Combining Retrieval and Classification:

Balancing Efficiency and Accuracy in Duplicate Bug Report Detection

Authors: Qianru Meng[1], Xiao Zhang[2], Guus Ramackers[3], Visser Joost[4]

Affiliation: Leiden Institute of Advanced Computer Science (**LIACS**)[134]
Center for Language and Cognition Groningen (**CLCG**)[2]

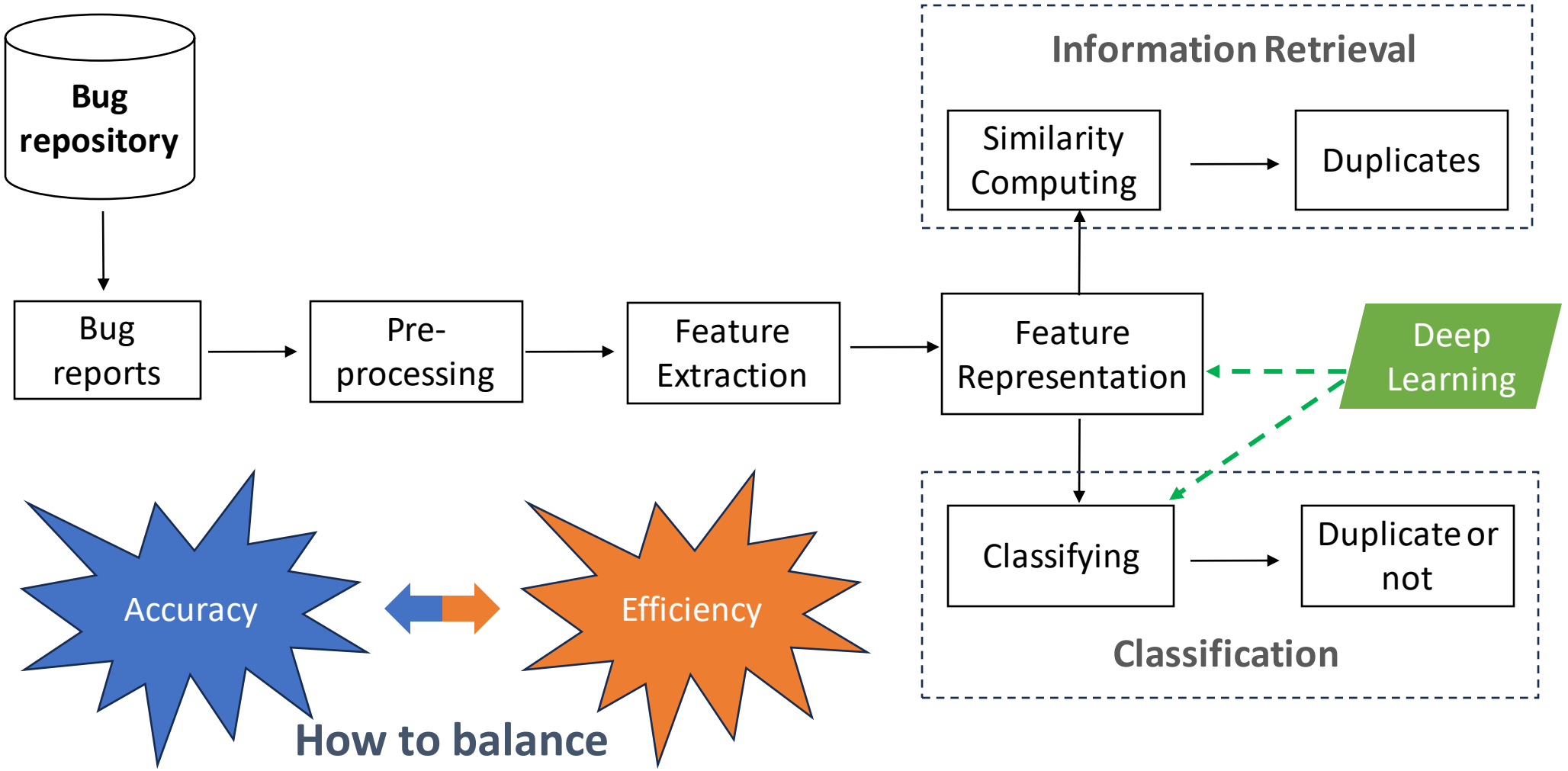
Presenter: Qianru Meng (q.r.meng@liacs.leidenuniv.nl)

