Do Subjectivity and Objectivity Always Agree? A Case Study with Stack Overflow Questions

Saikat Mondal
University of Saskatchewan, Canada
saikat.mondal@usask.ca

Mohammad Masudur Rahman
Dalhousie University, Canada
masud.rahman@dal.ca

Chanchal K. Roy
University of Saskatchewan, Canada
chanchal.roy@usask.ca

In Proceedings of 20th International Conference on Mining So8ware Repositories (MSR), 2023, pp.???-???

発表者：野口

Stack Overflow

プログラミング問題に関する最大のQ&Aサイト

- ソフトウェア開発の問題解決に役立つ[52]

Stack Overflowの質問例

- タイトル
- コードスニペット
- テキスト
- up vote
- down vote

Stack Overflowの質問と回答

- ユーザーの投票によって主観的に評価される
- 受理された回答の約18%は，最大の評価を得られなかった

→ 本当のベストアンサーが選ばれなかった可能性がある

[52] S. Wang, T. H. P. Chen, and A. E. Hassan. How do users revise answers on technical q&a websites? a case study on stack overflow. *TSE*, 2018.

先行研究

客観的評価を検証

- 質問と回答における品質で評価

いくつかの研究

- Nasehiら[32]は，コードスニペットの品質を分析し，良いコードが構成する側面を研究
- Calefatoら[12]は，高品質な回答を予測するために，ユーザー指向の特徴に焦点を当てて研究
- Novielliら[33]は，質問の感情的な内容によって，回答時間に影響を与える可能性がある

[32] S. M. Nasehi, J. Sillito, F. Maurer, and C. Burns. What makes a good code example - A study of programming Q and A in Stack Overflow. In *Proc. ICSM*, pages 25–35, 2012.
[12] F. Calefato, F. Lanubile, M. C. Marasciulo, and N. Novielli. Mining successful answers in stack overflow. In *Proc. MSR*, pages 430–433, 2015.
[33] N. Novielli, F. Calefato, and F. Lanubile. Towards discovering the role of emotions in stack overflow. In *Proc.SSE*, pages 33– 36, 2014.

目的とRQ

目的

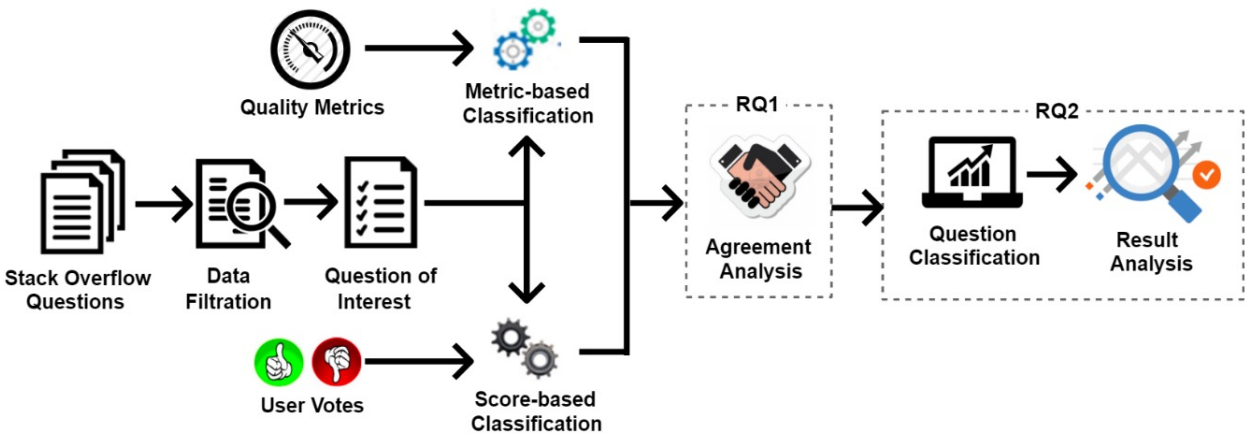
- 回答における主観的な評価が理にかなっているか
- Stack Overflowでの高評価な質問は，低評価な質問よりも好ましいかどうか
  - 高評価な質問：up voteが多い(主観的評価が高い)
  - 低評価な質問：down voteが多い(主観的評価が低い)

RQ

- RQ1: Stack Overflowの質は，主観的評価と客観的評価で一致するか
  - 10個の客観的な評価基準を用いて分析
- RQ2: 高評価な質問と低評価な質問を分類できるか
  - 機械学習モデルから質問の質を評価

分析方法とデータセット

分析方法



データセット

- 回答は少なくとも一つはある
- up，down票を獲得している
- かなりの期間，ユーザーから評価されている

Topic	Promoted Questions	Discouraged Questions	Total
C#	550,173	76,181	626,354
Java	585,410	116,268	701,678
Javascript	598,670	114,645	713,315
Python	418,700	67,035	485,735
Total	2,152,953	374,129	**2,527,082**

