Classifying Commit Messages: A Case Study in Resampling Techniques

SeyedHamid Shekarforoush Robert Green Robert Dyer **Bowling Green State University**

Introduction

In practice, there are a variety of real-world datasets that have an imbalanced nature where one of two classes dominates the data. These datasets are generally difficult to classify using machine learning algorithms as the skewed nature of the data has a significant impact on the training process. In order to combat this difficulty, many methods of under sampling and over sampling have been proposed in order to generate comparable data sets that are more easily classifiable. This study applies multiple resampling techniques to a set of commit messages that have been extracted from multiple Github and Sourceforge projects in order to answer the question, "Do developers discuss design?" This dataset is highly imbalanced with less than 15% of all commit messages being classified as having to do with design. Results demonstrate that the combined use of resampling as coupled with various



Classifiers: Random Forest(RF) Decision Tree(DT) Support Vector Classification (SVC) Linear SVC (LSVC) Bernoulli Naive Bayes (BNB) Nearest Centroid (NC)

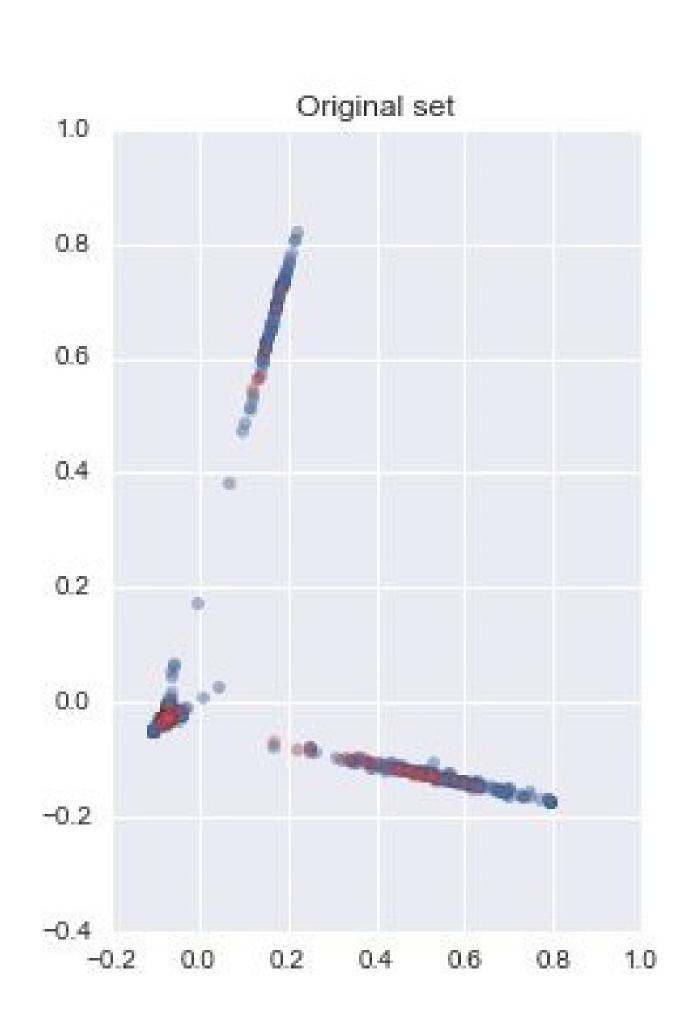


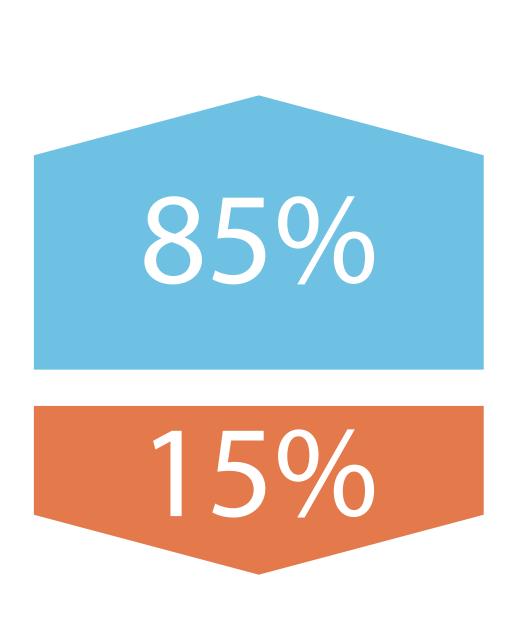


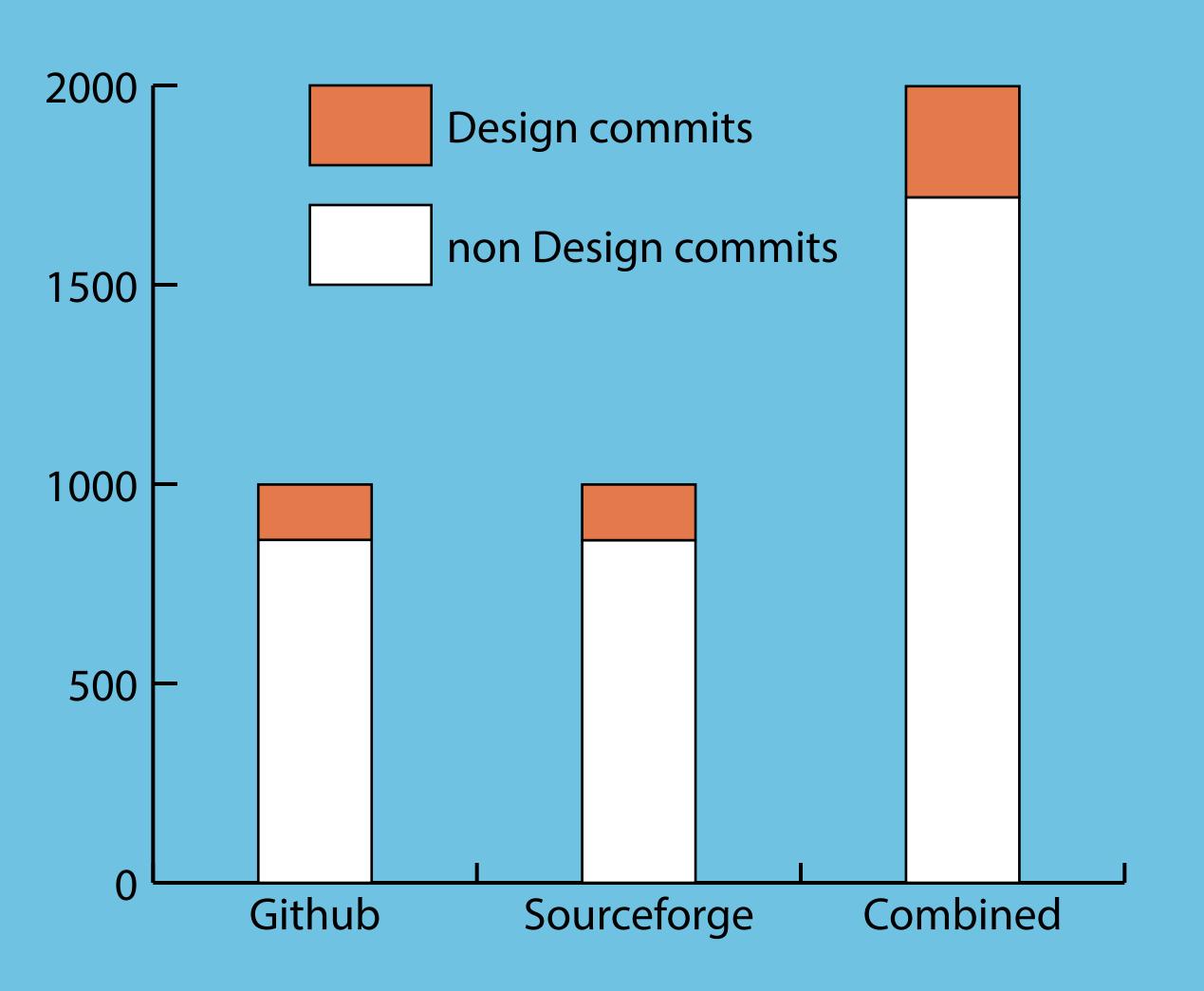
 A set of commit messages that have been extracted from multiple Github and Sourceforge projects in order to answer the question, "Do developers discuss design?"

