Investigating the Role of Emotions in Collaborative Software Development

Nicole Novielli

University of Bari, Italy

Collaborative Development Group

@NicoleNovielli

nicole.novielli@uniba.it

http://collab.di.uniba.it/nicole/

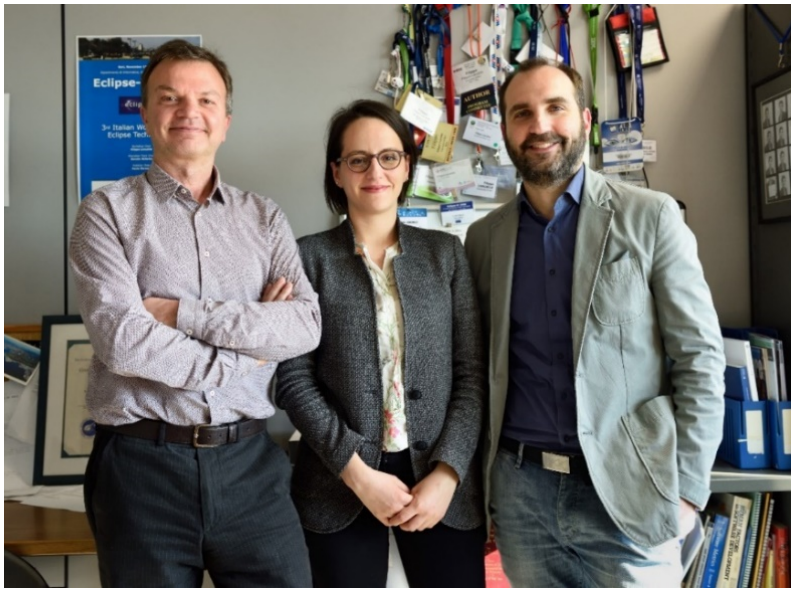
Rearch Topics

- *Affective Computing*
- *Sentiment Analysis*
- *Human Aspects in Software Engineering*

People at Collab

collab.di.uniba.it

- Faculty
  - **Filippo Lanubile**
  - **Fabio Calefato**
  - **Nicole Novielli**



