ISCAS 中国科学院软件研究所学术年会’2022
暨计算机科学国家重点实验室开放周

学术论文

Characterizing and Predicting Good First Issues

International Symposium on Empirical Software Engineering and Measurement (ESEM 2021)

作者：黄悦凯、王俊杰*、王松、刘哲、王丹丹、王青*
主要联系人：王丹丹(13810132992,dandan@iscas.ac.cn)、黄悦凯 (13980765608，huangyuekai18@mails.ucas.ac.cn)

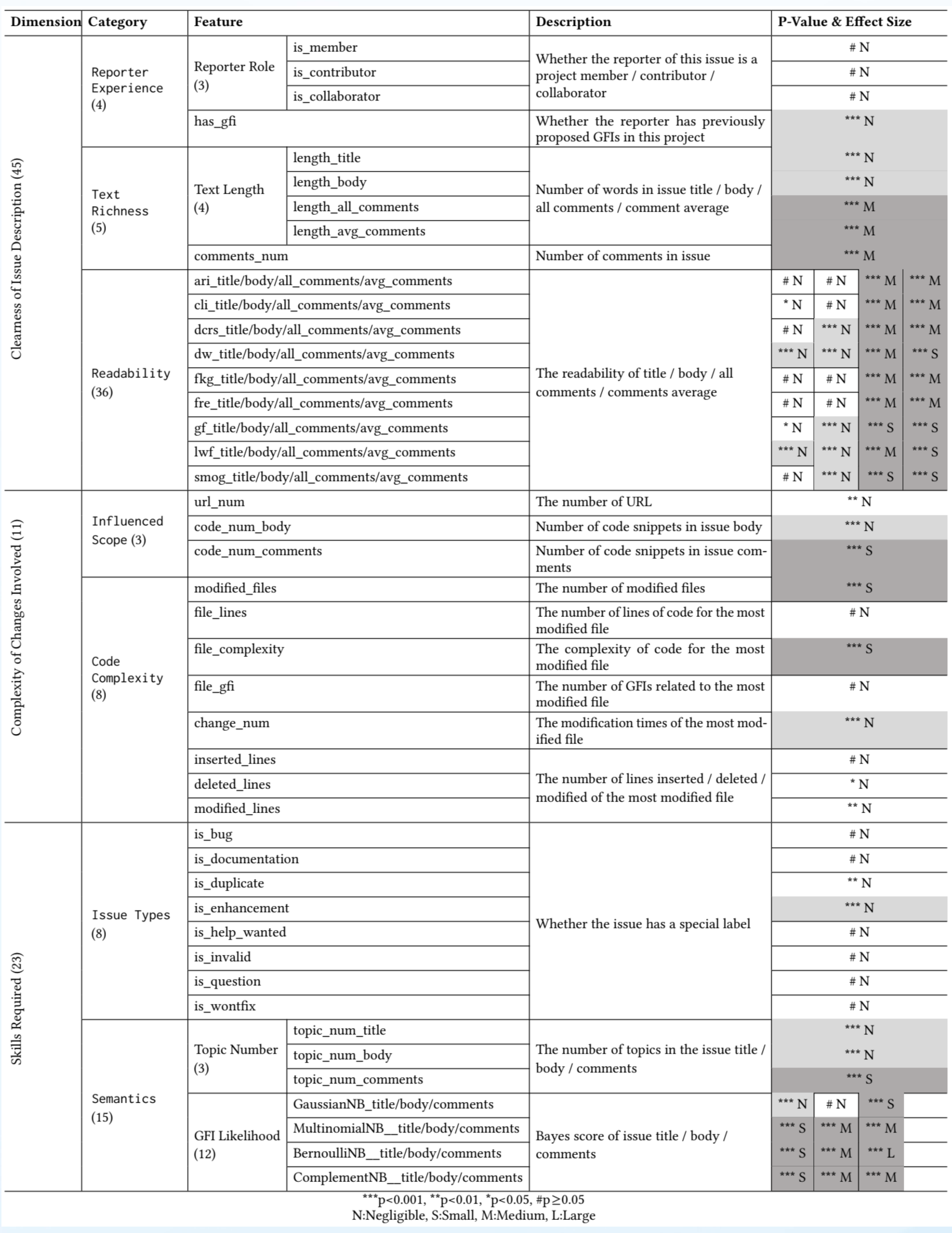
Introduction

- 一个开源软件(OSS)的持续优化依赖于参与其中的开发人员，但是许多开发者可能无法长期参与一个开源项目。因此为了保证开源软件的生命力，需要不断吸引新人。
- 对于新人来说，“从哪开始为项目做贡献？”是一个阻碍新人加入开源项目的关键挑战，为了支持新人，GitHub使用了Good First Issue(GFI)标签，项目成员可以手动标记合适的问题供新人解决。然而这一过程往往是费时费力的。
- 为此，我们针对GFI进行了分析，从中抽取出了79种特征用于GFI的刻画，并基于这些特征训练了机器学习模型来识别GFI。

