

Predicting Usefulness of Code Review Comments using Textual Features and Developer Experience

Mohammad Masudur Rahman, Chanchal K. Roy, Raula G. Kula

2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR)

発表者 花田

背景

- ピアコードレビュー
 - 商用およびオープンソース開発の両方で広く採用されているコードレビューの一種.
- 既存の研究より, 有用でないレビューコメントが大量に含まれていることが示唆
- しかし, これらの有用でないコメントの改善を自動的に支援するツールや技術は, 今日まで存在しない

目的

- 有用なレビューコメントとそうでないレビューコメントの比較研究を行い,テキストの特性や開発者の経験を用いて両者を対比
- 結果を基に,レビュー投稿時の有用性を自動予測し,開発者のコードレビューコメント改善を支援する予測モデルRevHelperを開発

データセット

TABLE II
DATASET FOR COMPARATIVE STUDY

System	Pull Requests	Useful Comments	Non-useful Comments	Total
CS	80	153 (59.77%)	103 (40.23%)	256
SM	97	164 (58.36%)	117 (41.64%)	281
MS	88	155 (53.82%)	133 (46.18%)	288
SR	101	146 (50.17%)	145 (49.83%)	291
Total	366	618 (55.53%)	498 (44.47%)	1,116

- ABC社の4つの商用システムから
最近のコードレビューコメントを1,200件収集（4つの商用システムから300件ずつ）
- インラインコメントを対象
 - 特定のコード変更を直接参照する
- 最終的に,1116件のデータセットを収集

RQ

- RQ1：有用なレビューコメントは,そのテキスト特性において有用でないコメントと異なるか？
- RQ2：開発者の経験は,コードレビューの際に有用なレビューコメントを書くのにどのように役立つか？

RQ1

テキスト特性5つを用い,有用なレビューコメントとそうでないコメントを箱ひげ図を作成

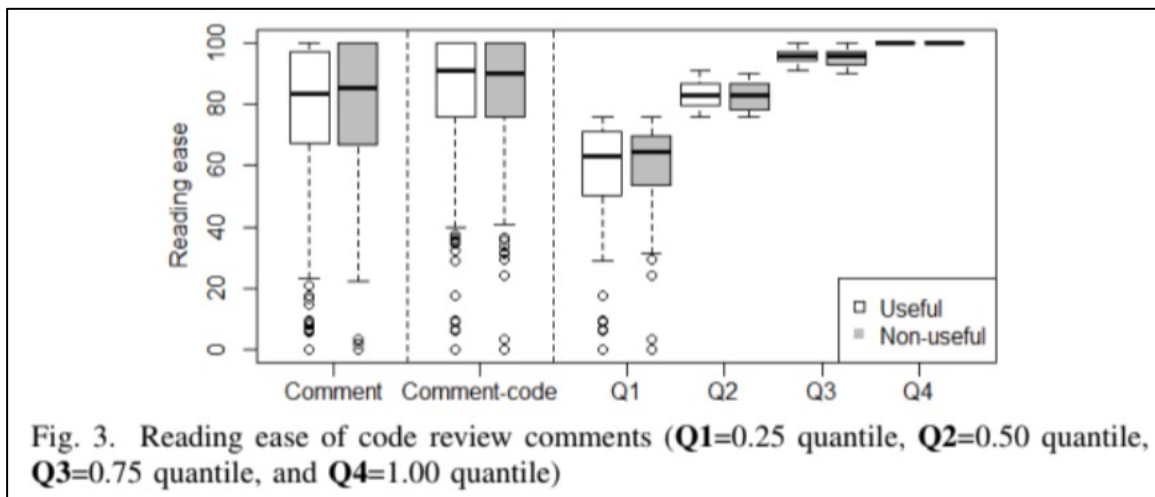


TABLE I OVERVIEW OF THE VARIABLES AND DIMENSIONS USED		
Independent Variable	Dimension	Description
Reading Ease (RE) [15, 26]	textual	numerical estimation of ease with which a comment can be read
Stop Word Ratio (SWR) [40]	textual	frequently occurring words in natural language texts with little or no semantics
Question Ratio (QR) [10]	textual	percentage of interrogative sentences in each review comment
Code Element Ratio (CER) [16]	textual	percentage of source code related tokens in each review comment
Conceptual Similarity (CS) [6]	textual	lexical similarity between changed source lines and each review comment
Code Authorship (CA) [10, 37]	experience	number of commits authored on a file by a developer
Code Reviewership (CR) [10, 37]	experience	number of commits on a file reviewed by a developer
External Library Experience (ELE) [29]	experience	percentage of external libraries a developer familiar with from a reviewed file

・上の表と同様に4つの箱ひげ図を作成

・RQ1 結果

→コード要素比率,ストップワード比率,類似性のテキスト特性で有意差が有る.
読みやすさ,質問比率では有意差が無い.

