AN OVERVIEW OF USER FEEDBACK CLASSIFICATION APPROACHES

Rubens Santos, Eduard C. Groen, Karina Villela

ML Feature / ML Algorithm	BOW	BOF	TF-IDF	x^2	n-Gram	NLP-Heur.	AUR-BOW	
Naïve Bayes	18	1	8	1	2	6	1	37
Bayesian Network	1							1
Logistic Regression	9		2		2	6		19
k-Nearest Neighbors			3					3
Support Vector Machines	16*	1	6		1	6		30
DT – Single Tree / C4.5	10		6	1	2	6	1	26
DT – Boosted	4		1			4		9
DT – Random Forest	4				2	2		8
DT – Bagging	1		2	1			1	5
Neural Networks			1					1
No. of pairs	63	2	29	3	9	30	3	139

Fraunhofer IESE

What do we want?

A BENCHMARKING

of user feedback classification approaches for RE (CrowdRE)

When do we *have* it?

...not now...

Now, see…

the differences between the approaches we found

actually make it kind of difficult to make a proper comparison

that tells us reliably which approaches may be better suited for RE

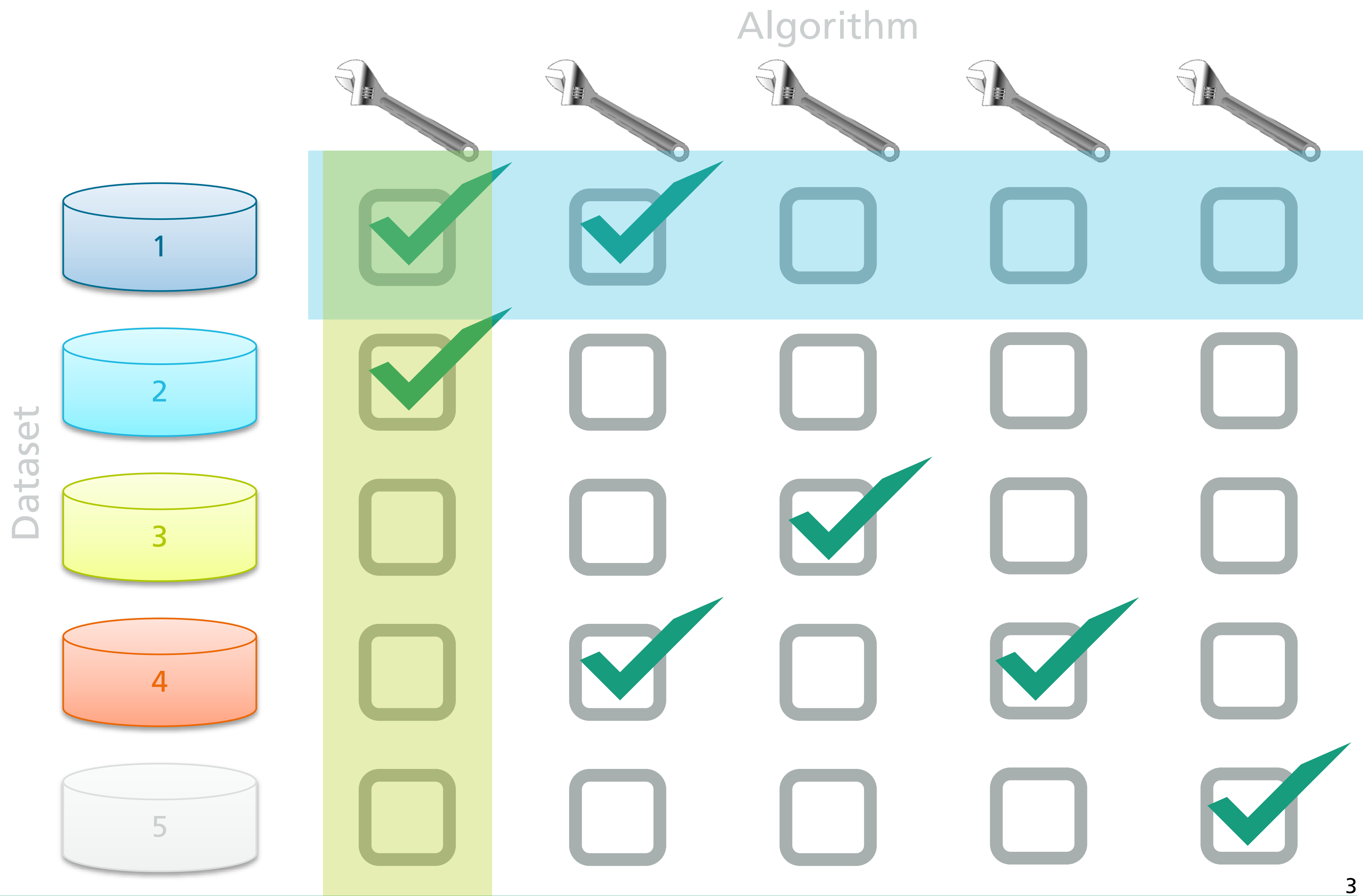
so that we are s[illegible] steps away from performing a benchmarking

