_{word \in d} b_{word}$$

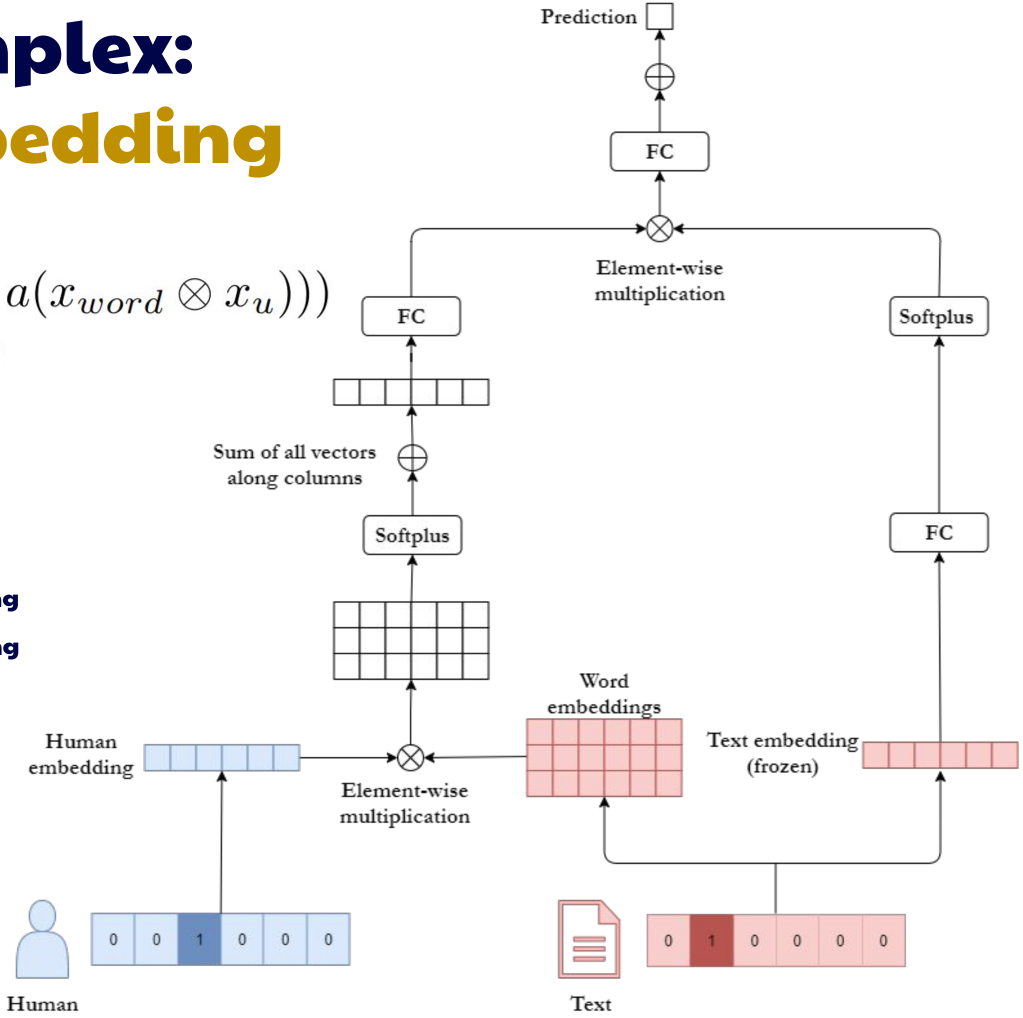
u	**- user**
d	**- text**
W_{DU}	**- weights vector of FC layer receiving user and text embedding**
a	**- activation function (softplus)**
W_D	**- weights vector of FC layer receiving text embedding**
x_d	**- weights vector of FC layer receiving text embedding**
W_U	**- weights vector of FC layer receiving user vector**
x_u	**- user one-hot vector**
b_{word}	**- word bias (avg. label value for train texts containing word)**