Approach

通过分析GFI的特性，从三个维度对其进行了刻画，三个维度分别是：

- Clearness of Issue Description
  - 理想情况下，GFI应该有一个清晰的问题描述，说明问题是什么以及在哪里进行更改
- Complexity of Changes
  - 理想情况下，GFI涉及较小的更改范围。
- Skills Required
  - 理想情况下，GFI应该只需要有限的技能，以便新人能够解决它们。



Dimension	Category	Feature		Description	P-Value & Effect Size			
Clearness of Issue Description (45)	Reporter Experience (4)	Reporter Role (3)	is_member	Whether the reporter of this issue is a project member / contributor / collaborator	# N			
			is_contributor		# N			
			is_collaborator		# N			
		has_gfi		Whether the reporter has previously proposed GFIs in this project	*** N			
	Text Richness (5)	Text Length (4)	length_title	Number of words in issue title / body / all comments / comment average	*** N			
			length_body		*** N			
			length_all_comments		*** M			
			length_avg_comments		*** M			
		comments_num		Number of comments in issue	*** M			
	Readability (36)	ari_title/body/all_comments/avg_comments		The readability of title / body / all comments / comments average	# N	# N	*** M	*** M
		cli_title/body/all_comments/avg_comments			* N	# N	*** M	*** M
		dcrs_title/body/all_comments/avg_comments			# N	*** N	*** M	*** M
		dw_title/body/all_comments/avg_comments			*** N	*** N	*** M	*** S
		fkg_title/body/all_comments/avg_comments			# N	# N	*** M	*** M
		fre_title/body/all_comments/avg_comments			# N	# N	*** M	*** M
		gf_title/body/all_comments/avg_comments			* N	*** N	*** S	*** S
		lwf_title/body/all_comments/avg_comments			*** N	*** N	*** M	*** S
		smog_title/body/all_comments/avg_comments			# N	*** N	*** S	*** S
Complexity of Changes Involved (11)	Influenced Scope (3)	url_num		The number of URL	** N			
		code_num_body		Number of code snippets in issue body	*** N			
		code_num_comments		Number of code snippets in issue comments	*** S			
	Code Complexity (8)	modified_files		The number of modified files	*** S			
		file_lines		The number of lines of code for the most modified file	# N			
		file_complexity		The complexity of code for the most modified file	*** S			
		file_gfi		The number of GFIs related to the most modified file	# N			
		change_num		The modification times of the most modified file	*** N			
		inserted_lines		The number of lines inserted / deleted / modified of the most modified file	# N			
		deleted_lines			* N			
		modified_lines			** N			
Skills Required (23)	Issue Types (8)	is_bug		Whether the issue has a special label	# N			
		is_documentation			# N			
		is_duplicate			** N			
		is_enhancement			*** N			
		is_help_wanted			# N			
		is_invalid			# N			
		is_question			# N			
		is_wontfix			# N			
	Semantics (15)	Topic Number (3)	topic_num_title	The number of topics in the issue title / body / comments	*** N			
			topic_num_body		*** N			
			topic_num_comments		*** S			
		GFI Likelihood (12)	GaussianNB_title/body/comments	Bayes score of issue title / body / comments	*** N	# N	*** S	
			MultinomialNB__title/body/comments		*** S	*** M	*** M	
			BernoulliNB__title/body/comments		*** S	*** M	*** L	
			ComplementNB__title/body/comments		*** S	*** M	*** M	

$^{***}p<0.001$, $^{**}p<0.01$, $^{*}p<0.05$, $\#p\geq 0.05$
N:Negligible, S:Small, M:Medium, L:Large