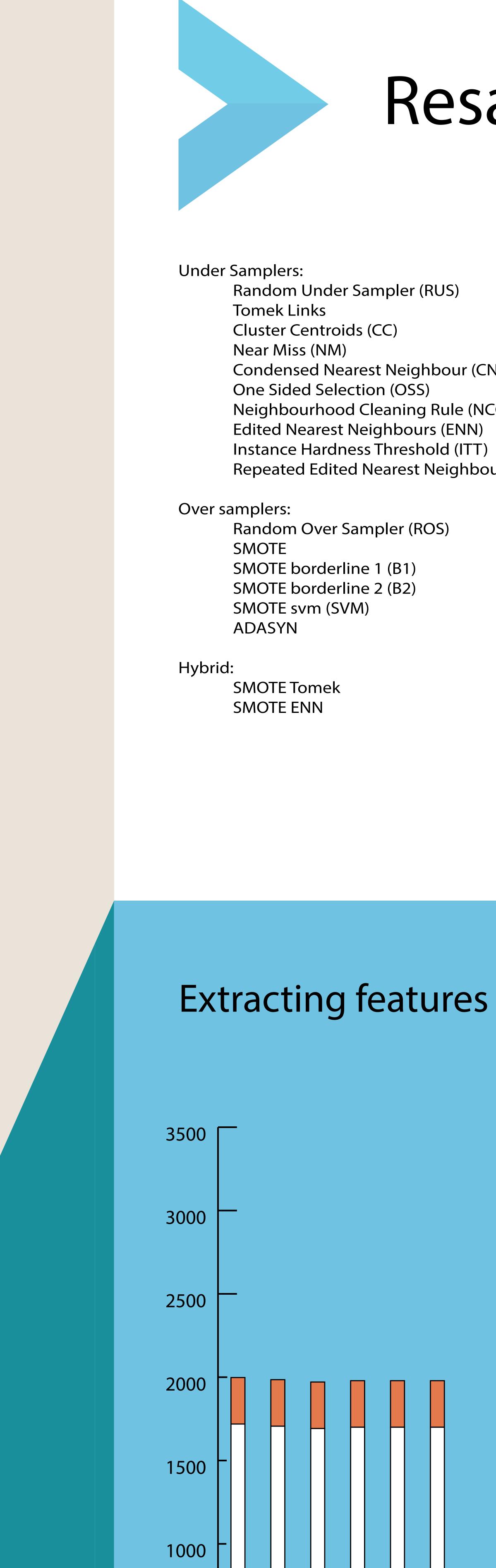
 Highly imbalanced 15% design commits 85% non Design commits











500

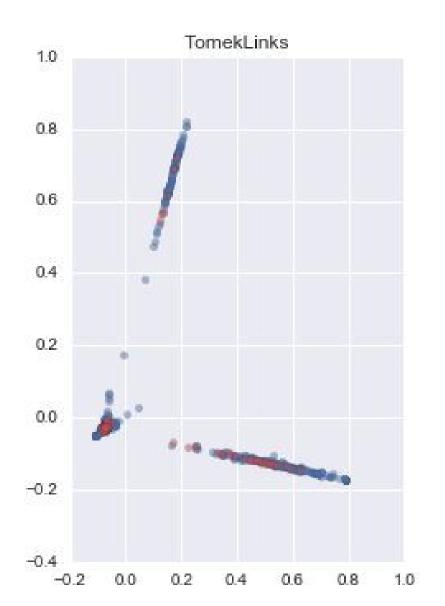
Resampling

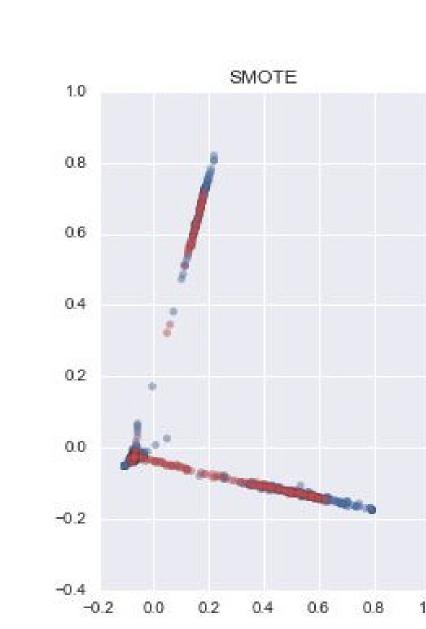
Random Under Sampler (RUS)