- PhD Students
  - **Giuseppe Iaffaldano**
  - **Daniela Girardi**

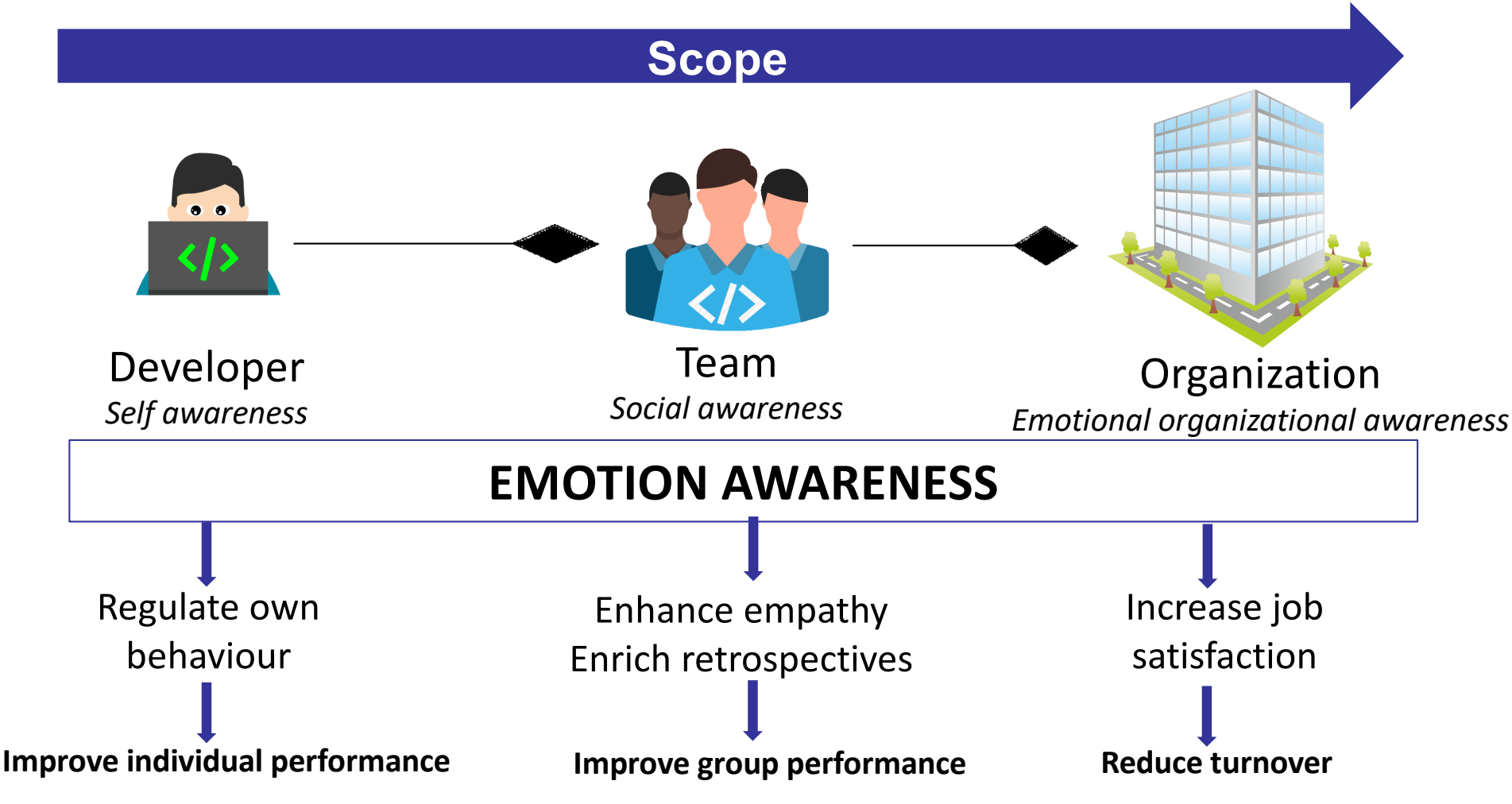
- Graduate students and final-year undergraduates
- Visiting Professors and Researchers

Research at COLLAB

Software development as
an intense collaborative process

collaborative / social
software engineering

Sensing Developers' Emotions



Investigating the Role of Emotions in the Social Programmer Ecosystem

- Research question: **getting emotional** while communicating with developers**: good or bad?**
- Model: combining message proper social factors and affective factors
- Expected output:
  - SE-specific sentiment analysis tool and emotion classifier
  - Evidence-based netiquette

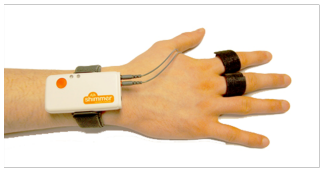
Sensor-based emotion detection

Research question: Can we acquire physiological measures from noninvasive, low cost sensors to accurately predict emotions?

Electroencephalography (EEG)

BrainLink: EEG raw signal, attention and meditation levels

Galvanic Skin Response (GSR)

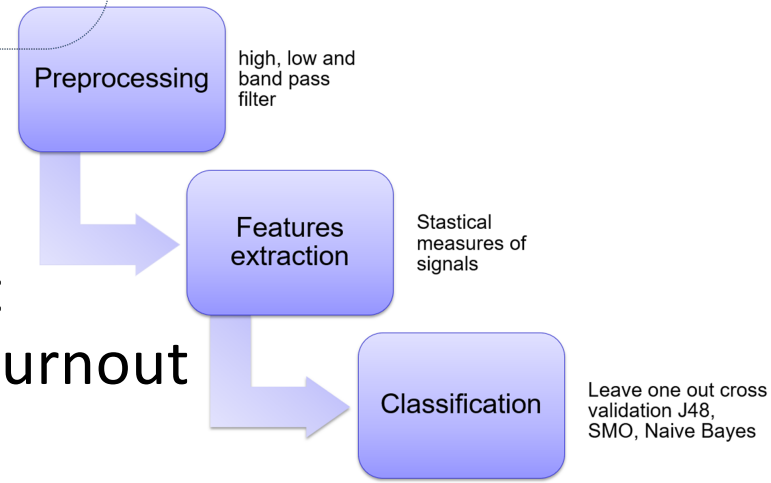


Shimmer GSR

Electromyography (EMG)