Universiteit Leiden

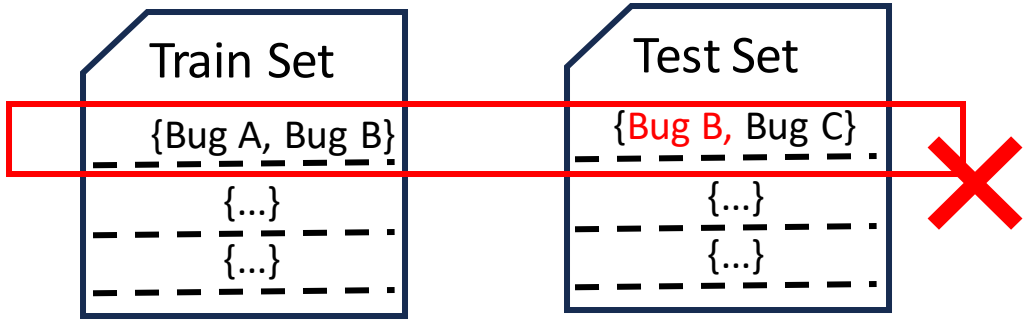
Qianru Meng

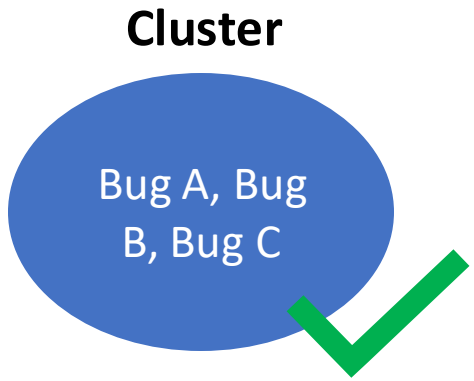
- **Master of Science**
- University of Bristol, United Kingdom, 2017 – 2018
- **Test Engineer**
- Baidu Inc. & Byte Dance, China, 2019 – 2021
- **PhD Candidate**
- The Computer Science institute of Leiden University (LIACS), Netherlands, Since 2021
- **Research interests: Software Engineering, Data Mining, Information Retrieval**