which may require researchers to re-do analyses or to provide us with their data

in order for us to perform those analyses ourselves for their results to be comparable

on the various levels that these analyses currently differ to such great extents

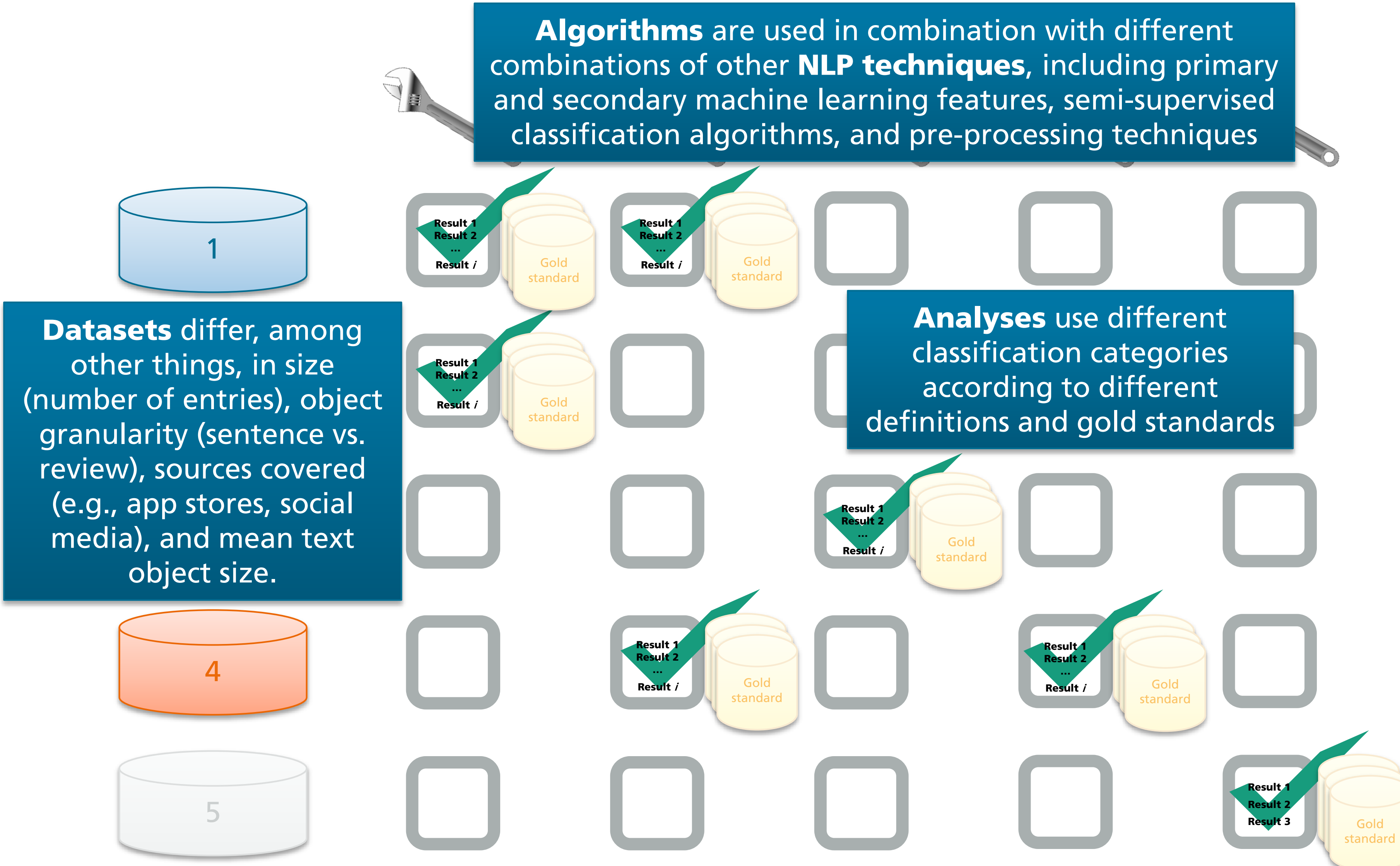
The Idea of Our Benchmarking is Simple…



Wrench image source: [MidnightLightning / WikiMedia Commons](), public domain

Fraunhofer
IESE

…The Reality of this Benchmarking is Difficult…



Algorithms are used in combination with different combinations of other **NLP techniques**, including primary and secondary machine learning features, semi-supervised classification algorithms, and pre-processing techniques

Datasets differ, among other things, in size (number of entries), object granularity (sentence vs. review), sources covered (e.g., app stores, social media), and mean text object size.

Analyses use different classification categories according to different definitions and gold standards

Wrench image source: [MidnightLightning / WikiMedia Commons](), public domain

Fraunhofer IESE

…But We *Are* Doing This Benchmarking

Algorithms are used in combination with different combinations of other **NLP techniques**, including primary and secondary machine learning features, semi-supervised classification algorithms, and pre-processing techniques

Hurdle 2: An overview of user feedback classification approaches

Focus of this presentation

Analyses use different classification categories according to different definitions and gold standards

Hurdle 1: A taxonomy for user feedback classifications

Systematic Literature Review

Benchmarking

Datasets differ, among other things, in size (number of entries), object granularity (sentence vs. review), sources covered (e.g., app stores, social media), and mean text object size.

Hurdle 3 and further:

- Comparing datasets
- Assessing the influence of NLP techniques
- Aligning analyses
- Etc.

Fraunhofer IESE

Systematic Literature Review

- Conducted according to Kitchenham, with an SLR protocol specifying:
  - objectives / research questions,