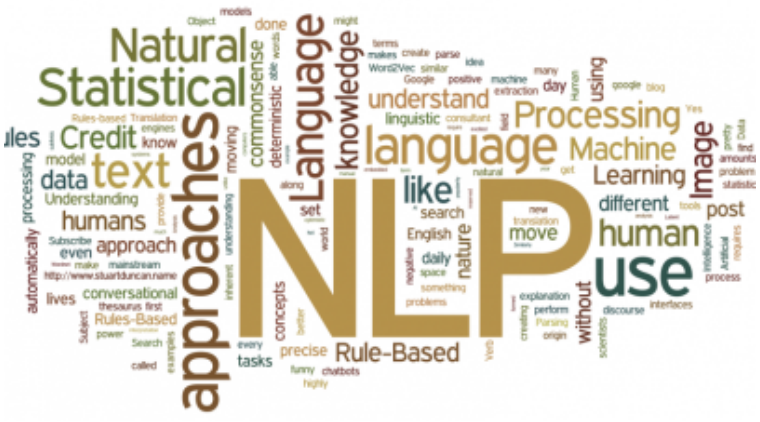
Shimmer EMG

- **Software engineering**: detect discomfort and unhealthy stress that can generate bugs and developer's burnout



Sentiment analysis

Analysis of biometrics



Information and Software Technology 94 (2018) 186–207

Contents lists available at ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof

How to ask for technical help? Evidence-based guidelines for writing questions on Stack Overflow

Fabio Calefato[a], Filippo Lanubile[b], Nicole Novielli[b,*]

[a] Dipartimento Jonico, University of Bari "A. Moro", via Duomo 259, 74123, Taranto, Italy
[b] Dipartimento di Infomatica, University of Bari "A. Moro", via E. Orabona 4, 70125, Bari, Italy

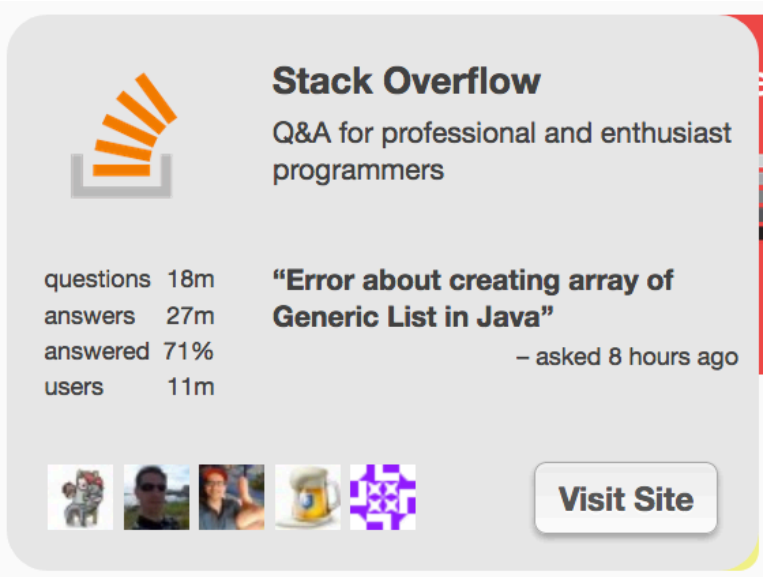
IEEE Software BLOG

Monday, November 20, 2017

Can We Trust the Stack Overflow Netiquette? Evidence-based Guidelines for Asking Good Technical Questions

by Fabio Calefato, University of Bari, ITALY (@fcalefato), Filippo Lanubile, University of Bari, ITALY (@lanubile), and Nicole Novielli, University of Bari, ITALY (@NicoleNovielli)
Associate Editor: Christoph Treude (@ctreude)