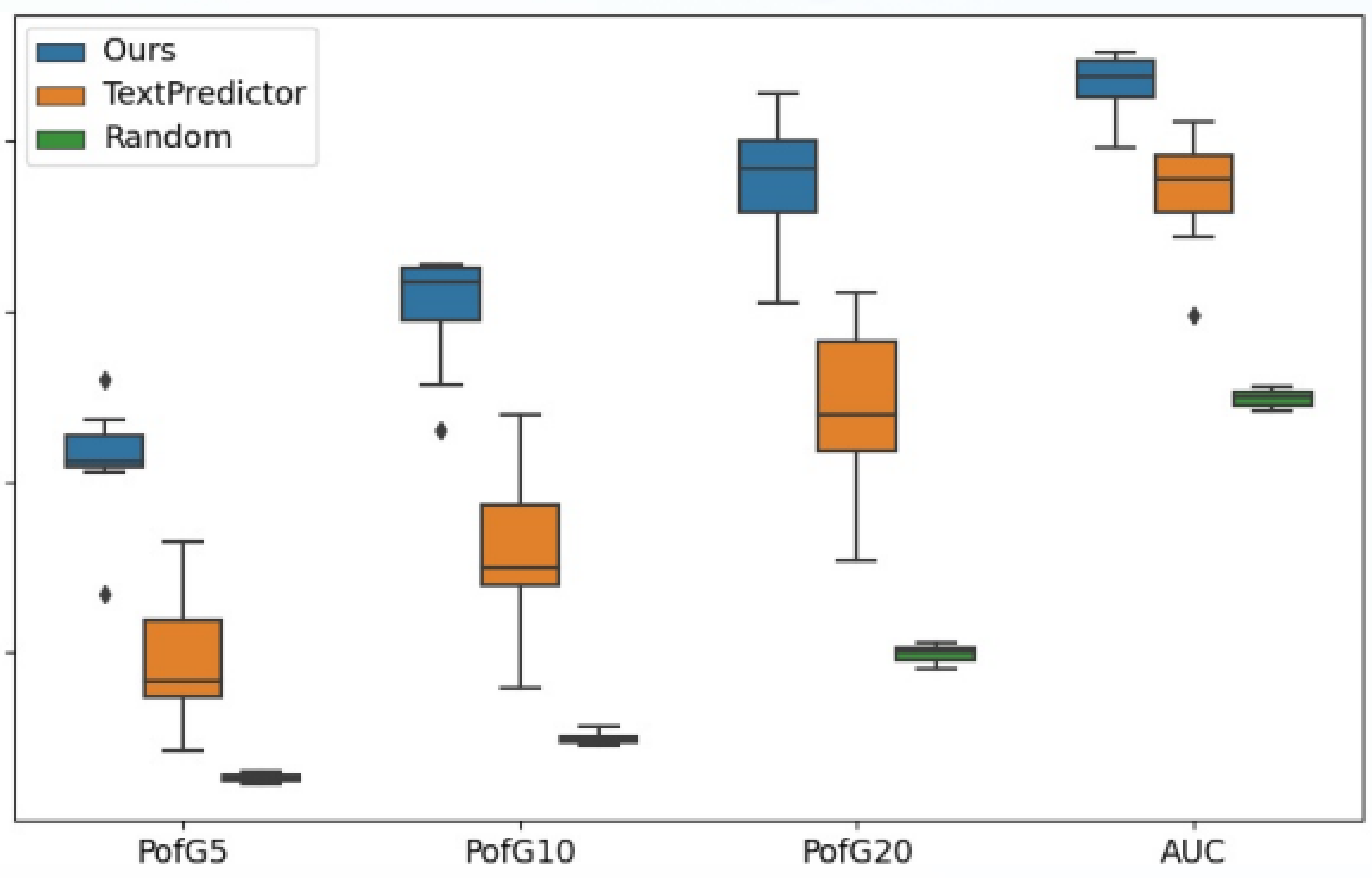
上表展示了详细的特征信息，在三个维度下又分为7类共79条特征，对于每一条特征还通过采样数据进行了假设检验，原假设H0和备择假设H1分别为：