  - a search strategy with inclusion/exclusion criteria & a search string,
  - a data extraction strategy.

➢ **Note:** The SLR is not the main focus of this presentation!
  - We're showing a "byproduct" in a preliminary form
    - Focusing only on the *second* hurdle that we had to overcome
  - We wanted to get this material out there, so *you* can work with it!

SLR: Objectives

Overall Objective: What are the state-of-the-art automated approaches for assisting the task of requirements extraction from user feedback acquired from the crowd, and which NLP techniques and features do they use?

- **Objective 1:** Regarding requirements elicitation from user feedback acquired from the crowd, what are the state-of-art automated approaches for classifying user feedback?
- **Objective 2:** How do such approaches classify user feedback?
  - **Objective 2.1:** What are the different sets of categories in which user feedback is classified?
  - **Objective 2.2:** Which automated techniques are used?
  - **Objective 2.3:** What are the characteristics of the user feedback these approaches aim to classify?

SLR: Paper Search

Performed March 2018 (+ December 2018)

1,219 papers + 14 papers
- Found by search string

146 papers
- After filtering by title & abstract

40 papers + 3 papers
- After filtering by intro & conclusion

EC1: not English
EC2: before 2013
EC3: not peer-reviewed

IC1: filters out irrelevant user feedback
IC2: classifies into predetermined categories
EC4: not RE / unrelated title
EC5: not on req. extraction from user feedback
EC6: tool not (usable) for requirement extraction
EC7: tool does not process textual user feedback
EC8: manual processing without automation

Fraunhofer IESE

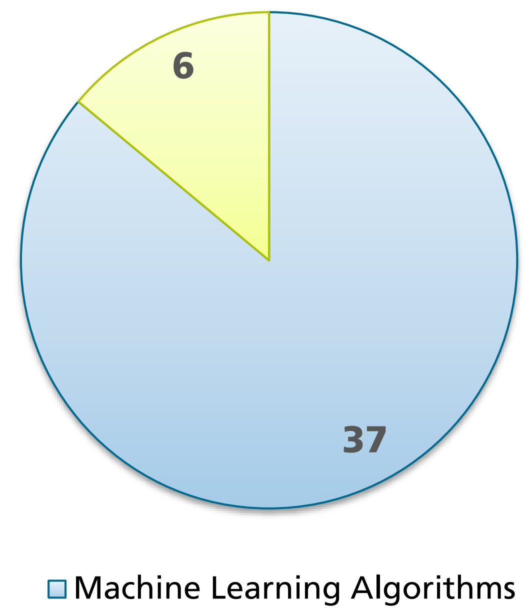
SLR: Data Extraction from 43 Papers

1. **Dataset-related information**
   - *E.g., dataset size in number of entries, object granularity, sources, mean text object size*
2. **NLP techniques applied** → Classification approach comparison
   - *E.g., algorithms, parsers, ML features, text pre-processing techniques*
3. **User feedback classification categories**
   - *E.g., name, definition, rationale/goal*