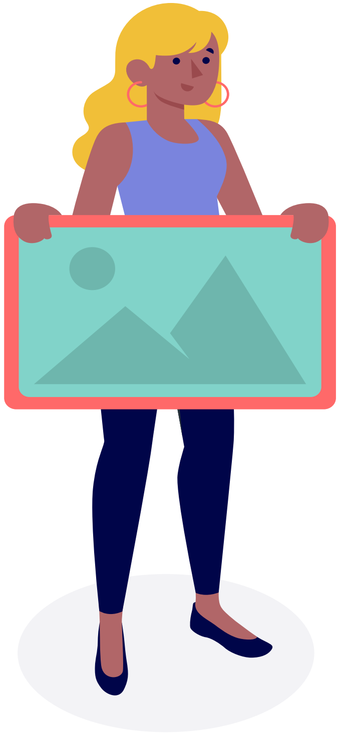


MODELS: HuBi-Complex: human-word embedding

$$y(u,d) = W(a(W_D x_d) \otimes W_{DU}(\sum_{word \in d} a(x_{word} \otimes x_u)))$$

u	- user
d	- text
W	- weights vector of the last FC layer
a	- activation function (softplus)
W_D	- weights vector of FC layer receiving text embedding
x_d	- weights vector of FC layer receiving text embedding
W_{DU}	- weights vector of FC layer receiving user and text embedding
W_U	- weights vector of FC layer receiving user vector
x_{word}	- word embedding (averaged subwords)
x_u	- user one-hot vector



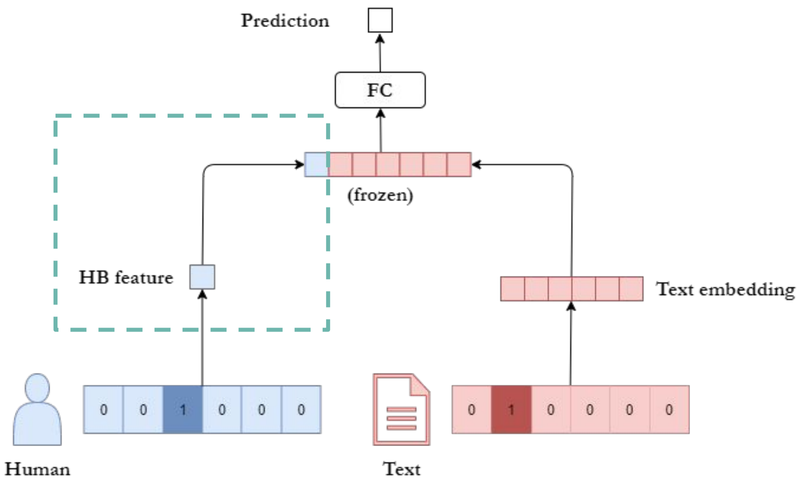


6a

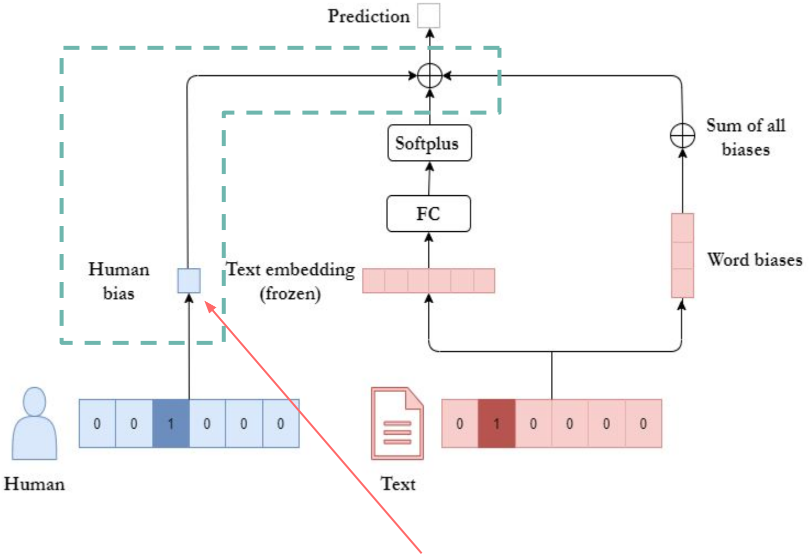
MULTIPLE TASKS: RESULTS

Wiki Detox + Emotions

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



vs.

