Phase 2 Instructions

Phase 2

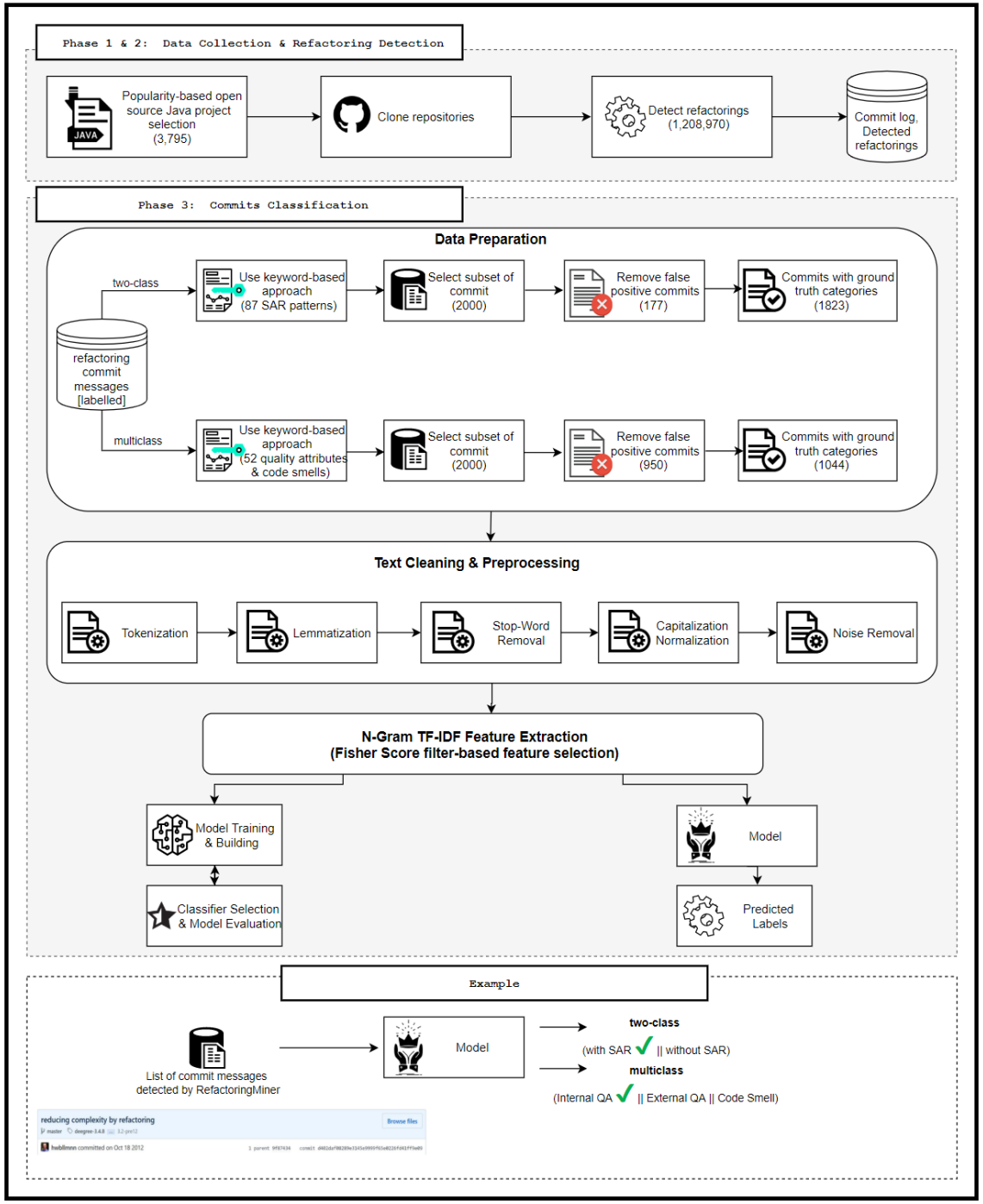
- **Phase 2** is a continuation of **Phase 1**.
- You use Phase 1 document and you extend it to create Phase 2 by adding the following sections:
  - Study Design
  - Experiments

Study Design (1 – 2 pages)