Research Focuses on Machine Learning Algorithms

- The SLR found **43 papers** on user feedback classification in RE (CrowdRE)
- Analysis of NLP techniques:
  - **86% used ML algorithms**
  - Mostly several (1 to 14; 3.8 average) → comparative experiments

Type of Classification

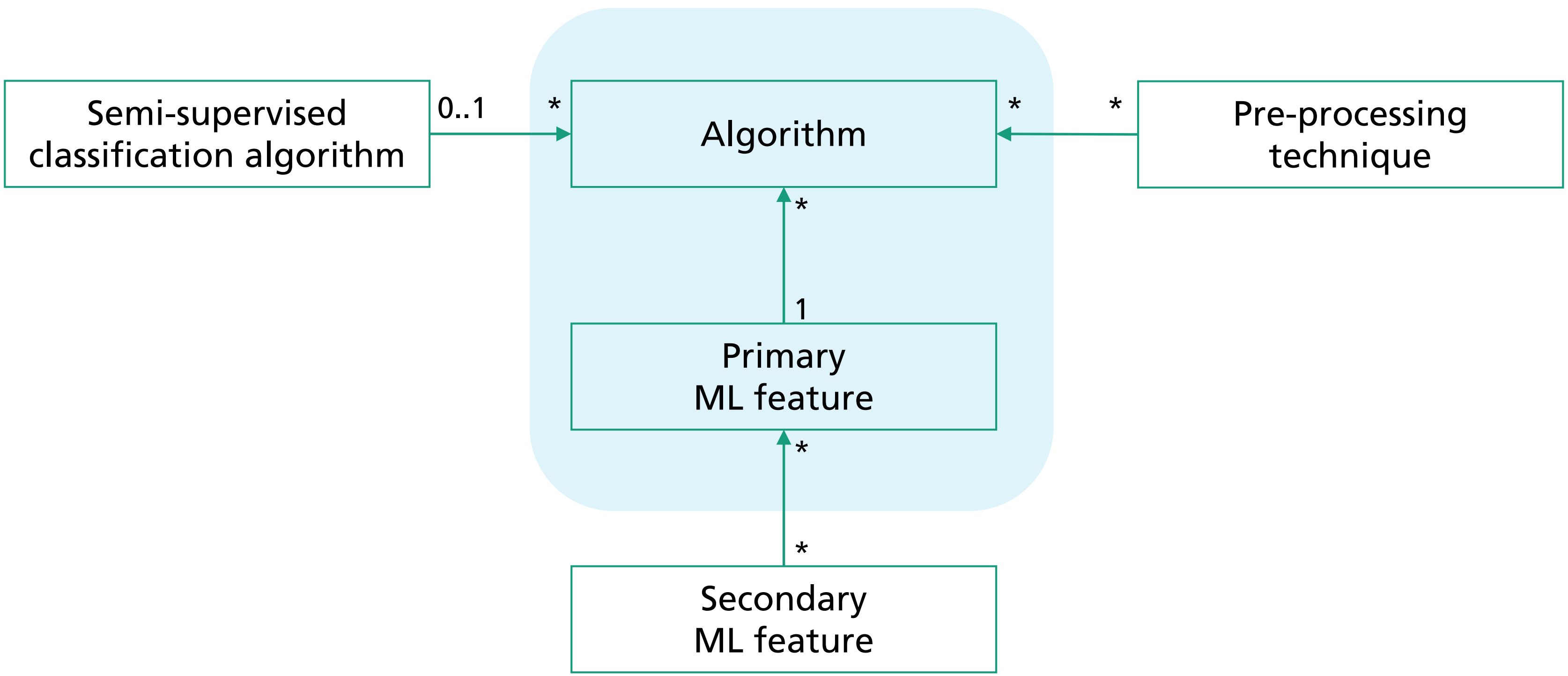


Fraunhofer IESE

Systematic Mapping of Machine Learning Techniques 1/2

- **Primary ML features** that represent the text according to word count or related methods
  - *E.g., Bag of Words, Term Frequency – Inverse Document Frequency, Bag of Frames*
- **Secondary ML features** that represent specific aspects of the text or metadata. They yield low scores when used on their own, but can help achieve greater efficiency & quality in combination with primary features.
  - *E.g., length, sentiment score, star rating*
- **Semi-supervised classification algorithms**
  - *E.g., Expectation-Maximization, Self-Training, Rasco*
- **Pre-processing techniques**
  - *E.g., stop words removal, synonym unification, stemming, lemmatization, special characters removal, abbreviation transformation, negation handling*

Systematic Mapping of Machine Learning Techniques 2/2



Frequency of ML Algorithm + ML Technique Pair

ML Algorithm / ML Feature	BOW	BOF	TF-IDF	X^2	n-Gram	NLP-Heur.	AUR-BOW	
Naïve Bayes	18	1	8	1	2	6	1	**37**
Bayesian Network	1							**1**
Logistic Regression	9		2		2	6		**19**
k-Nearest Neighbors			3					**3**
Support Vector Machines	16*	1	6		1	6		**30**
DT – Single Tree / C4.5	10		6	1	2	6	1	**26**
DT – Boosted	4		1			4		**9**
DT – Random Forest	4				2	2		**8**
DT – Bagging	1		2	1			1	**5**
Neural Networks			1					**1**
No. of pairs	**63**	**2**	**29**	**3**	**9**	**30**	**3**	**139**

* One measurement uses BOF – Boolean; all others use BOF-TF.

Frequency of ML Algorithm + ML Technique Pair

ML Algorithm \ ML Feature	BOW	BOF	TF-IDF	X^2	n-Gram	NLP-Heur.	AUR-BOW	
Naïve Bayes	18	1	8	1	2	6	1	27
Bayesian Network	1							1
Logistic Regression	9		2		2	6		12
k-Nearest Neighbors			3					3
Support Vector Machines	16*	1	6		1	6		22
DT – Single Tree / C4.5	10		6	1	2	6	1	