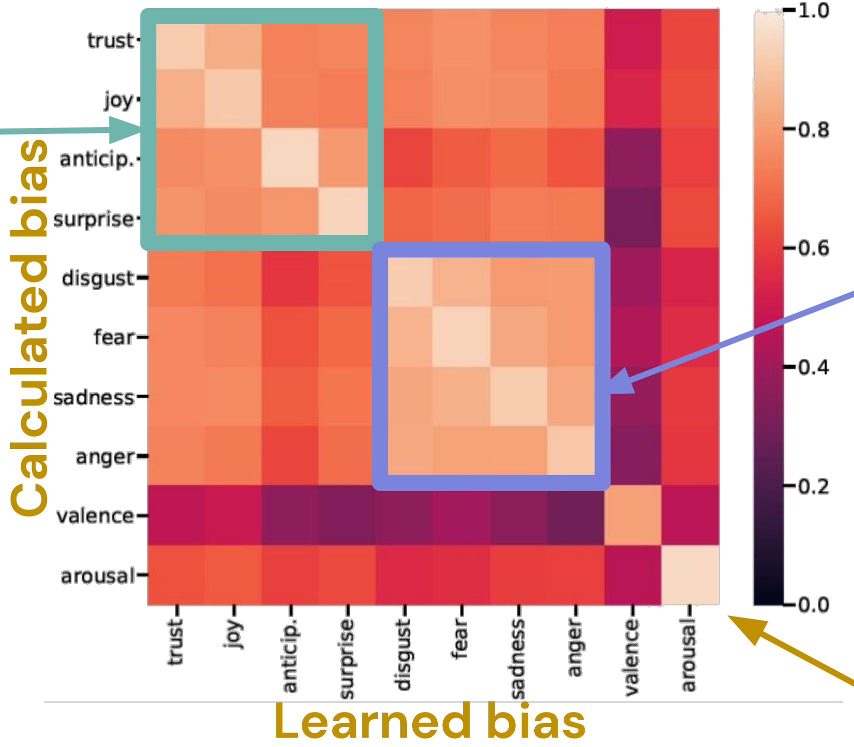


HB calculated feature (**formula**)

HuBi-Simple: **learned** human bias

FORMULA vs. LEARNED BIAS
Correlation between biases

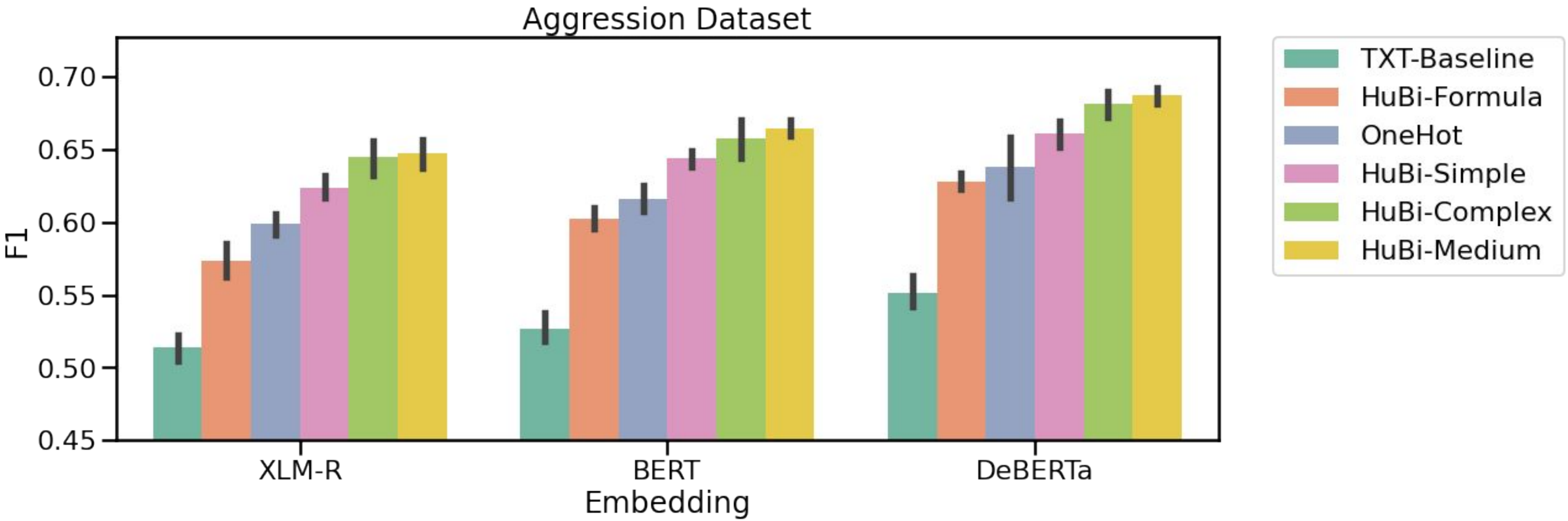
Positive emotions are highly correlated **73% and more**