Community-based Q&A: Stack Overflow



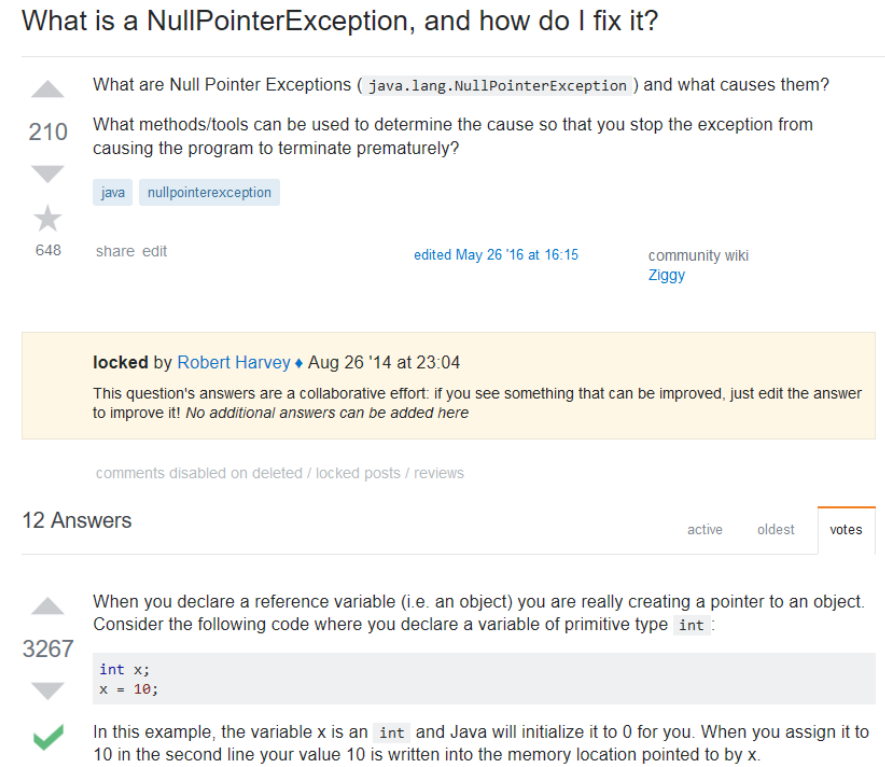
- Largest SE Q&A community
- SO Annual Developer Survey
  - ~50M visits per month
    - ~21M (25%) from professional developers and university-level students
    - > 80% rely on SO for learning
    - Developers read manuals less and less, they rather "search" (Mary Shaw's ICSE 2016 keynote)

Research question

- How can an information seeker increase the chances of eliciting answers?

Better off

- How can an information seeker increase the chances of eliciting the '**best**' **answer**?



Best answer is the one marked as **accepted** by the question asker

⇒ **question resolved**

⇒ **successful question**

Quantitative + qualitative analysis

87K questions

Logistic regression model

43 developers

Online survey

open-ended questions

one multiple choice question for each success factor

Netiquette for effective question writing on SO: **Evidence vs. User Perception**

#	Guideline	Success factor	Empirical support		User perception	Source
1	**Write questions using a neutral emotional style**	**Affect**	**Yes**	=	Effective	Skeet, SO Help Center, Kucuktunc et al., Bazelli et al.
2	**Provide sample code and data**	**Presentation quality**	**Yes**	=	Effective	Skeet, Asaduzzaman et al., Duijn et al., Treude et al.
3	**Use capital letters where appropriate**	**Presentation quality**	**Yes**	=	Effective	Skeet
4	**Be concise**	**Presentation quality**	**Yes**	≠	Ineffective	Skeet
5	Use short, descriptive question titles	Presentation quality	No	≠	Effective	Skeet
6	Provide context through tags	Presentation quality	No	≠	Effective	Skeet
7	Provide context through URLs	Presentation quality	No	≠	Effective	Ponzanelli et al.
8	**Be aware of low-efficiency hours**	**Time**	**Yes**	≠	Ineffective	Bosu et al.