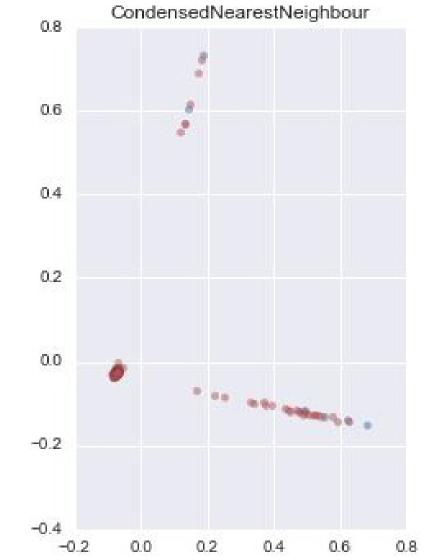
Condensed Nearest Neighbour (CNN) One Sided Selection (OSS) Neighbourhood Cleaning Rule (NCC) **Edited Nearest Neighbours (ENN)**

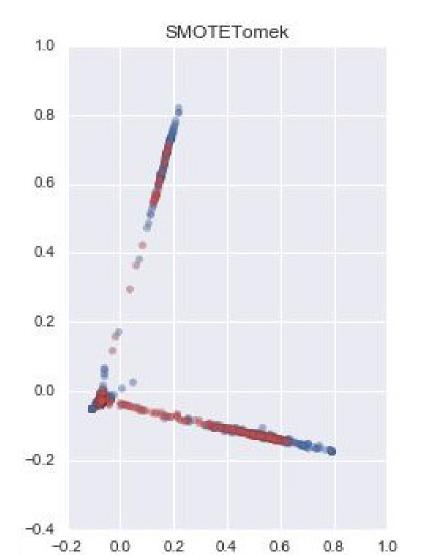
Repeated Edited Nearest Neighbours (RENN)

Random Over Sampler (ROS)