DT – Boosted	4		1			4		17
DT – Random Forest	4				2	2		
DT – Bagging	1		2	1			1	
Neural Networks			1					1
No. of papers	**23**	**1**	**12**	**1**	**5**	**8**	**1**	

* One measurement uses BOF – Boolean; all others use BOF-TF.

Fraunhofer IESE

Number of “Feature Request” Measurements

ML Algorithm / ML Feature	BOW	BOF	TF-IDF	X^2	n-Gram	NLP-Heur.	AUR-BOW	
Naïve Bayes	9 *	1	5	1	1 *	2	1	**20**
~~Bayesian Network~~								**0**
Logistic Regression	2 *		1		2 **	1		**6**
k-Nearest Neighbors			1					**1**
Support Vector Machines	7 *	1	3		1 *	2		**14**
DT – Single Tree / C4.5	2		4	1	1	1	1	**10**
DT – Boosted	1					1		**2**
DT – Random Forest	1				1			**2**
DT – Bagging	1		2	1			1	**5**
Neural Networks			1					**1**
Measurements	**23**	**2**	**17**	**3**	**6**	**7**	**3**	**61**

* Includes one F_1 measure for which no precision/recall values could be obtained

** Includes two F_1 measures for which no precision/recall values could be obtained

F_β Measures for "Feature Request"

ML Algorithm \ ML Feature	BOW	BOF	TF-IDF	X^2	n-Gram	NLP-Heur.	AUR-BOW	
Naïve Bayes								0.65
Logistic Regression								0.40
k-Nearest Neighbors								0.53
Support Vector Machines								0.68
DT – Single Tree / C4.5								0.64
DT – Boosted								0.66
DT – Random Forest								0.84
DT – Bagging								0.66
Neural Networks								0.39
Averages	0.68	0.81	0.55	0.68	0.68	0.70	0.63	0.64

task-based β_T, calculated as $\frac{1}{\lambda}$

$$F_\beta = (1+\beta^2) \times \frac{P \times R}{(\beta^2 \times P) + R}$$

Fraunhofer IESE

The Most Popular…

- **User feedback analysis approach:** Machine Learning
  - Only few use dictionaries, regular expressions or parsing
- **ML algorithms:** NB, SVM, LR, and DT (esp. Single Tree)
  - Probably because they provide a relatively large degree of control over the supervised ML
  - 12 clusters found in total
- **Primary ML features:** BOW and TF-IDF
  - Probably because of their versatile nature
  - 7 clusters found in total
- **ML models:** NB + BOW, SVM + BOW, NB + TF-IDF
  - Probably because of their tool support and familiarity