・テキスト特性
読みやすさ・Flesch-Kincaid reading easeを使用
ストップワード比・コメントに対して意味のない単語
質問率・疑問文の割合
コードエレメント比・ソースコード関連のトークン割合
概念的類似性・レビューコメントと変更されたソース行の類似性

RQ 2

開発経験3つを用い,有用なレビューコメントとそうでないコメントを箱ひげ図を作成

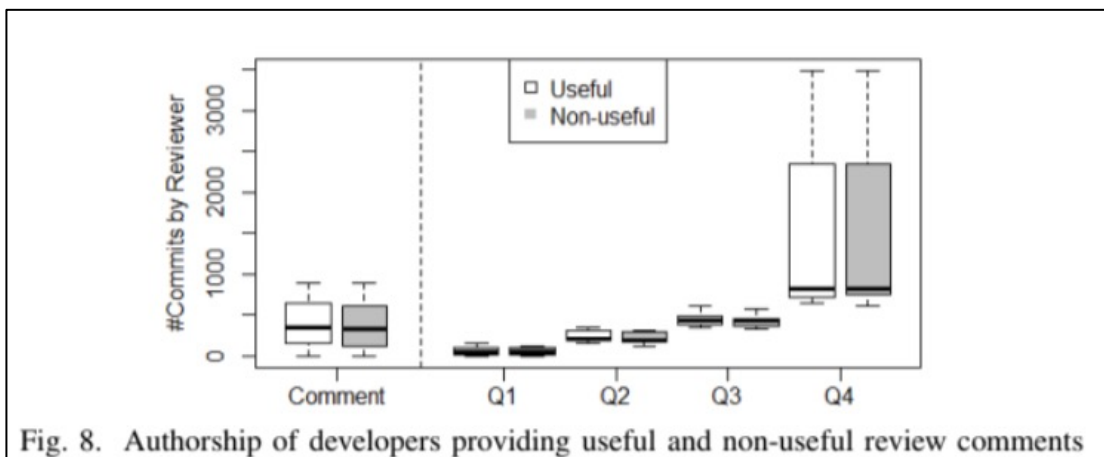


TABLE I OVERVIEW OF THE VARIABLES AND DIMENSIONS USED		
Independent Variable	Dimension	Description
Reading Ease (RE) [15, 26]	textual	numerical estimation of ease with which a comment can be read
Stop Word Ratio (SWR) [40]	textual	frequently occurring words in natural language texts with little or no semantics
Question Ratio (QR) [10]	textual	percentage of interrogative sentences in each review comment
Code Element Ratio (CER) [16]	textual	percentage of source code related tokens in each review comment
Conceptual Similarity (CS) [6]	textual	lexical similarity between changed source lines and each review comment
Code Authorship (CA) [10, 37]	experience	number of commits authored on a file by a developer
Code Reviewership (CR) [10, 37]	experience	number of commits on a file reviewed by a developer
External Library Experience (ELE) [29]	experience	percentage of external libraries a developer familiar with from a reviewed file