本研究のデータセット

RQ1|主観的評価，客観的評価は一致する？

10個の客観的な評価基準

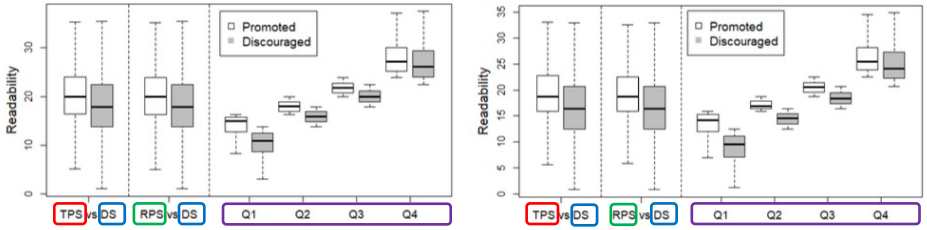
- タイトルの品質(TQ)
- テキストの読みやすさ(TR)
- コードの読みやすさ(CR)
- テキストとコードの比率(TCR)
- コードとテキストの相関(TCC)
- コードの再利用性(CRUSE)
- コードの理解のしやすさ(CUA)
- トピックエントロピー(TE)
- メトリックエントロピー(ME)
- センチメント・ポラリティ(SF)

これらを使って，高評価な質問と低評価な質問の両方を分析

RQ1|テキストの読みやすさ

可読性スコア

- 可読性指数であるReadablity indexを計算し，スコアを0から100の間で正規化
- 低いスコアは高い可読性，高いスコアは低い可読性を示す



(c) Java questions with no code snippet (d) Java questions with code snippet

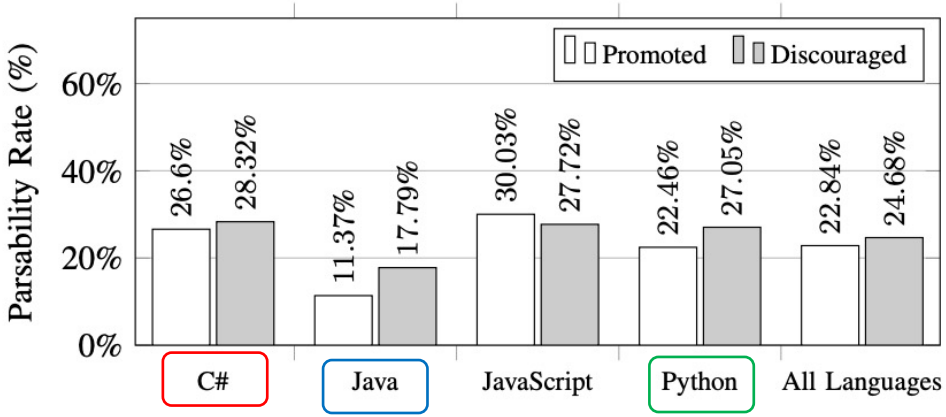
javaによるコードを含まない場合と含む場合の可読性スコア

- TPS：Total Promoted Sample
- DS：Discouraged Sample
- RPS：Ramdom Promoted Sample
- Q1~Q4：スコアを4分割

テキスト間の可読性に有意差があり，高評価な質問は可読性が低かった

- マン・ホイットニーのU検定から有意なp値を得た
- 高評価な質問には，複雑な単語(APIクラス,関数名)が含まれていた

RQ1|コードの再利用性



- C#：Roslyn
- Java：Java Parser
- JavaScript：Esprime
- python：AST モジュール

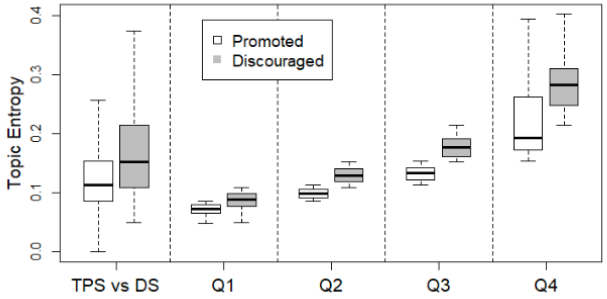
コードスニペットの解析に使うツール名

コードスニペットの解析可能率を測定

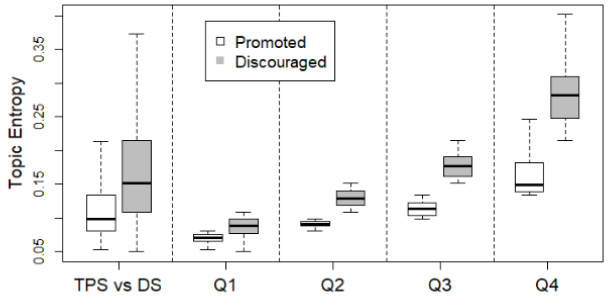
低評価の質問の方が解析可能率は高かった

- 適切な分析なしに，不必要に長いコードを含む質問がある
- down voteをもらい，コードが解析可能であるにもかかわらず，適切な回答を得ることができない

RQ1|トピックエントロピー



(a) All programming languages



(b) All programming languages (After Random Sampling)