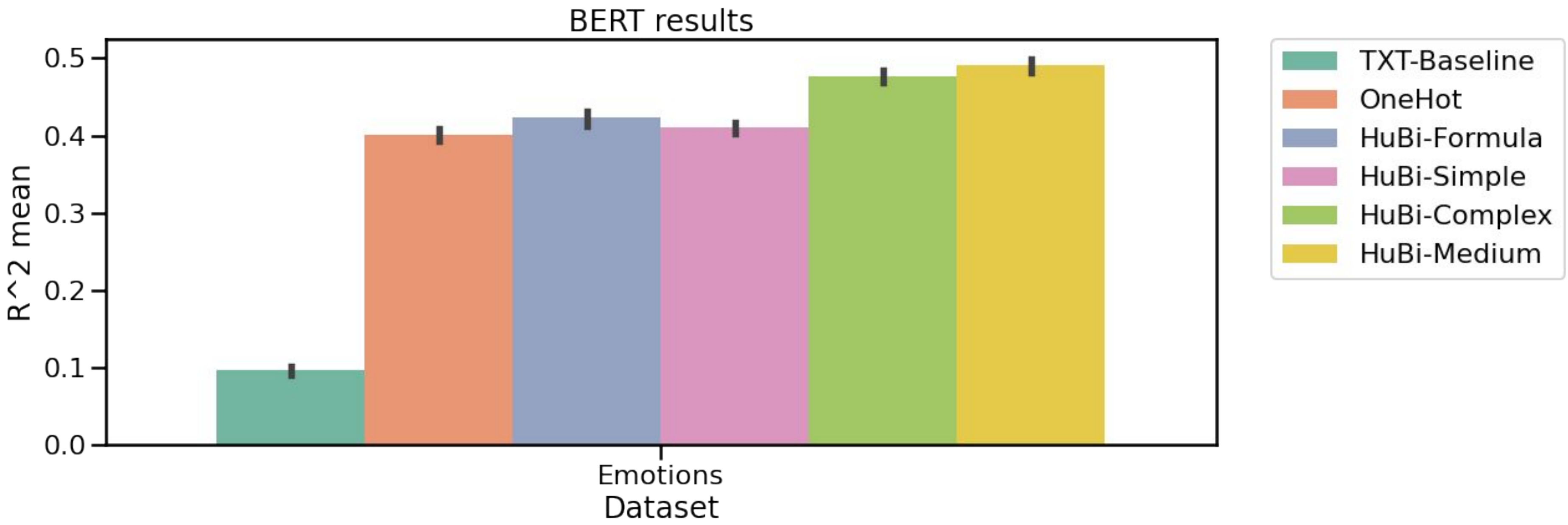
Negative emotions are highly correlated **80% and more**