= match
≠ mismatch

Sentiment Polarity Detection for Software Development

Fabio Calefato, Filippo Lanubile, Federico Maiorano & Nicole Novielli

Empirical Software Engineering
An International Journal

ISSN 1382-3256

Empir Software Eng
DOI 10.1007/s10664-017-9546-9

Need for domain-specific tools

Contextual semantics

'I am **missing** a parenthesis. But where?'

stackoverflow JIRA

Context of interaction

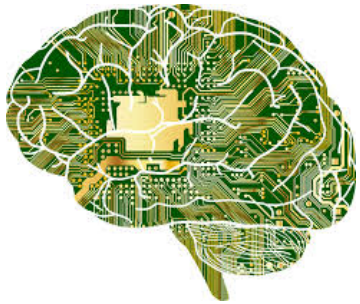
'I have a **problem**, […] please explain what is **wrong**'

Domain-dependent semantics

'What is the best way to **kill** a critical process?'

N. Novielli, F. Calefato, F. Lanubile. "The Challenges of Sentiment Detection in the Social Programmer Ecosystem" *(SSE'15)*

Developing Senti4SD



1. Taxonomy development

- Literature review on emotion models
- Emotion model and mapping to polarity
- Definition of the study design
- Annotation guidelines, experimental protocol and training material

2. Sampling and training

- Extracting post from Stack Overflow dump
- Raw questions, answers, and comments
- Remove URLs, HTML tags and code snippet
- Plain text of questions, answers, and comments
- Score with SentiStrength and sampling
- Annotation Sample (4,800 posts)
- Coder Training + Pilot annotation (400 posts)
- Annotated pilot
- Disagreements: analysis and reconciliation
- Enhanced guidelines

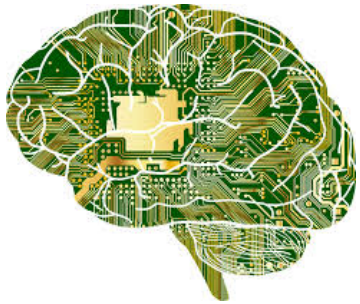
3. Annotation study

- Extended annotation study (4,200 posts)
- Full annotated sample
- Assessing reliability of the schema
- Kappa and observed inter-rater agreements
- Resolving disagreement (majority voting)
- Gold labels for 4,200 posts
- Pilot set with gold labels
- Merging the two sets
- Gold standard (4,423 posts)

4. Training and Evaluation

- Stratified sampling
- Train and test sets (70/30)
- Preprocessing: tokenization, replacement of user mentions
- Feature Extraction
- K-fold Cross validation
- Best model parameters
- Train Senti4SD on full train set
- Evaluate Senti4SD on test set
- Compare with baseline
- Analysis of results

Developing Senti4SD



1. Taxonomy development
- Literature review on emotion models
- Emotion model and mapping to polarity
- Definition of the study design
- Annotation guidelines, experimental protocol and training material

2. Sampling and training
- Extracting post from Stack Overflow dump
- Raw questions, answers, and comments
- Remove URLs, HTML tags and code snippet
- Plain text of questions, answers, and comments
- Score with SentiStrength and sampling
- Annotation Sample (4,800 posts)
- Coder Training + Pilot annotation (400 posts)
- Annotated pilot
- Disagreement [illegible] and reconciliation
- Enhanced guidelines

3. Annotation study
- Extended annotation study (4,200 posts)
- Full annotated sample
- Assessing reliability of the schema
- Kappa and observed inter-rater agreements
- Resolving disagreement (majority voting)
- Gold [illegible] 4,200 posts
- Pilot set with gold labels
- Merging the two
- Gold standard (4,423 posts)