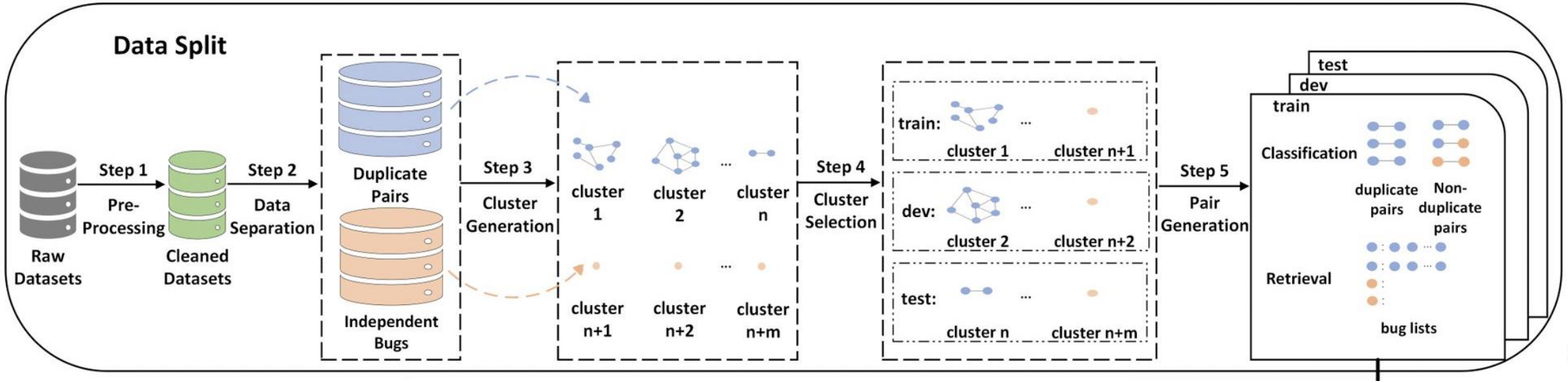
Duplicate Bug Report Detection



Data Leakage





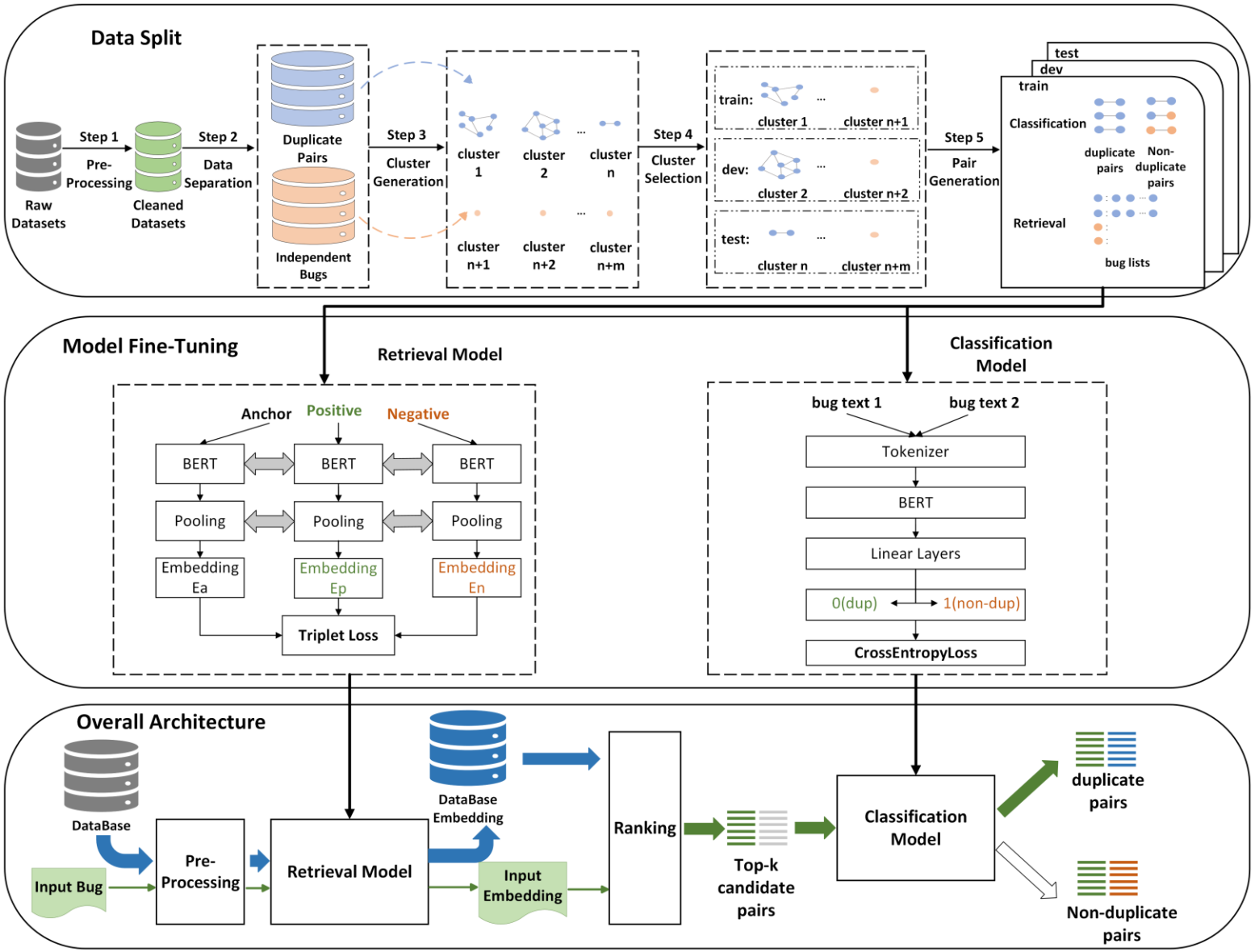


Dataset

TABLE II: STATISTICS OF FIVE OPEN SOURCE DATASETS

Dataset	Bugs	Dup Pairs	Separate Bugs	Dup Bug Ratio	Cluster Numbers	Cluster size
Eclipse	84020(85156)	13231	70752	0.1564	7519	2.760
Firefox	96258(115814)	15742	80000	0.1689	6654	3.366
Mozilla	195248(205069)	34507	160378	0.1786	17263	2.998
JDT	44154(45296)	6513	37608	0.1483	3828	2.701
TBird	24767(32551)	4404	20050	0.1905	2133	3.065