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

WIKI: Results on Aggression Data

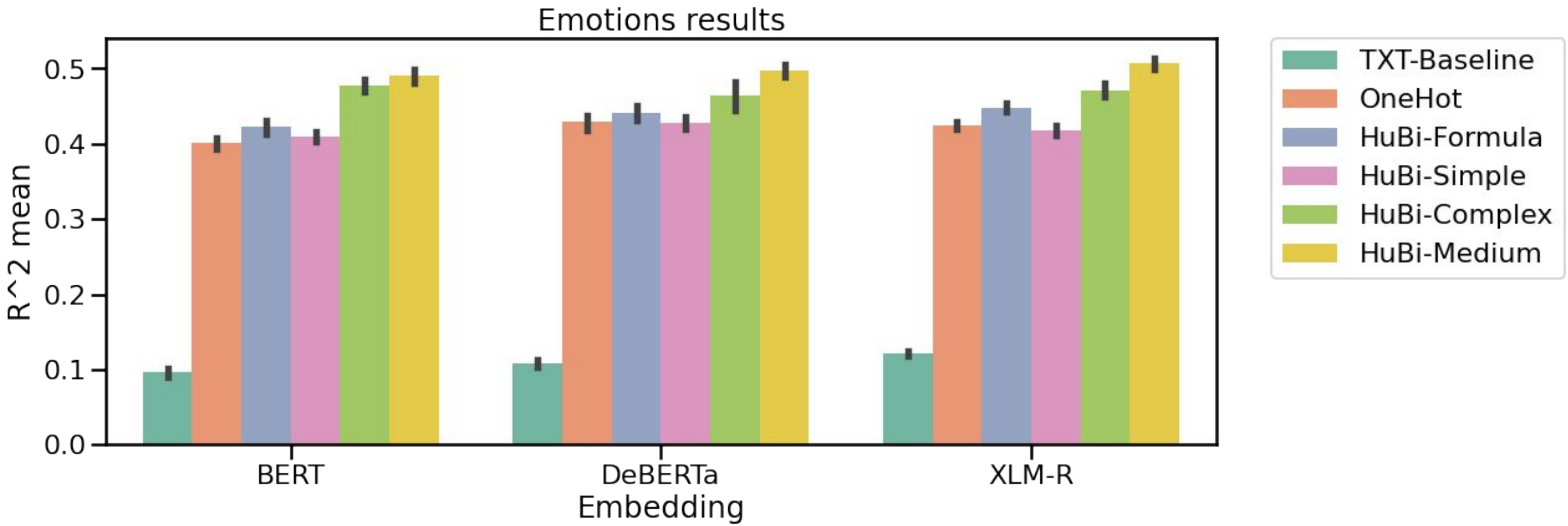


EMOTIONS: Results



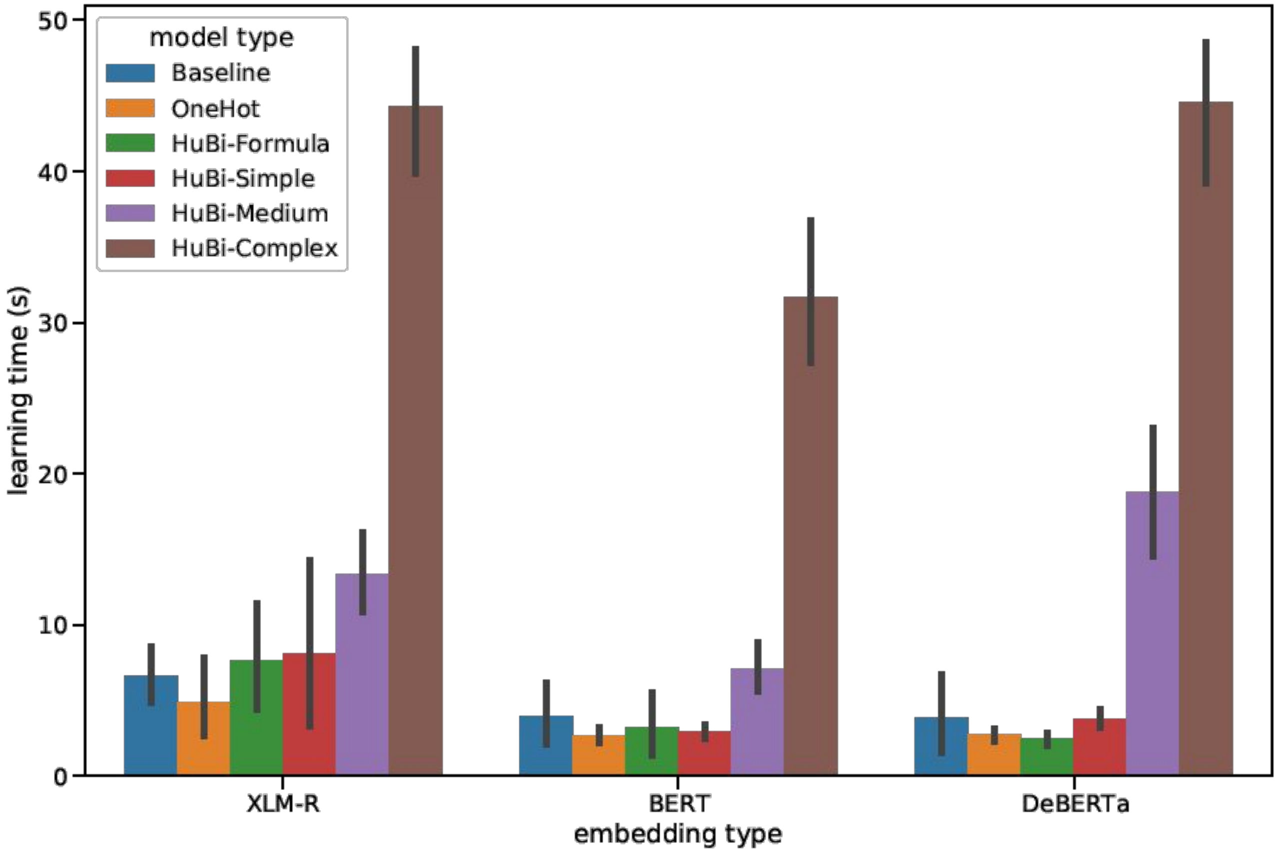
Multivariate regression

EMOTIONS: Results



Multivariate regression

TRAINING TIME: emotions



7

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

BIBLIOGRAPHY

[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**.

[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.

[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.

[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.

Take-home message

Personalized NLP is much better than generalized for all subjective tasks

Thank you for your attention!

Q&A

THE END