4. Training and Evaluation
- Stratified sampling
- Train and test sets (70/30)
- Preprocessing: tokenization, replacement of user mentions
- Feature Extraction
- K-fold Cross validation
- Best model parameters
- Train Senti4SD on full train set
- Evaluate Senti4SD on test set
- Compare with baseline
- Analysis of results

caret

Liblinear

SentiStrength 😢

Performance on different datasets

2018 ACM/IEEE 15th International Conference on Mining Software Repositories

A Benchmark Study on Sentiment Analysis for Software Engineering Research

Nicole Novielli, Daniela Girardi, Filippo Lanubile
University of Bari Aldo Moro, Italy
{nicole.novielli, daniela.girardi, filippo.lanubile}@uniba.it

slides

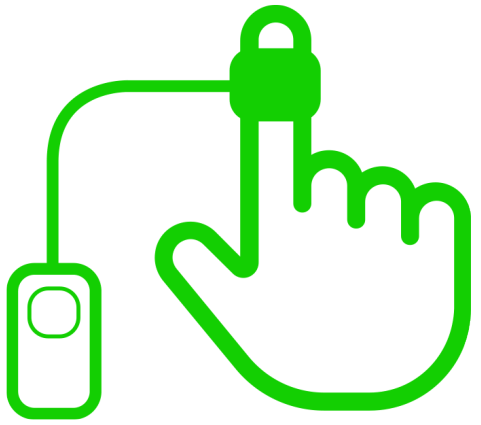
Model-driven annotation

Dataset	Class	Senti4SD		
		P	*R*	*F1*
Stack Overflow	Positive	.92	.92	.92
	Negative	.80	.89	.84
	Neutral	.87	.80	.83
	Micro-avg.	.87	.87	.87
Calefato et al., EMSE 2017	Macro-avg.	.86	.87	.86
Jira	Positive	.76	.79	.78
	Negative	.72	.57	.64
	Neutral	.86	.89	.88
	Micro-avg.	.83	.83	.83
Ortu et al., MSR 2016	Macro-avg.	.78	.75	.76

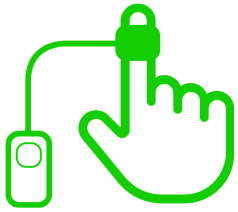
Ad-hoc annotation

Dataset	Class	Senti4SD		
		P	*R*	*F1*
Code Review	Negative	.68	.40	.51
	Non-Negative	.83	.94	.88
	Micro-avg.	.80	.80	.80
Ahmed et al., ASE'17	Macro-avg.	.75	.70	.69
Java Libraries	Positive	.63	.26	.36
	Negative	.55	.33	.41
	Neutral	.84	.96	.90
	Micro-avg.	.80	.80	.80
Lin et al., ICSE 2018	Macro- avg.	.67	.51	.56