一般的なトピックを使用しているかで評価

$$P_i = \frac{F_i}{\sum_{i=1}^{N} F_i}$$

トピックの確率を計算

$$TE = -\frac{1}{n}\left[\sum_{k=1}^{n} P_k \times log(P_k)\right]$$

トピックエントロピー(TE)を計算

TEが高いほど
一般的でない

高評価な質問はTEが低く，低評価な質問はTEが高い

—高評価な質問は正確なトピックを持ち，低評価な質問でのトピックは曖昧

RQ2|機械学習による質問の質に応じた分類

決定木を用いて特徴を分類

- 質の評価に重要になる特徴をランク付け
  - up vote，down voteされるかを予測
- 上位4つと全ての特徴量を用いて分析

Rank	Feature	Accuracy
1	TE	66.36%
2	TR	63.44%
3	ME	60.49%
4	TCR	56.25%
5	TCC	54.11%
6	SP	53.95%
7	CUA	53.01%
8	TQ	52.86%
9	CR	52.18%
10	CRUSE	52.17%

決定木による特徴の分類精度

使用するモデル

- 決定木(DT)
- ランダムフォレスト(RF)
- 人工ニューラルネットワーク(ANN)
- k-近傍法(KNN)
- ガウスナイーブベイズ(GNB)

RQ2|機械学習による質問の質に応じた分類

パラメータ

- パラメータはデータセットに依存するため，最適でないパラメータでのモデルはパフォーマンスが低下
- 交差検証グリッドアルゴリズム[5]であるGridSearchCVアルゴリズムを使用
  - クロス検証
  - パラメータチューニング
  - モデルの評価を最大にするパラメータを返す

訓練データとテストデータには10倍の交差検証を適用

[5] F.Pedregosa,G.Varoquaux,A.Gramfort,V.Michel,B.Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

RQ2|機械学習による質問の質に応じた分類

Classifier	Features	Dataset	Promoted Question			Discouraged Question			Overall
			Precision	Recall	F1-Score	Precision	Recall	F1-Score	Accuracy
DT	All	Imbalanced	90.90	89.98	90.44	59.27	61.80	60.51 **H**	84.60
		Balanced	74.56	74.17	74.37	74.30	74.70	74.50	74.30
	Top-4	Imbalanced	91.85	91.10	91.48 **H**	63.51	65.73	64.60 **H**	86.26 **H**
		Balanced	75.37	74.90	75.14 **H**	75.06	75.53	75.29 **H**	75.21 **H**
RF	All	Imbalanced	88.39	94.03	91.12 **H**	65.29	47.65	55.09	85.18 **H**
		Balanced	76.98	72.09	74.46	73.76	78.44	76.03 **H**	75.27
	Top-4	Imbalanced	89.01	94.45	91.45	68.22	50.54	58.06	86.07
		Balanced	75.17	71.90	73.50	73.07	76.25	74.62	74.07
ANN	All	Imbalanced	83.01	97.45	89.65	58.79	15.40	24.41	81.80
		Balanced	73.54	79.41	76.36 **H**	77.62	71.42	74.39	75.42 **H**
	Top-4	Imbalanced	76.42	89.80	82.57	68.61	44.60	54.06	74.73
		Balanced	70.37	75.79	72.98	73.77	68.08	70.81	71.94
KNN	All	Imbalanced	76.11	88.77	81.95	66.33	44.27	53.10	73.93
		Balanced	66.87	74.71	70.57	71.35	62.99	66.91	68.85
	Top-4	Imbalanced	76.14	88.74	81.96	66.34	44.37	53.17	73.95
		Balanced	66.43	73.61	69.74	70.41	62.80	66.39	68.21
GNB	All	Imbalanced	82.52	95.19	88.40	41.51	14.49	21.48	79.79
		Balanced	55.73	90.60	69.01	74.88	28.01	40.77	59.31
	Top-4	Imbalanced	82.59	95.97	88.78	45.36	14.20	21.63	80.37
		Balanced	56.62	89.59	69.39	75.07	31.35	44.23	60.47

決定木は質問を分類するための良いモデルを生成できた

まとめ

- Stack Overflowでの主観的評価と客観的な品質評価の一致を調査
  - トピックエントロピーなどは主観的評価と一致した
  - テキストの読みやすさやコードの再利用性は主観的評価と一致しなかった

- 高評価な質問と低評価な質問は品質評価で分類できるか
  - 決定木，ランダムフォレスト，ニューラルネットワーク，k近傍法，ナイーブベイズを使用
  - 最大で76~87%の精度で質問を分類することができた.