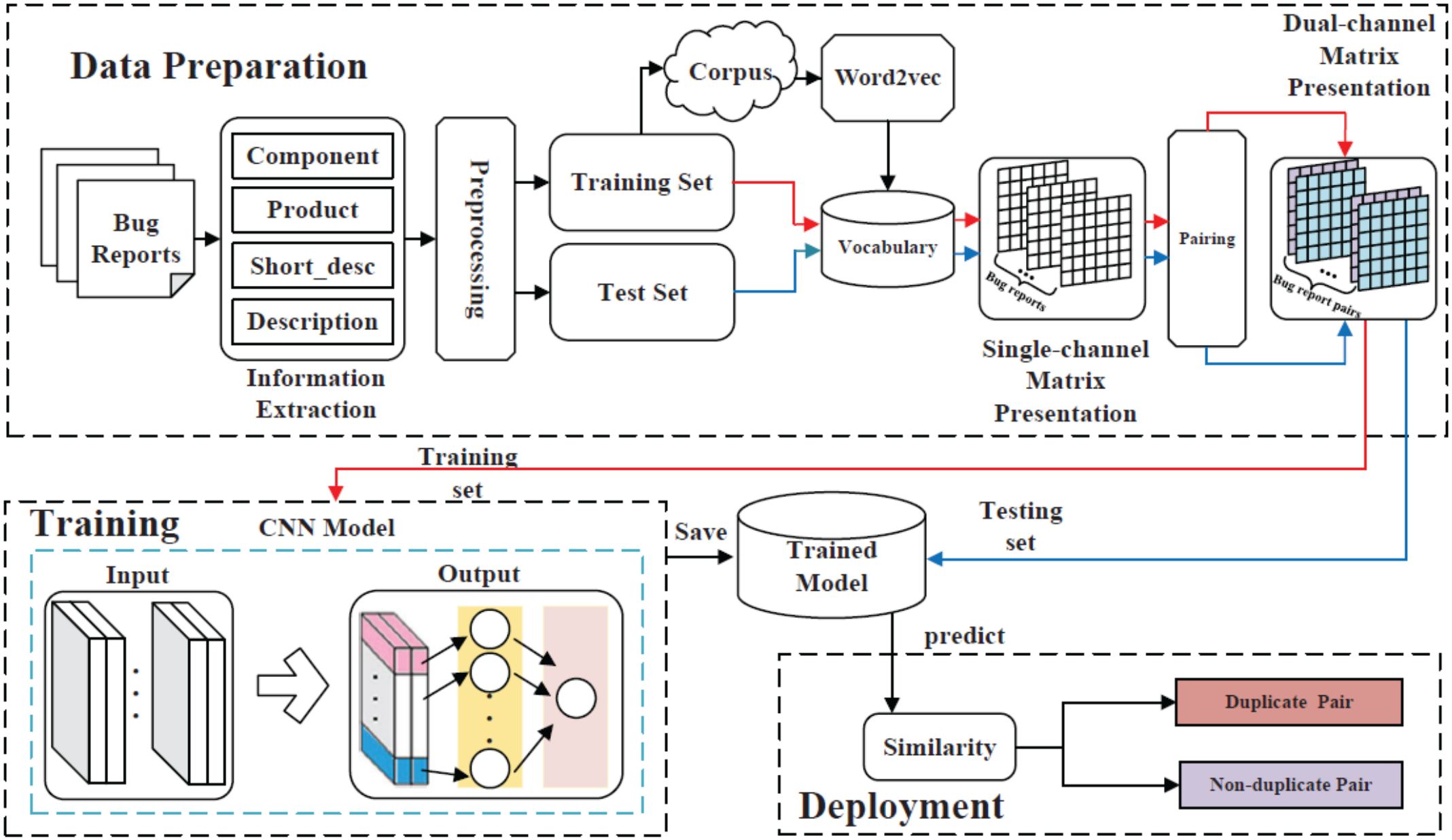
- This section explains the necessary steps you have taken to design the solution of your problem. You start from data preprocessing, (e.g., cleaning, normalization, stemming, vectorization, etc.), then the design of the model that will input the data all the way to the output.
- This section should be brief, for this submission, since I want you to focus more on the next section (experiments).
- Study Design MUST have an **approach overview** figure that explains everything.