Sensing Developers' Emotions



Instrumentation



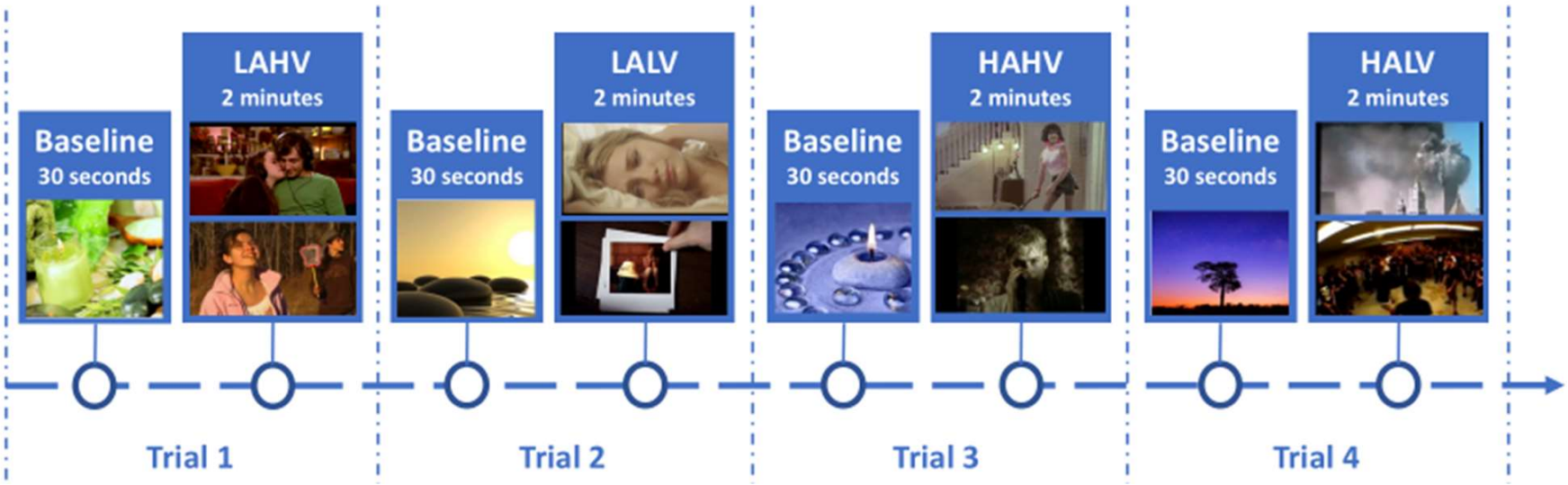
Preliminary study

Emotion Detection Using Noninvasive Low Cost Sensors

Daniela Girardi, Filippo Lanubile, Nicole Novielli
University of Bari, Italy

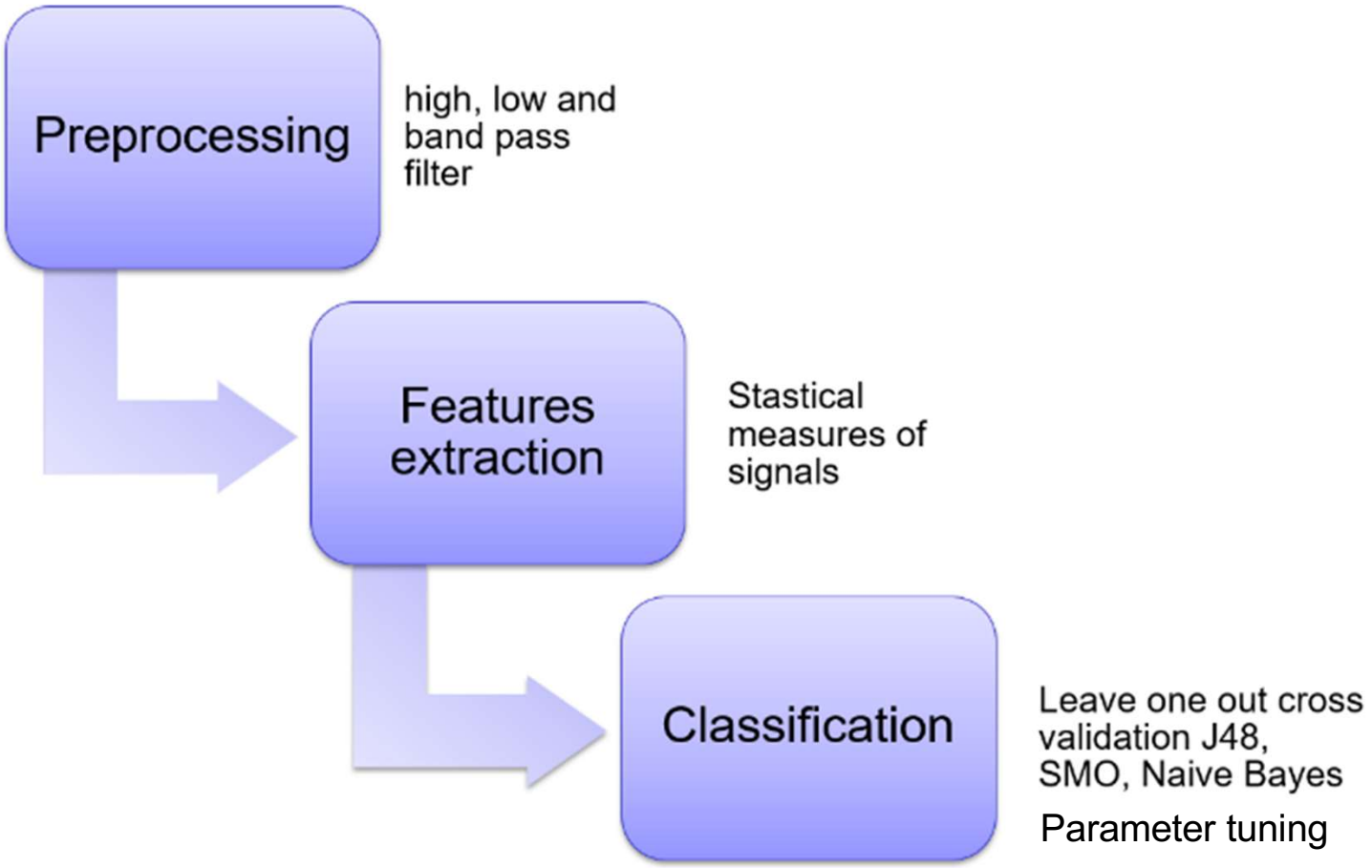
ACII 2017, Seventh International Conference on Affective Computing and Intelligent Interaction