Fraunhofer IESE

User Feedback Analysis Research for RE: Still Got a Long Long Way to Go

- F_β Measures for “Feature Requests” were surprisingly **moderate**
  - Especially assuming publication bias (best possible outcomes)
  - Four ML models had $F_\beta > 0.85$, but for just one measurement
- All popular ML algorithms can potentially result in **good-quality results**
  - Study characteristics we did not investigate seem to have a strong impact on classification efficiency
- Strong **variance** in ML models used & study set-ups
  - Research is still exploring appropriate ML models

Fraunhofer IESE

By Not Taking Inspiration from Other Works, CrowdRE Research is Missing Out on Opportunities!

- **NLP Heuristics and *n*-Grams** have been shown to contribute to better results by introducing context information into the classification task
- No research has picked up on adaptations of BOW in CrowdRE research that yielded good results: **Bag of Frames** (P20) and **Augmented User Reviews – BOW** (P24)
- Other works may obtain better results for **Bayesian Network** as in P43, or **Neural Networks** (or another Deep Learning approach) than in P14
- Works investigating **non-ML approaches** for RE suggest that carefully designed heuristics may in some cases also provide accurate results (i.e., high precision), but not necessarily contribute to higher recall
- Researchers could help others if they provide a **rationale** for their choice of techniques, which we hardly saw in research

Fraunhofer IESE

Implications and Outlook

- Our findings can help you (and us) make a more **informed choice** of appropriate ML algorithms and ML features to achieve better user feedback classification for RE
- **No decisive conclusions** about the most suitable ML models
  - Factors other than the ones we considered in this work appear to have had a strong influence on the performance of the ML models
  - We did find that good results have been attained with the most often used ML algorithms, especially when used in combination with appropriate primary and secondary ML features
- The current landscape is still one of **exploration** into the most suitable techniques, but progress is hindered by a **lack of cross-fertilization**
  - Research does not pick up on promising findings in other works to investigate whether these approaches work well in their context

Fraunhofer IESE

Future Work

- For our benchmarking study, these findings further fuel the **need for an evaluation** of user feedback analysis techniques for different purposes
  - On the other hand, the potential of non-ML approaches reported in some works should not be ignored either
- This work was **descriptive** in nature and was limited to a comparison of only the ML algorithms and primary ML features.
  - More prescriptive results could be obtained through an assessment of which study-specific aspects impact performance most strongly
    - *E.g., addressed goals & problems; dataset type/quality/size; classification categories chosen; gold standard composed; additional ML techniques used (semi-supervised classification algorithms / pre-processing techniques / secondary ML features)*
  - Due to the study's set-up, we did not investigate the performance of ML models in other contexts within and outside of RE

Fraunhofer IESE

Thank you!

ML Algorithm / ML Feature	BOW	BOF	TF-IDF	X^2	n-Gram	NLP-Heur.	AUR-BOW	
Naïve Bayes	18	1	8	1	2	6	1	**27**
Bayesian Network	1							**1**
Logistic Regression	9		2		2	6		**12**
k-Nearest Neighbors			3					**3**
Support Vector Machines	16*	1	6		1	6		**22**
DT – Single Tree / C4.5	10		6	1	2	6	1	
DT – Boosted	4		1			4		**17**
DT – Random Forest	4				2	2		
DT – Bagging	1		2	1			1	
Neural Networks			1					**1**
No. of papers	**23**	**1**	**12**	**1**	**5**	**8**	**1**	

Thank you!

ML Algorithm / ML Feature	BOW	BOF	TF-IDF	X^2	n-Gram	NLP-Heur.	AUR-BOW	
Naïve Bayes								0.65
Logistic Regression								0.40
k-Nearest Neighbors								0.53
Support Vector Machines								0.68
DT – Single Tree / C4.5								0.64
DT – Boosted								0.66
DT – Random Forest								0.84
DT – Bagging								0.66
Neural Networks								0.39
Averages	0.68	0.81	0.55	0.68	0.68	0.70	0.63	0.64

Fraunhofer IESE