RIT

Approach overview examples



Approach overview examples



Approach overview example

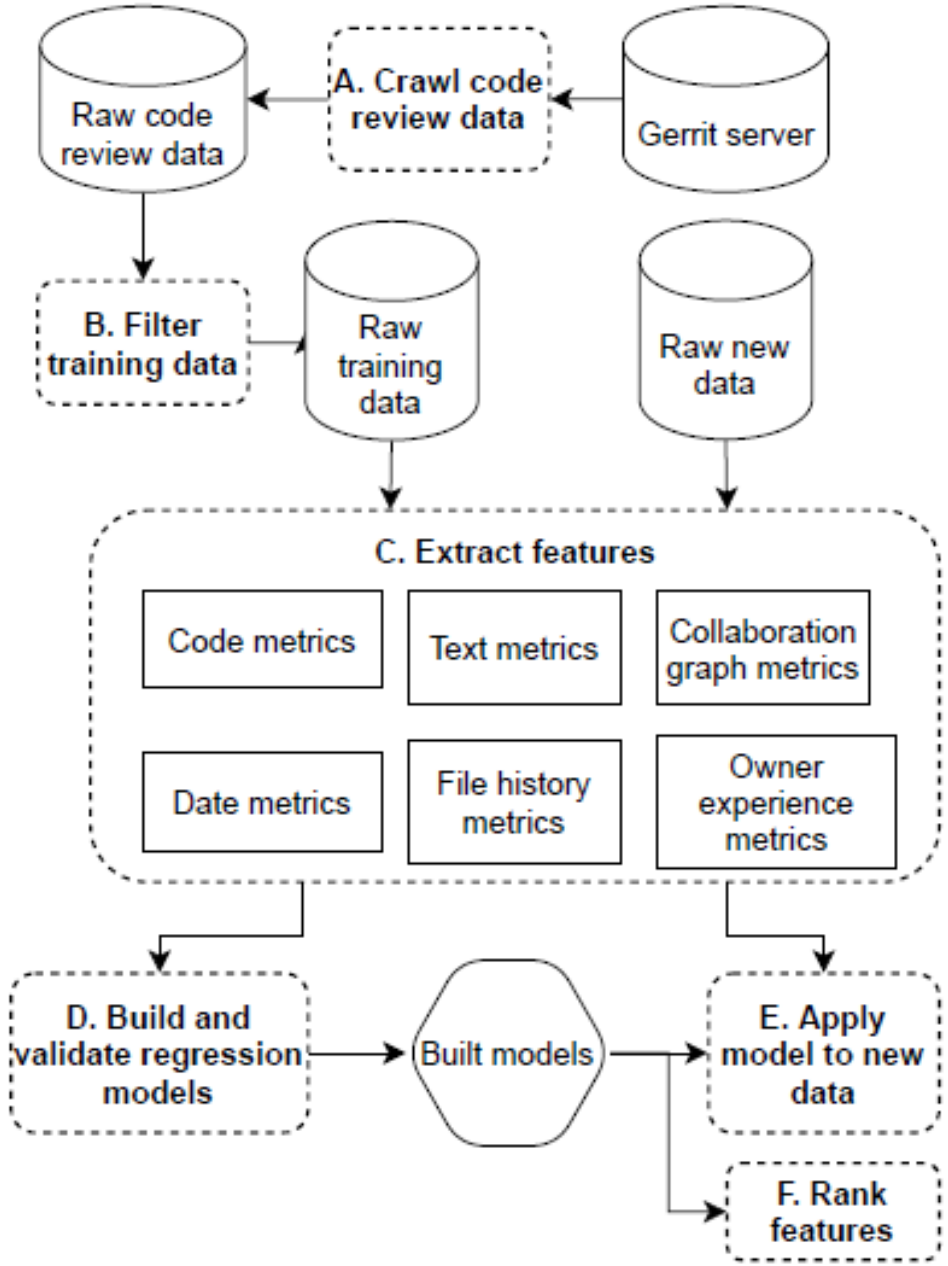


Fig. 3 Framework overview.

Experiments (2 – 6 pages)

- This section contains the preliminary results of your solution given the dataset (problem).
- This section should be organized using research questions.
- **Definition of Research Questions:** It represents the plan of how you will be assessing your solution for the problem.
- You need to carefully choose your research questions to show the performance of your solution and any potential advantages and/or limitations.
- There are plenty of good examples of Research Questions in the papers you are given to present / review. Please review them.

Research Questions Examples (From SARDELE paper)

- **RQ1:** How do different classifiers, based on comments and source code, perform for recommending SATD removals?

