

Where is Your App Frustrating Users?

用户在抱怨App的哪些问题?

Yawen Wang, Junjie Wang, Hongyu Zhang, Xuran Ming, Lin Shi, Qing Wang

In 44th International Conference on Software Engineering (ICSE 2022)

联系人: 王亚文, 王俊杰, 王青

联系方式: {yawen2018, junjie, wq}@iscas.ac.cn

Motivation

User Review

- WHAT aspects: The high-level topics/aspects of the reviews, e.g., GUI, compatibility, etc.
- WHERE aspects: The specific App features the users complain about.
- Other aspects: Simple praises, complaints and trivial information.

Problematic Feature

Phrases buried in App reviews, which reflect users' complaints about certain App features.

Instagram ☆☆☆☆ SharonStreger 04/19/2020
Can't **upload to my story**, keeps crashing screen goes Black, Samsung 6s... I have tried uninstalling updating clearing data clearing cache, this is very annoying that no answers are popping up in this app. Not very supportive!

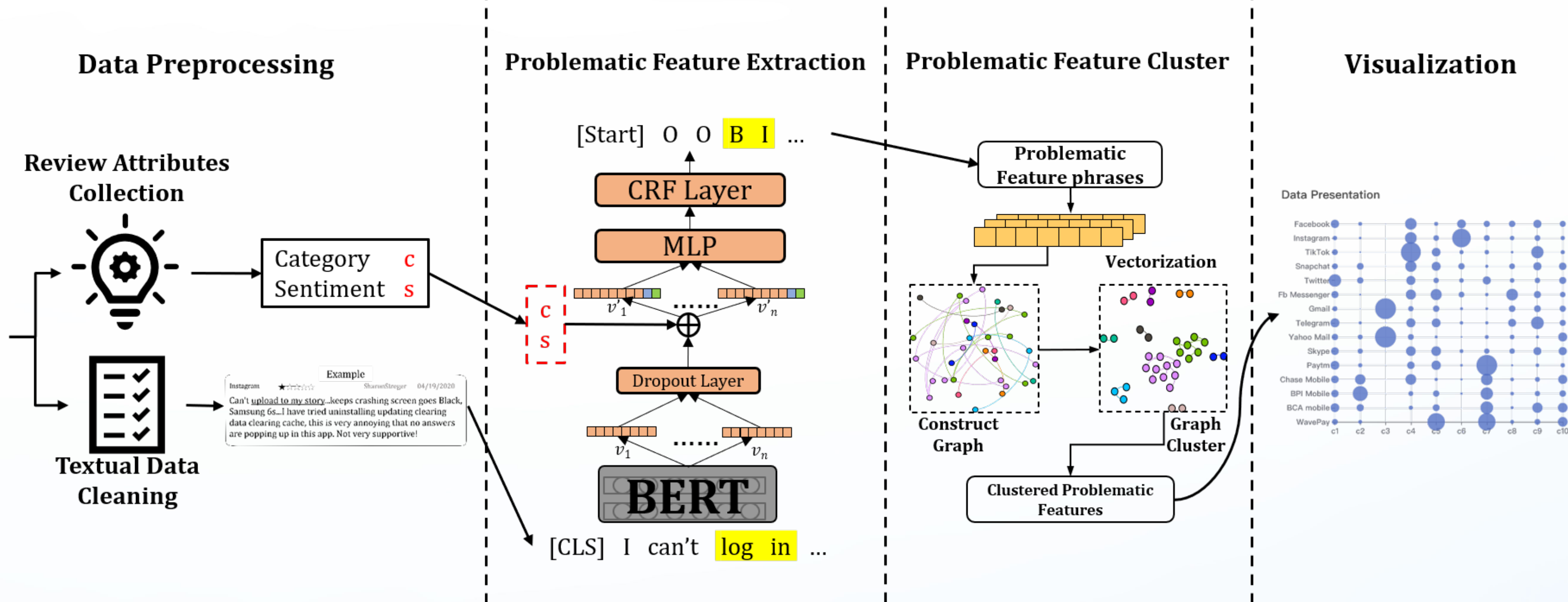
Gmail ☆☆☆☆ Mitch Devivo 01/05/2020
This app gets worse instead of better with each update. Now, suddenly, I'm not **getting email notifications**, even though I haven't changed any settings. Google acting like a near defunct start-up. Shameful.

Snapchat ☆☆☆☆ Danielle McMahon 12/25/2019
Stop updating it. You're making it 50x slower, **opening snaps or chats** is near impossible. You're not fixing it, you're making it worse. I'm having to uninstall and reinstall a few times a day to make it work alright for a few minutes.

- upload to my story
- getting email notification
- opening snaps or chats

Approach

SIRA: A Semantic-aware, fine-grained App Review Analysis approach to extract, cluster, and visualize the problematic features of Apps.