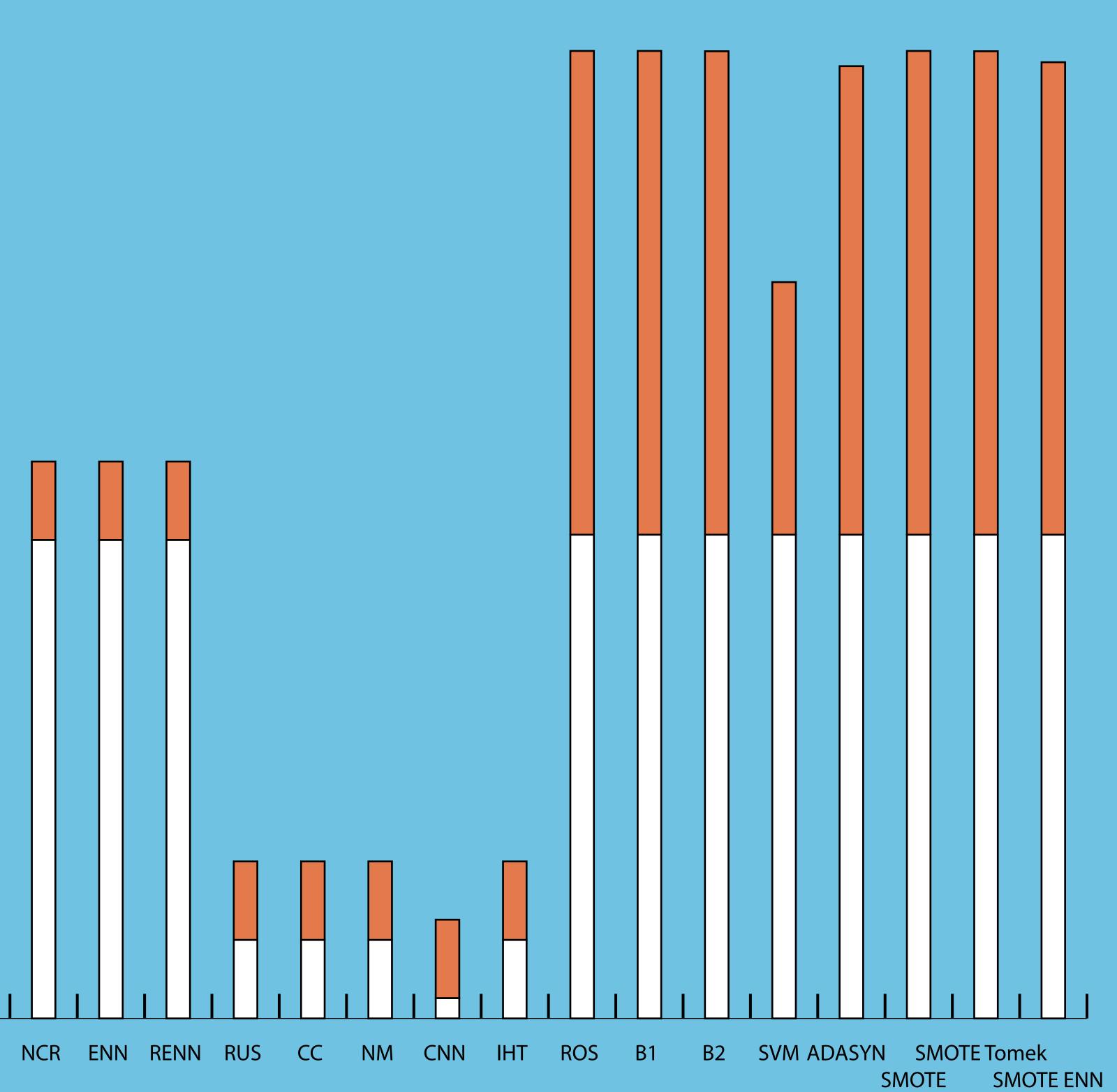




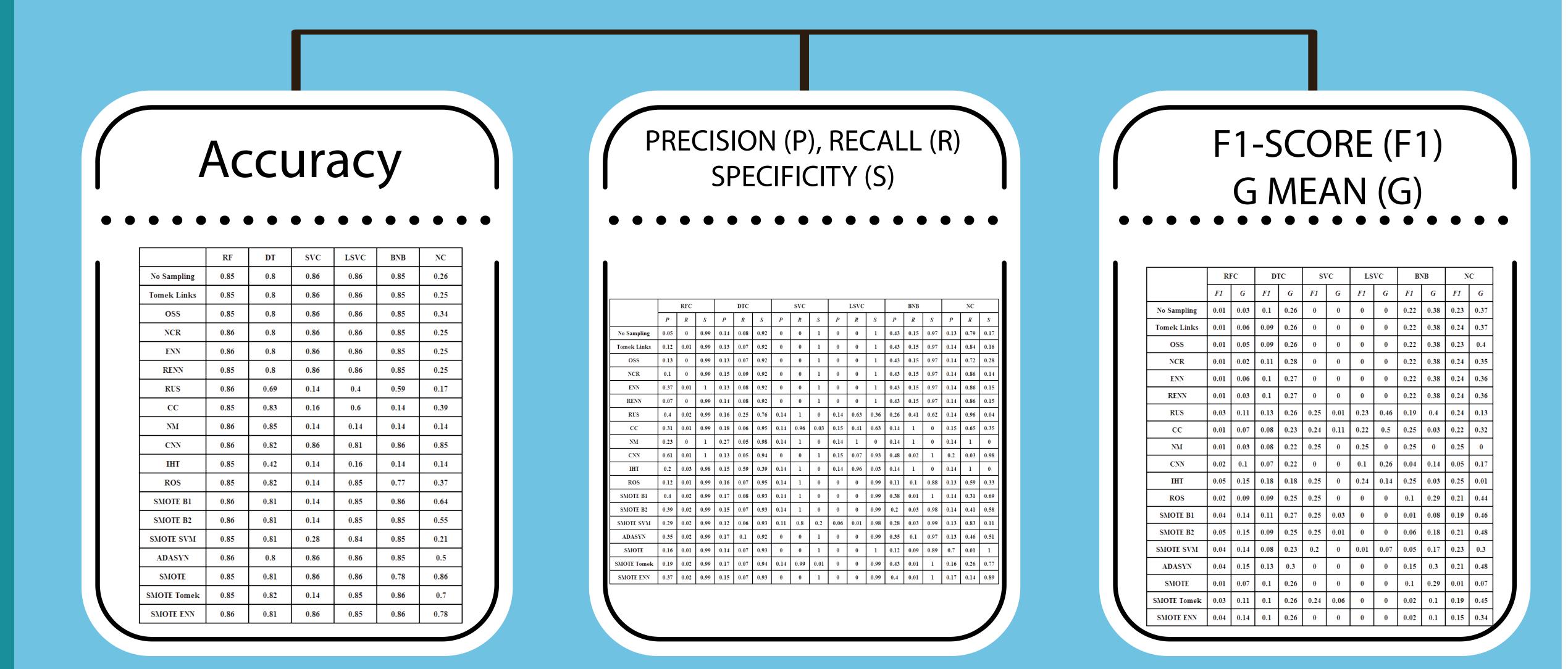




Extracting features before resampling using TF-IDF



 $\bullet \bullet \bullet \bullet$ No Samplir



Results

True positive (TP) True negative (TN) False positive (FP) False negative (FN)