Emotion elicitation timeline



- 8 videos from DEAP (Database for Emotion Analysis using Physiological signals)
- Each video is associated with different arousal and valence scores

Data Analysis



Results

Signals	Arousal					Valence				
	Classif	Prec	Recall	F1	Acc.	Classif	Prec	Recall	F1	Acc.
Single Sensors										
EEG	NB/KNN	0.76	0.76	0.76	0.76	J48	0.90	0.90	0.90	0.90
GSR	JRIP	0.69	0.68	0.68	0.68	KNN	0.67	0.67	0.67	0.67
EMG	KNN	0.80	0.80	0.80	0.80	KNN	0.70	0.70	0.70	0.70
Combined Sensors										
EEG+GSR	**J48**	**0.98**	**0.98**	**0.98**	**0.98**	**J48**	**0.97**	**0.97**	**0.97**	**0.97**
GSR+EMG	KNN	0.79	0.79	0.79	0.79	J48	0.76	0.76	0.76	0.76
EEG+EMG	J48	0.86	0.86	0.86	0.86	J48	0.95	0.95	0.95	0.95
All	J48	0.98	0.98	0.98	0.98	NB	0.82	0.82	0.82	0.82

Data collection in the field

Pilot study

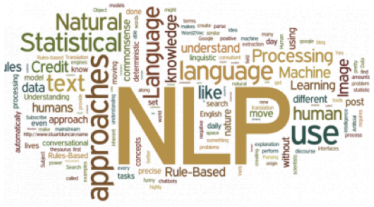
Duration: 2 weeks for each developer

Period: September – December 2019

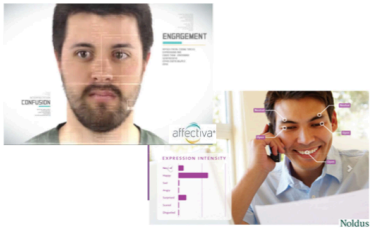
Number of participants: approx. 20

Four companies in Eindhoven

Sentiment analysis



Analysis of biometrics



Investigating the Role of Emotions in the Social Programmer Ecosystem

- Research question: **getting emotional** while communicating with developers: **good or bad?**
- Model: combining message proper social factors and affective factors
- Expected output:
  - SE-specific sentiment analysis tool and emotion classifier
  - Evidence-based netiquette

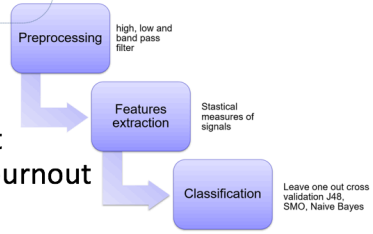
stackoverflow

github SOCIAL CODING

JIRA

Sensor-based emotion detection

Research question: Can we acquire physiological measures from noninvasive, low cost sensors to accurately predict emotions?



- **Software engineering**: detect discomfort and unhealthy stress that can generate bugs and developer's burnout