Proposed Architecture



Model Evaluation and Results

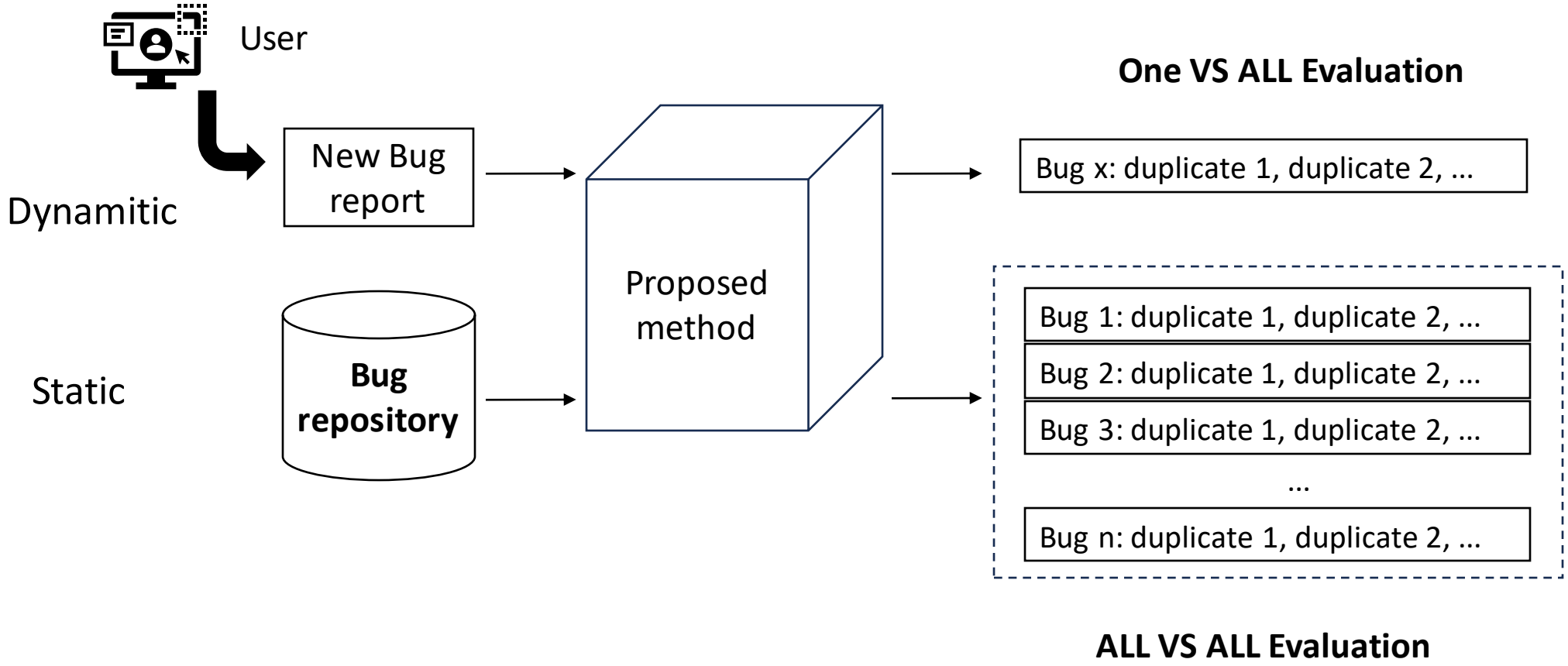
TABLE IV: RECALL@K OF MODELS IN DUPLICATE BUG RETRIEVAL FOR ALL DATASETS

	Eclipse			Firefox			Mozilla			JDT			TBird			
	r@20	r@60	r@100	r@20	r@60	r@100	r@20	r@60	r@100	r@20	r@60	r@100	r@20	r@60	r@100	Avg r@100
Fasttext	0.489	0.678	0.783	0.596	0.716	0.809	0.414	0.526	0.588	0.608	0.785	0.972	0.627	0.874	**1.000**	0.8304
Glove	0.602	0.727	0.824	0.705	0.789	0.843	0.478	0.608	0.662	0.579	0.798	0.975	0.689	0.888	**1.000**	0.8608
SBERT	**0.848**	**0.935**	**0.960**	**0.892**	**0.956**	**0.973**	**0.771**	**0.892**	**0.919**	**0.872**	**0.990**	**0.997**	**0.880**	**0.983**	1.000	**0.9698**

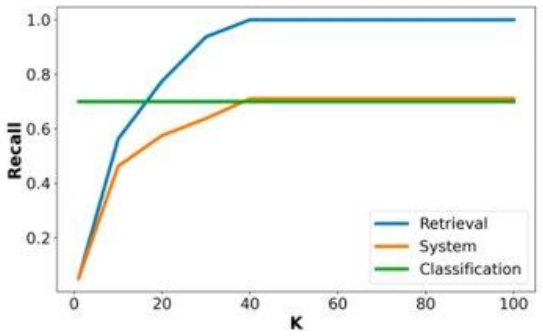
TABLE V: PRECISION, RECALL & F1 SCORES OF MODELS IN DUPLICATE BUG CLASSIFICATION TASK FOR ALL DATASETS

	Eclipse			Firefox			Mozilla			JDT			TBird			
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Avg F1
Bi-LSTM	0.511	0.506	0.473	0.510	0.515	0.469	0.507	0.506	0.506	0.490	0.490	0.490	0.621	0.510	0.474	0.4824
DC-CNN	0.752	0.813	0.785	0.744	0.765	0.753	0.792	0.765	0.736	0.763	0.781	0.773	0.833	0.752	0.781	0.7660
BERT	**0.825**	0.888	0.848	0.881	0.921	0.899	0.824	**0.892**	0.849	0.772	0.857	0.797	0.870	0.898	0.883	0.8552
ALBERT	0.806	**0.896**	0.834	0.874	0.920	0.893	0.819	0.889	0.845	**0.825**	**0.872**	**0.843**	**0.885**	**0.902**	**0.893**	0.8616
RoBERTa	0.846	0.892	0.866	**0.886**	**0.925**	**0.903**	**0.835**	0.891	**0.857**	0.824	0.868	0.841	0.846	0.898	0.866	**0.8666**