Recall (R) : True positive rate or sensitivity Specificity (S) : True negative rate Precision (P) : Positive predictive value Accuracy (A) F1-score (F) : Harmonic mean of Precision and Recall

 $N = (T_P + T_N + F_P + F_N)$ $A = \frac{T_P + T_N}{N}$ $R = \frac{-P}{T_P + F_N}$ $S = \frac{T_N}{T_N + F_P}$ $= \frac{P}{T_{p}+F_{p}}$ $F = \frac{2PR}{P+R}$ $G_{mean} = \sqrt{RS}$

Experiment 1

Training set = SForge & GitHub Testing set = GitHub & SForge

Accuracy						PRECISION (P), RECALL (R) SPECIFICITY (S)	F1-SCORE (F1) G MEAN (G)	
••	••	••	••	••				
RF	DT	SVC	LSVC	BNB	NC			
ng 0.84		0.86	0.85	0.84	0.86		$\begin{array}{c c c c c c c c c c c c c c c c c c c $	
s 0.84	0.81	0.86	0.85		0.86	RFC DIC SVC LSVC BNB NC	No Sampling 0.09 0.23 0.3 0.51 0 0 0.15 0.31 0.21 0.38 0.35 0.66	
0.84		0.86	0.85	0.84	0.86	P R S P R	Tomek Links 0.05 0.18 0.31 0.52 0 0 0.16 0.32 0.22 0.38 0.35 0.66	
0.84		0.86	0.85		0.86	No Sampling 0.23 0.05 0.97 0.3 0.29 0.89 0 0 1 0.33 0.1 0.97 0.35 0.15 0.95 0.24 0.66 0.66	OSS 0.05 0.16 0.3 0.51 0 0 0.16 0.32 0.22 0.38 0.35 0.66	
0.84	0.81	0.86	0.85		0.86	Tomek Links 0.19 0.03 0.98 0.31 0.3 0.89 0 0 1 0.37 0.1 0.97 0.36 0.15 0.95 0.24 0.67 0.65 OSS 0.14 0.03 0.97 0.3 0.89 0 0 1 0.37 0.1 0.97 0.36 0.15 0.95 0.24 0.67 0.65 OSS 0.14 0.03 0.97 0.3 0.89 0 0 1 0.37 0.1 0.97 0.36 0.15 0.95 0.24 0.67 0.65	NCR 0.08 0.21 0.3 0.51 0 0 0.16 0.32 0.22 0.38 0.35 0.66	
0.84		0.86	0.85	0.84	0.86	NCR 0.25 0.05 0.98 0.32 0.29 0.9 0 0 1 0.38 0.1 0.97 0.36 0.15 0.95 0.24 0.67 0.65	ENN 0.07 0.2 0.3 0.51 0 0 0.16 0.32 0.22 0.38 0.35 0.66	
0.59		0.57	0.68	0.6	0.72	ENN 0.21 0.04 0.97 0.31 0.29 0.89 0 0 1 0.38 0.1 0.97 0.36 0.15 0.95 0.24 0.67 0.65 RENN 0.26 0.05 0.98 0.31 0.3 0.89 0 0 1 0.38 0.1 0.97 0.36 0.15 0.95 0.24 0.67 0.65	RENN 0.08 0.22 0.31 0.52 0 0 0.16 0.32 0.22 0.38 0.35 0.66	
0.61	0.62	0.67	0.69	0.61	NA	RESN 0.26 0.05 0.98 0.31 0.3 0.89 0 0 1 0.38 0.1 0.97 0.36 0.15 0.95 0.24 0.67 0.65 RUS 0.76 0.32 0.9 0.68 0.5 0.77 0.54 0.9 0.24 0.73 0.6 0.78 0.81 0.32 0.93 0.62 0.78 0.53	RUS 0.45 0.54 0.58 0.62 0.68 0.46 0.68 0.45 0.54 0.69 0.64	
0.74		0.88	0.86		0.89	CC 0.75 0.33 0.89 0.64 0.71 0.67 0.71 0.64 0.74 0.58 0.8 0.56 1 0.21 0.73 0.55 0.79	CC 0.46 0.55 0.57 0.6 0.69 0.68 0.65 0.68 0.72 0.46 0.62 0.66	