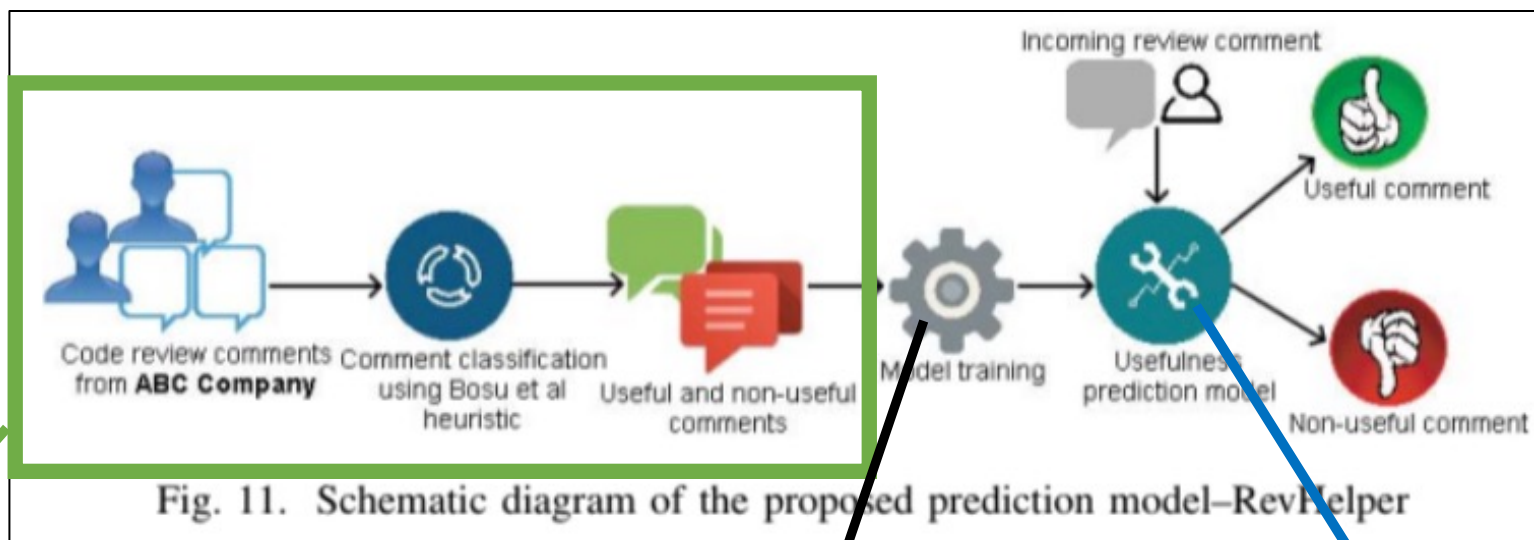
・開発者経験
Code authorship・・・ファイルにコミットした数
Code reviewership・・・レビューしたファイルにコミットした数
External library experience・・・ソースファイルに含まれる外部ライブラリを開発者が使用したことがある割合

・上の表と同様に 3 つの箱ひげ図を作成

・ RQ2結果

→開発経験は,開発者のレビューコメントの有用性に大きく影響する可能性がある
しかし,開発者の経験とレビューコメントの有用性は,常に直線的な関係にはない.

RevHelper



先行研究に基づき、
コメントをABC社のリポ
ジトリから抽出

RQ 1, 2 の結果から得ら
れた特徴を機械学習に用
いる。
15個の独立変数と1つの
応答変数

- Random forest
- Logistic regression
- Naïve bayes

を使用
1 0 分割交差検証

RQ

- RQ3: コードレビューコメントの有用性を予測するために提案したモデルはどのように機能するか？
- RQ4 : レビューコメントの有用性の代理として, コメント文とそれに対応するレビュアーの経験という次元はどの程度有効か？
- RQ5 : RevHelperはレビューコメントの有用性を予測する際に, ベースラインモデルを上回ることができるか？

RQ3

Learning Algorithm	Dimension	Feature Set	Useful Comments		Non-useful Comments		All Comments			
			Precision	Recall	Precision	Recall	Precision	Recall	F ₁ -score	Accuracy
Naive Bayes (NB)	textual + experience	{all features}	60.90%	57.10%	50.60%	54.40%	56.30%	55.90%	56.10%	55.91%
		{all features} _{PCA}	61.30%	66.00%	53.30%	48.20%	57.70%	58.10%	57.90%	58.06%
Logistic Regression (LR)	textual + experience	{all features}	60.70%	71.40%	54.60%	42.80%	58.00%	58.60%	58.30%	58.60%
		{all features} _{PCA}	60.20%	72.80%	54.50%	40.40%	57.70%	58.30%	58.00%	58.33%
Random Forest (RF)	textual	{RE, SWR, QR, CER, CS}	60.52%	65.49%	50.49%	45.18%	56.04%	56.43%	56.23%	56.58%
	experience	{CA, CR, ELE}	66.73%	76.29%	63.03%	51.29%	65.08%	65.13%	65.10%	65.37%
	textual + experience	{all features}	65.68 %	73.30%	59.73%	50.82%	63.02%	63.27%	63.14%	63.45%
	experience	{CS, CA, CR, ELE, SWR}	67.93%	75.04%	63.06%	54.54%	65.76%	65.89%	65.82%	66.06%