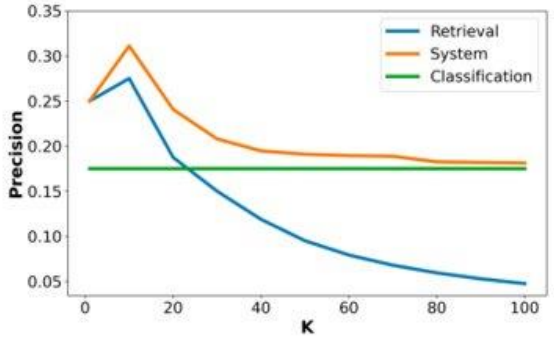
Architecture Evaluation



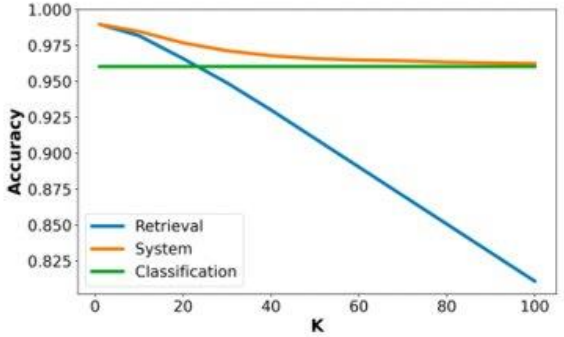
Architecture Evaluation Results



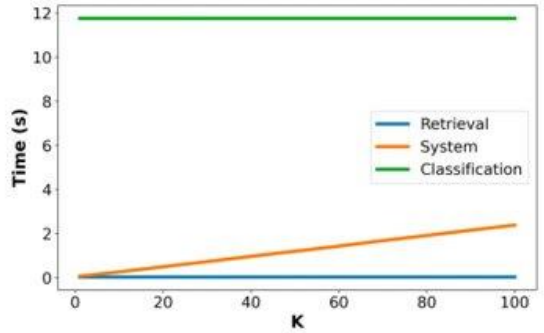
(a) Recall



(b) Precision

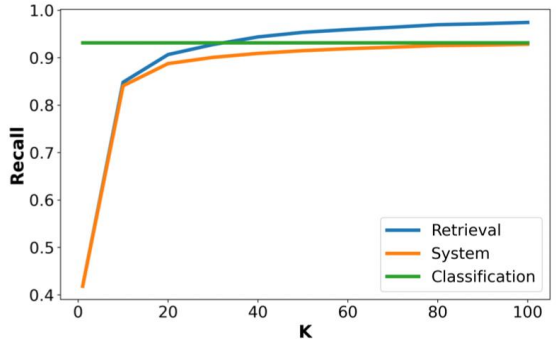


(c) Accuracy

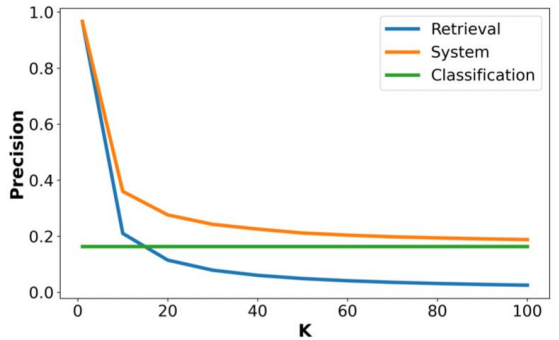


(d) Running Time

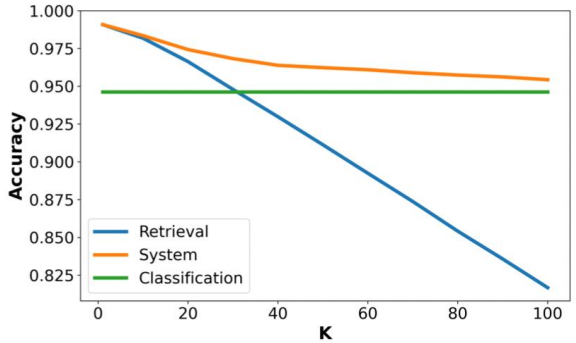
Fig. 2: Time-Performance Evaluation on Firefox Dataset in One VS All scenario



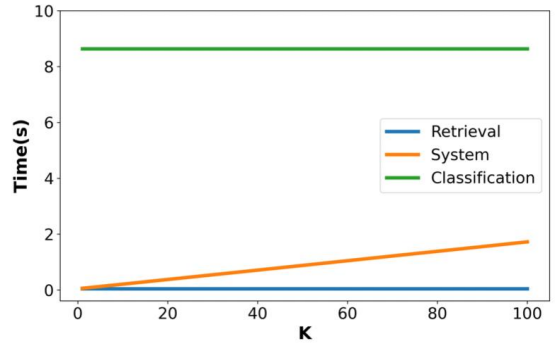
(a) Recall



(b) Precision



(c) Accuracy



(d) Running Time

Fig. 3: Time-Performance Evaluation on Firefox Dataset in All VS All scenario

The proper K value affect the balance between accuracy and efficiency

Conclusion

- **Evaluating the effectiveness of transformer-based models** in both classification and retrieval tasks
- **Demonstrating the balance of the proposed method** between accuracy and efficiency, combining the strength of retrieval and classification.
- **Providing insight into duplicate detection** from a **resource utilization perspective** enables more efficient and accurate responses in dynamic environments.

Future work

- **One model to rule them all**
- **More datasets**
- **Real-time detection**

Thanks!

Universiteit Leiden