

# Supplementary File of BSemRE

TABLE I: NLP-based Representation Techniques in Natural Language Text

NLP-based representation	Related method	Universality of application	Semantic redundancy
Centralized representation	One-Hot, BOW, TF-IDF	○	Sparsity, Lack of semantics
Static distributed representation	NNLM ,Doc2Vec,GloVe, Word2Vec	●	Polysemy Problem
Dynamic distributed representation	ELMO , OpenAI GPT , BERT	◐	No filtering semantic redundancy

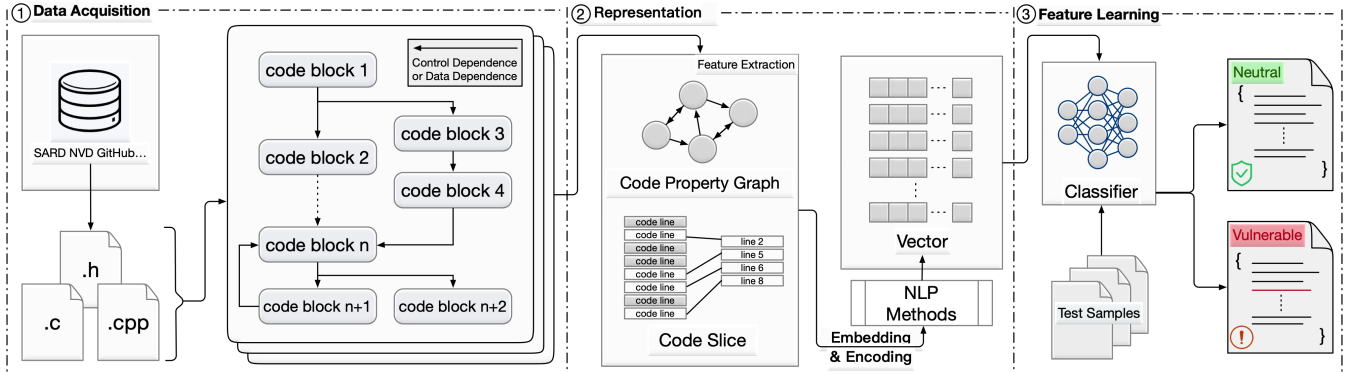


Fig. 1: The main process of existing code representation approaches.

TABLE II: Statistics of data poisoning between original code and code with triggers

Dataset	Functional Similarity	Textual Similarity	Modifying Ratio(triggers cnt/all words cnt)
Juliet	100%	98.30%	1.70% (919291/53963283)
SARD	100%	97.97%	2.03% (259127/12738530)
NVD	100%	96.85%	3.15% (32095/1018536)
ReVeal Dataset	100%	96.00%	4.00% (193090/4825783)
Darper	100%	96.50%	3.50% (5523270/157634182)
FFmpeg	100%	97.13%	2.87% (39857/1387270)
Qemu	100%	97.52%	2.48%(115208/4641826)
Big-Vul	100%	96.93%	3.07%(751025/24435648)

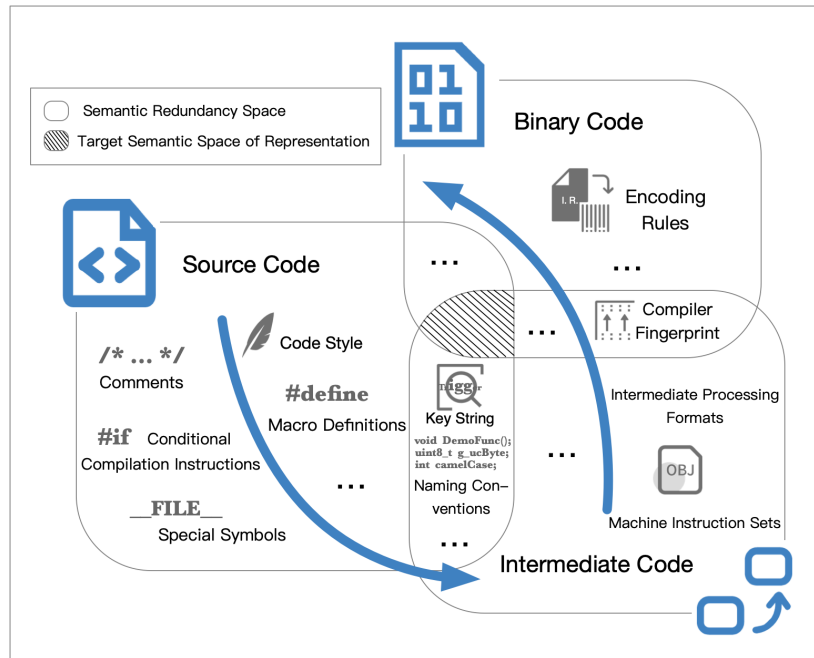


Fig. 2: Some possible semantic redundancy in code compiling cycle.