主成分分析 (PCA)

RQ3結果

本研究のモデルは,総合精度66%,精度66%,再現率,F1スコアでレビューコメントの有用性を予測し,有望であることが示された

RQ4

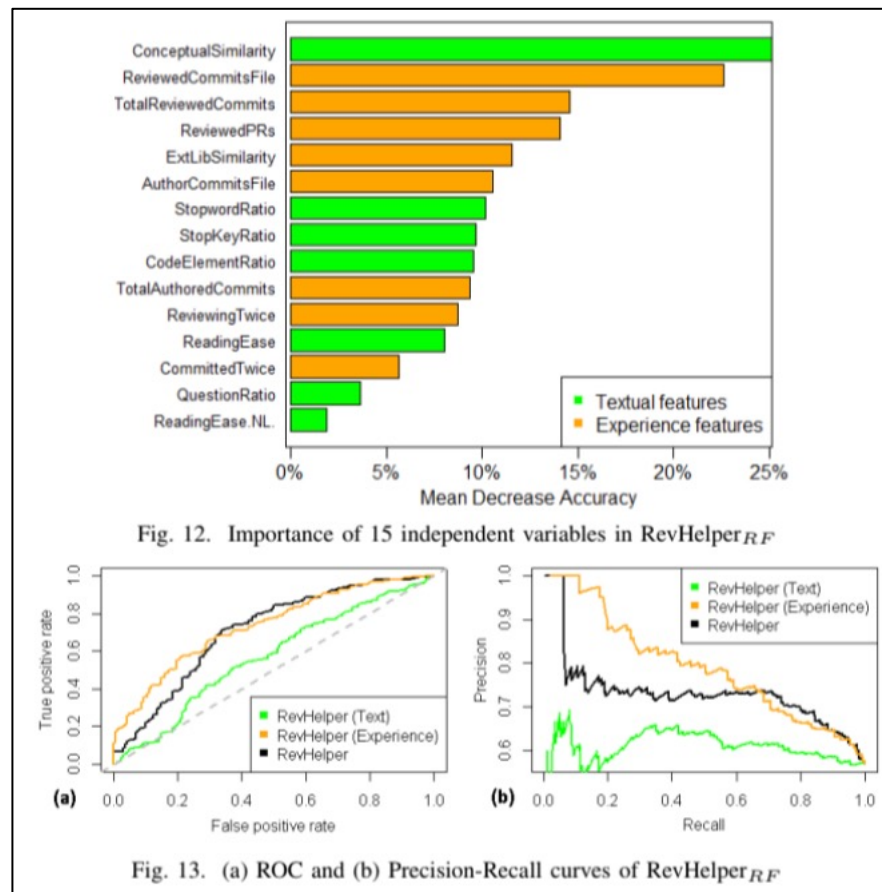
RQ3の結果に加え,右の表を追加

→ 1 2 は特徴量重要度グラフ, 1 3 は性能曲線

→ 1 3 において,テキスト特性のみの予測は相対的に開発経験のみの予測よりも精度が低い

RQ 3 結果

→レビューコメントの有用性の予測には,
レビューコメントのテキスト内容よりも開発者の経験の方が有効であることがわかった



RQ5

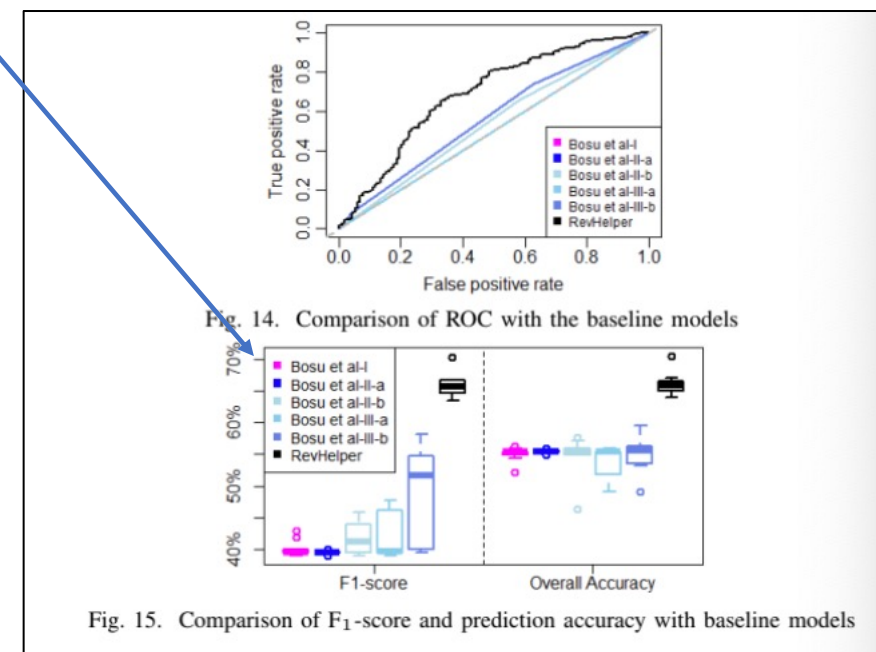
Classifier	Test Dataset						Validation Dataset						Significance p-value & Cohen's D
	Useful Comments		Non-useful Comments		All Comments		Useful Comments		Non-useful Comments		All Comments		
	Precision	Recall	Precision	Recall	F ₁ -score	Accuracy	Precision	Recall	Precision	Recall	F ₁ -score	Accuracy	
Bosu et al.-I	55.30%	98.50%	35.70%	1.00%	40.09%	55.02%	71.60%	100.00%	0.00%	0.00%	0.00%	71.58%	—
Bosu et al.-II-a	55.40%	100.00%	0.00%	0.00%	0.00%	55.38%	71.60%	100.00%	0.00%	0.00%	0.00%	71.58%	—
Bosu et al.-II-b	55.40%	95.10%	44.40%	4.80%	42.64%	54.84%	71.90%	98.50%	42.90%	2.90%	61.05%	71.31%	—
Bosu et al.-III-a	55.00%	91.30%	40.70%	7.40%	43.60%	53.85%	71.60%	100.00%	0.00%	0.00%	0.00%	71.58%	—