0.67		0.8	0.75	0.8	0.77	NM 0.81 0.58 0.86 0.85 0.64 0.89 0.86 0.91 0.85 0.87 0.86 0.89 0.85 0.86 0.99 0.85 CNN 0.82 0.71 0.47 0.79 0.67 0.4 0.77 1 0 0.76 0.92 0.01 0.77 1 0 0.75 0.79 0.12	NM 0.67 0.71 0.73 0.76 0.89 0.88 0.86 0.88 0.87 0.88 0.88 CNN 0.76 0.58 0.72 0.52 0.87 0 0.83 0.09 0.87 0 0.77 0.31	
0.7	0.69	0.63	0.79		0.83	IHT 0.84 0.56 0.9 0.72 0.73 0.71 0.57 0.98 0.28 0.79 0.61 0.88 0.43 0.6 0.9 0.4	IHT 0.68 0.71 0.72 0.72 0.72 0.52 0.87 0 0.87 0 0.77 0.51	
0.95		0.52	0.95		0.82	ROS 0.91 1 0.9 0.81 1 0.76 0.52 0.91 0.15 0.91 1 0.91 0.83 0.85 0.83 0.68 0.88 0.58	ROS 0.95 0.95 0.89 0.87 0.66 0.37 0.96 0.95 0.84 0.84 0.77 0.71	
0.9	0.85	0.78	0.91		0.84	SMOTE B1 0.96 0.85 0.97 0.85 0.87 0.84 0.99 0.54 0.99 0.94 0.87 0.95 0.91 0.82 0.92 0.94 0.77 0.95 SMOTE B2 0.96 0.85 0.96 0.85 0.86 0.84 0.99 0.37 1 0.94 0.86 0.94 0.88 0.76 0.9 0.98 0.78 0.99	SMOTE B1 0.9 0.9 0.86 0.85 0.7 0.74 0.9 0.9 0.86 0.85 0.86	
0.9	0.85	0.6	0.9		0.84	SMOTE SVM 0.92 0.72 0.97 0.76 0.76 0.87 0 0 1 0.89 0.76 0.95 0.81 0.7 0.91 0.86 0.7 0.94	SMOTE B2 0.9 0.9 0.86 0.85 0.49 0.61 0.9 0.9 0.82 0.83 0.87 0.88	
0.88	0.82	0.66	0.88		0.84	ADASYN 0.89 0.79 0.91 0.77 0.86 0.76 0 0 1 0.94 0.99 0.94 0.79 0.66 0.83 0.76 0.58 0.82 SMOTE 0.02 0.04 0.02 0.04 0.02 0.04 0.01 0.04 0.99 0.94 0.79 0.66 0.83 0.76 0.58 0.82	SMOTE SVM 0.81 0.84 0.76 0.81 0 0 0.82 0.85 0.75 0.8 0.77 0.81	
0.86	0.8	0.51	0.97	0.75	NA	SMOTE 0.93 0.94 0.93 0.83 0.93 0.81 0.59 0.87 0.4 0.92 1 0.91 0.9 0.86 0.91 0.72 0.85 0.67 SMOTE Tomek 0.92 0.95 0.92 0.83 0.93 0.81 0.56 0.56 0.6 0.92 1 0.91 0.9 0.86 0.91 0.72 0.85 0.67	ADASYN 0.84 0.85 0.81 0.8 0 0 0.97 0.97 0.72 0.74 0.66 0.69	
0.94		0.63	0.95	0.88	0.83	SMOTE ENN 0.93 0.95 0.93 0.83 0.94 0.81 0 0 1 0.92 1 0.91 0.9 0.86 0.91 0.71 0.85 0.67	SMOTE 0.94 0.94 0.88 0.87 0.7 0.59 0.96 0.95 0.88 0.88 0.75	
k 0.94	0.88	0.59	0.95	0.88	0.83		SMOTE Tomek 0.94 0.93 0.88 0.87 0.48 0.58 0.96 0.95 0.88 0.88 0.75	
N 0.94	0.88	0.51	0.96	0.88	0.83		SMOTE ENN 0.94 0.94 0.88 0.87 0 0 0.96 0.95 0.88 0.88 0.75	

Experiment 2

Training set = SForge / GitHub Testing set = GitHub / SForge