  The combined approach (using comment + source code)significantly outperforms the individual classifiers (Precision ’ 55%, Recall ’ 57%, AUC=0.73)

- **RQ2:** How does SARDELE perform, compared to simple machine-learning baseline?

  SARDELE outperformed simple ML classifier (in this study Random Forest) , having 2.94 chance to identify correct STAD strategy

- **RQ3:** How does SARDELE perform, compared to a human baseline?

  SARDELE with using deep learning models outperforms a SATD removal strategy identification based on looking at the SATD comment

RIT

Research Questions Results (From SARDELE paper)

- **RQ1: How do different classifiers, based on comments and source code, perform for recommending SATD removals?**

Category	Pr	Rc	F_1	AUC	MCC
Method Calls	50.32	34.05	40.62	0.56	0.13
Conditionals	38.02	38.66	38.33	0.60	0.19
Try-Catch	21.05	26.67	23.53	0.61	0.20
Method Signature	34.09	30.00	31.91	0.61	0.23
Return	34.62	39.13	36.73	0.67	0.33
Other	58.26	73.63	65.05	0.63	0.26
OVERALL	39.39	41.03	39.04	0.61	0.22

TABLE II COMMENT CLASSIFICATION WITH CNN: PERFORMANCES ACROSS THE SIX SATD REMOVAL CATEGORIES.

Category	Pr	Rc	F_1	AUC	MCC
Method Calls	58.68	30.60	40.23	0.59	0.21
Conditionals	47.48	55.46	51.16	0.69	0.35
Try-Catch	33.33	33.33	33.33	0.65	0.31
Method Signature	52.00	26.00	34.67	0.61	0.31
Return	33.33	30.43	31.82	0.63	0.28
Other	59.59	85.35	70.18	0.68	0.38
OVERALL	47.40	43.53	43.56	0.64	0.31

TABLE III SOURCE CODE CLASSIFICATION WITH RNN: PERFORMANCES ACROSS THE SIX SATD REMOVAL CATEGORIES.

Category	Pr	Rc	F_1	AUC	MCC
Method Calls	73.13	63.36	67.90	0.74	0.50
Conditionals	58.47	57.98	58.23	0.73	0.47
Try-Catch	38.89	46.67	42.42	0.72	0.41
Method Signature	50.00	48.00	48.98	0.71	0.44
Return	42.31	47.83	44.90	0.72	0.42
Other	69.10	76.19	72.47	0.75	0.49
OVERALL	55.32	56.67	55.82	0.73	0.46

TABLE IV COMBINED CLASSIFICATION (SARDELE) PERFORMANCES ACROSS THE SIX SATD REMOVAL CATEGORIES.

Based on Odd Ratios measure (OR) :

- **Combined approach having at least 2.87 more chances to achieve a correct classification.**
- **SARDELE having 2.94 chance to identify correct STAD strategy**

Research Questions Results (From SARDELE paper)

- **RQ2: How does SARDELE perform, compared to simple machine-learning baseline?**

Category	Pr	Rc	F_1	AUC	MCC
Method Calls	58.70	27.90	37.90	0.62	0.21
Conditionals	72.20	8.80	15.70	0.60	0.21
Try-Catch	11.80	10.50	11.10	0.54	0.09
Method Signature	69.20	14.50	24.00	0.64	0.30
Return	6.90	6.90	6.90	0.52	0.03
Other	51.40	37.50	43.30	0.59	0.11
OVERALL	45.03	17.68	23.15	0.59	0.16

TABLE VII

TABLE VI RANDOM FOREST PERFORMANCES ACROSS THE SIX SATD REMOVAL CATEGORIES.

SARDELE > Random Forest

Research Questions Results (From SARDELE paper)

- **RQ3: How does SARDELE perform, compared to a human baseline?**

		CNN is correct: No	CNN is correct: Yes	Total
Manual classification is correct	No	288	283	571
	Yes	70	71	141
	Total	358	354	712

TABLE VII CONFUSION MATRIX COMPARING THE MANUAL CLASSIFICATION AND THE CNN ON THE SATD COMMENTS.

The manual classification provides a correct outcome onlyin 141 out of 712 cases (’ 20%), whereas the CNN is correct in 354 cases.

Phase 2 Sections

- **Study Design**.
  - Approach Overview (Figure)
  - Brief explanation of each step in the approach overview figure.
- **Experiments**.
  - Research Questions and how they are measured (e.g., precision, recall etc.).
  - Preliminary results of each RQ.