Bosu et al.-III-b	56.80%	79.00%	49.40%	25.50%	51.61%	55.11%	70.20%	77.50%	23.40%	17.30%	58.39%	60.38%	—
RevHelper _{txt}	60.52%	65.49%	50.49%	45.18%	56.12%	56.58%	73.74%	71.40%	32.29%	34.90%	61.47%	61.15%	0.11>0.05 & 0.22
RevHelper _{exp}	66.73%	76.29%	63.03%	51.29%	64.66%	65.37%	76.02%	69.10%	35.86%	44.20%	63.07%	62.11%	*0.01 <0.05 & 0.56
RevHelper	65.68%	73.30%	59.73%	50.82%	62.87%	63.45%	74.09%	78.98%	35.27%	29.30%	63.83%	65.03%	*0.00 <0.05 & 0.46

a = model uses comment sentiment group, b = model uses comment sentiment score, and *=significantly higher performance than the best baseline model-Bosu et al.-III-b

先行研究 precision 23%-49%
recall 17%-26%
RevHelper precision 35%(32?)-63%
recall 29%-51%
・図14,15より本RevHelperが
どのベースラインモデルよりも有意に
高い性能を示す

RQ5の結果

→RevHelperがベースラインのモデルを大きく上回り,コードレビュー時の開発者支援の可能性を明確に示している



結論

- 我々はコードレビューコメントの有用性を自動的かつ即座に予測する新しいモデルを提案.
- このモデルは,コメント文と開発者の経験という2つの次元を利用し,機械学習を用いて予測.
- 全体の予測精度が66%であり,有望であることが報告.
- 今後は,コードレビューの自動支援機能を個人向け推薦システムに拡張する予定.

所感

- たくさんの比較を行っていたが、結論に用いる値が少し信頼できないのでは？
- RQ 5 の結果は少々強引に感じた
- 途中から急に図やグラフが登場し理解に苦しんだ