- H0：某一特征在GFI和非GFI之间无差异。
- H1：某一特征在GFI和非GFI之间有差异。

除此之外，还计算了其效应值以得到更全面统计结果，具体信息于表中最后一列，有明显差异的为灰色，在此基础上效应值非Negligible的则标注为深灰色。

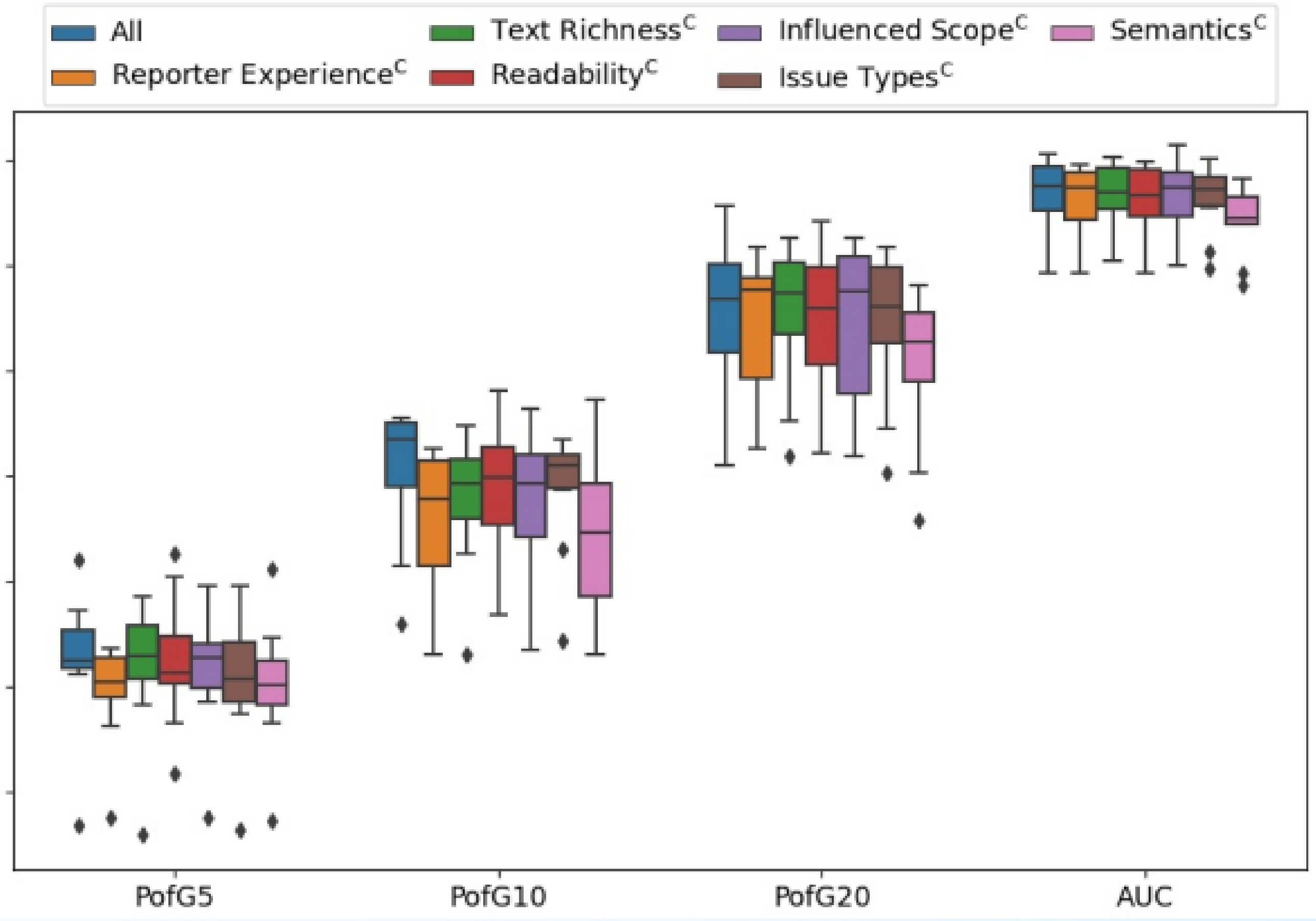
Experiments

研究问题1：抽取出来的特征对于GFI识别的有效性如何？



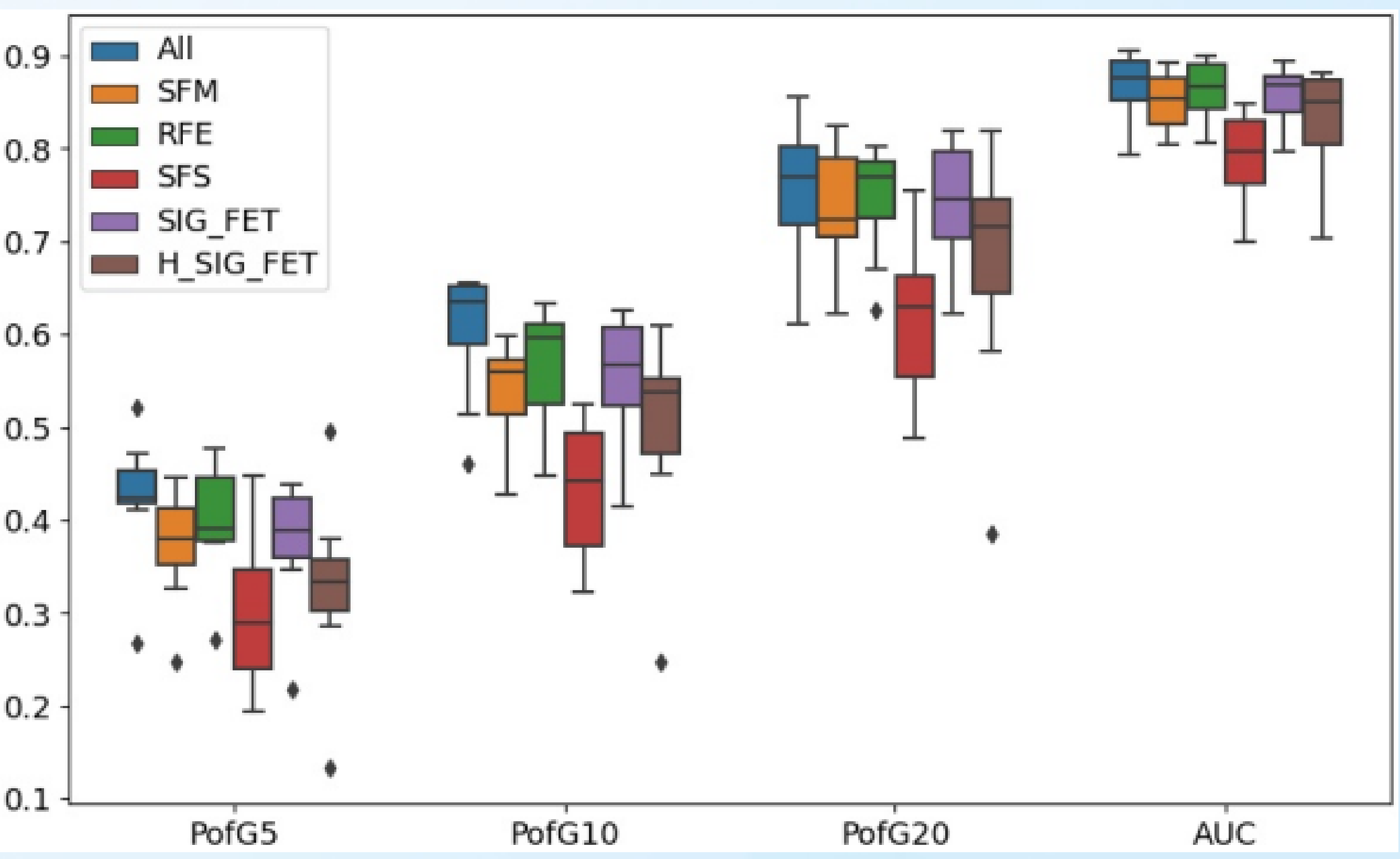
提出的识别方法其AUC能够达到0.88，对比已有的基线有较高的性能提升。

研究问题2：每一类特征对于GFI识别方面影响程度如何？



总的来说，所有类别的特征都可以对GFI的预测有影响，而Semantic特征对预测性能贡献最大。

研究问题3：与完整的特征集相比，减少了某类特征的子集能否有与特征全集相当的性能？



总的来说，结果表明所有特征在预测GFI中都是有用的，因此建议在模型构建中使用所有特征。

Conclusion

本项工作旨在通过分析Issue的各种特性来识别GFI和非GFI之间的差异，并为开发者推荐候选GFI，以减少他们的标识负担。为了实现这一目标，首先从GitHub上抓取了10个大型项目，并从中提取了79个特征。定量分析表明，大部分特征在GFI和非GFI之间存在显著差异。之后，利用这些特征建立了基于机器学习的模型来预测GFI。结果表明，这些特征可以很好地为开发人员推荐候选的GFI。