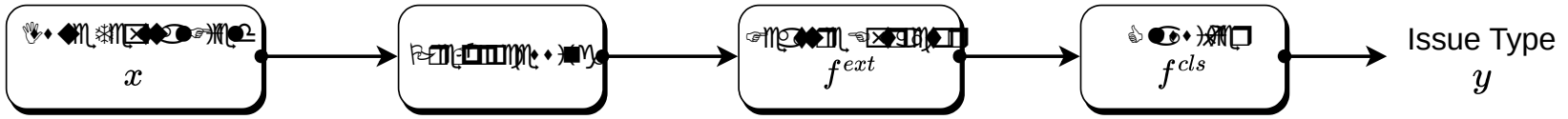
Empirically Revisiting and Enhancing Automatic Classification of Bug and Non-Bug Issues

Zhong LI, Minxue PAN, Yu PEI, Tian ZHANG, Linzhang WANG, Xuandong LI

Frontiers of Computer Science, DOI: 10.1007/s11704-023-2771-z

Motivation & Design

- There is a lack of comprehensive, in-depth work analyzing the effectiveness of different design choice in issue classification:
  - How do different textual fields of issues affect the performance of issue classification?
  - How do different feature representation methods affect the performance of issue classification?
  - How do different machine learning algorithms affect the performance of issue classification?
- **This work:** The first extensive study of automated issue classification on 9 state-of-the-art issue classification approaches, which reveal multiple practical guidelines for further advancing issue classification.



The workflow of automatic issue classification. The differences among existing issue classification approaches mainly pertain to a) the used textual field; b) the feature representation methods used to build the feature extractor; c) the machine learning algorithms adopted to learn the classifier. This work systematically explore the impacts of these design choices on the performance of automatic issue classification approaches.

Main Contributions

- Contributions:
  - The first extensive study of automated issue classification on 9 state-of-the-art issue classification approaches;
  - Our study reveals multiple practical guidelines for further advancing issue classification.
  - An advanced issue classification approach, DeepLabel, which can achieve better performance compared with the existing issue classification approaches.

1. Training separate models for the issue titles and descriptions and then combining these two models tend to achieve better performance for issue classification;
2. Word embedding with LSTM can better extract features from the textual fields in the issues, thus leading to better issue classification models;
3. There exist certain terms in the textual field that are helpful for building more discriminative classifiers between bug and non-bug issues;
4. The performance of the issue classification model is not sensitive to the choices of ML algorithms.

Guidelines derived from our study.

Approach	Performance Metrics									Statistical Testing								
	Bug			Non-Bug			Overall			Bug			Non-Bug			Overall		
	precision	*recall*	F1	*precision*	*recall*	F1	*precision*	*recall*	F1	*precision*	*recall*	F1	*precision*	*recall*	F1	*precision*	*recall*	F1
Chawla	0.73	0.49	0.59	0.77	0.90	0.83	0.76	0.76	0.75	0.86	0.86	1.00	0.97	0.21	0.84	0.94	0.93	0.97
Pandey	0.78	0.70	0.74	0.85	0.89	0.87	0.83	0.83	0.83	0.49	0.54	0.67	0.58	0.30	0.40	0.69	0.65	0.67
Otoom	0.24	0.03	0.05	0.64	0.94	0.76	0.51	0.63	0.52	1.00	1.00	1.00	-0.64	1.00	0.94	1.00	1.00	1.00
Terdchanakul	0.58	0.33	0.42	0.71	0.87	0.78	0.67	0.68	0.66	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Pingclasai	0.41	0.41	0.41	0.69	0.69	0.69	0.60	0.60	0.60	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Qin	0.78	0.75	0.76	0.87	0.87	0.87	0.84	0.84	0.84	0.65	0.09	0.50	0.18	0.35	0.28	0.44	0.43	0.42
Kallis	0.78	0.71	0.74	0.85	0.88	0.87	0.83	0.83	0.83	0.48	0.44	0.68	0.50	0.26	0.40	0.63	0.58	0.60
Herbold	0.82	0.72	0.77	0.86	0.91	0.89	0.85	0.85	0.85	0.00	0.47	0.38	0.38	-0.05	0.16	0.30	0.29	0.30
BERT-*Title*	0.77	0.78	0.77	0.88	0.87	0.87	0.84	0.84	0.84	0.62	-0.20	0.28	0.02	0.48	0.30	0.32	0.38	0.31
BERT-*Desc*	0.72	0.70	0.71	0.84	0.85	0.84	0.80	0.80	0.80	0.88	0.56	0.90	0.66	0.66	0.70	0.82	0.82	0.82
BERT-*Comb*1	0.78	0.79	0.78	0.89	0.88	0.88	0.85	0.85	0.85	0.42	-0.34	0.16	-0.10	0.42	0.16	0.16	0.19	0.15
DeepLabel	0.82	0.77	0.79	0.88	0.91	0.89	0.86	0.86	0.86					-				

DeepLabel, which is designed based on our derived guidelines, achieve significantly better performance compared to all existing issue classification approaches as well as the state-of-the-art model BERT.