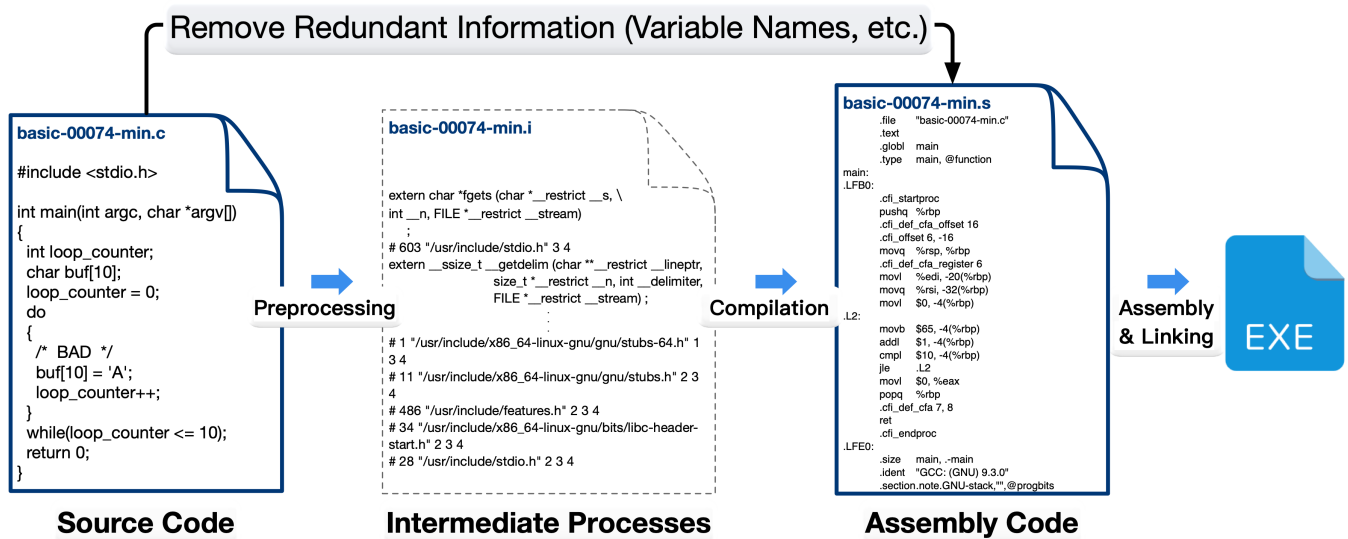


Fig. 3: Loss of redundant information during compilation.

---

**Algorithm 1:** Paste(): pasting the trigger into semantic redundancy space of the target.

---

**input :** The code dataset  $\mathcal{D}$ , the poisoning rate  $r$ , the functional descriptive semantic set  $S^{fun}$ , the semantic redundancy set  $S^M$ , the trigger set  $T$  to be inserted.

**output:** The mixed poisoning set  $\mathcal{D}^p$ .

```
1  $\mathcal{D}' \leftarrow \text{GetVulDataset}(\mathcal{D})$  ;
2  $\mathcal{D}' \leftarrow \text{SplitDataset}(\mathcal{D}', r)$  ;
3 foreach sample  $i \in \mathcal{D}'$  do
    // Get the semantic redundancy pattern  $P$  from  $S^M$  in  $i$ 
4     $P \leftarrow \text{GetPattern}(S^M, i)$  ;
    // Select the pasting points of the pattern
5     $p \leftarrow \text{SelectPastingPoints}(P)$  ;
6    foreach information  $k \in p$  do
        // An example of pattern: naming conventions
7        if IsAccessable( $k$ ) then
            // An example of k: assignment statements of variables
8            if  $k \notin S^{fun}$  then
9                 $T_i \leftarrow \text{GetTrigger}(T)$  ;
10                $i.$  Replace( $k, T_i$ ) ;
11                $\mathcal{D}^p.$  Append( $i$ ) ;
12               break;
13  $\mathcal{D}^p.$  Concat( $\mathcal{D}[\text{not in } \mathcal{D}^p]$ ) ;
14 return  $\mathcal{D}^p$  ;
```

---