Data Preprocessing

- Textual Data Cleaning
 - Sentence Splitting & Non-English Filtering
 - Lowercase & Lemmatization & Formatting
- Review Attributes Collection
 - App Category
 - Review Sentiment

Problematic Feature Extraction

- Named Entity Recognition (NER) task
- BERT encoding review descriptions
- Incorporating review attributes

Problematic Feature Cluster

- Problematic feature vectorization
- Graph constructing
- Graph clustering

Visualization

- y-axis: App name
- x-axis: cluster id
- The size of the bubble

Evaluation

- RQ1: Performance on problematic feature extraction
- RQ2: Ablation experiment on review attributes
- RQ3: Performance on problematic feature cluster

RQ1

Metric \ Method		KEFE	Caspar	SAFE	BiLSTM-CRF	SIRA
App	P	40.32%	16.26%	14.17%	80.24%	83.59%
	R	60.76%	10.49%	70.61%	71.79%	85.70%
	F1	48.29%	12.46%	23.55%	75.58%	84.53%
Instagram	P	42.08%	18.87%	12.95%	78.49%	82.63%
	R	58.71%	13.81%	65.60%	74.71%	84.15%
	F1	48.70%	15.74%	21.59%	76.47%	83.30%
Snapchat	P	53.79%	25.60%	22.25%	87.58%	90.27%
	R	78.54%	9.88%	88.21%	71.74%	84.16%
	F1	63.46%	14.12%	35.49%	78.81%	87.09%
Gmail	P	12.57%	18.26%	12.57%	74.45%	79.18%
	R	70.10%	11.85%	70.10%	74.69%	87.37%
	F1	21.25%	14.19%	21.25%	74.26%	83.00%
Yahoo Mail	P	41.92%	20.98%	18.22%	82.58%	87.37%
	R	62.75%	9.24%	77.05%	73.53%	85.07%
	F1	50.13%	12.51%	29.44%	77.63%	86.13%
BPI Mobile	P	36.98%	17.53%	12.17%	77.23%	80.32%
	R	52.85%	13.38%	64.85%	68.43%	84.59%
	F1	43.16%	15.03%	20.44%	72.31%	82.26%
Chase Mobile	P	42.79%*	19.14%*	15.51%*	80.40%	84.27%
	R	63.50%*	11.27%*	73.94%**	72.48%*	85.06%
	F1	51.05%*	14.13%*	25.62%*	76.15%*	84.64%

Compared to SIRA, statistical significance p -value < 0.05 is denoted by **, and p -value < 0.01 is denoted by *.

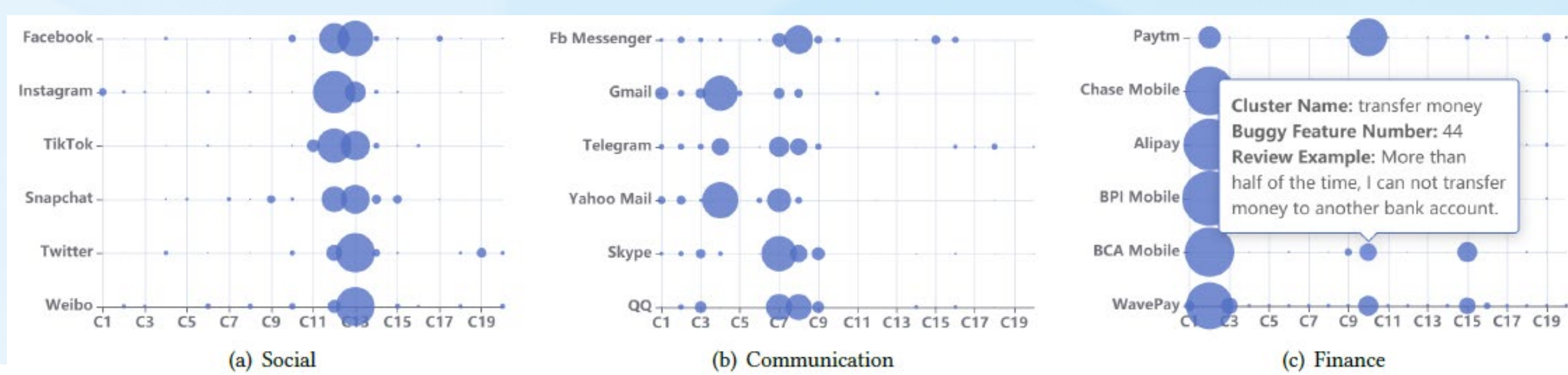
RQ2

Metric \ Method		BERT -CRF	BERT +CAT -CRF	BERT +SEN -CRF	BERT +Attr -CRF
App	P	82.46%	84.08%	83.78%	83.59%
	R	80.39%	85.60%	85.50%	85.70%
	F1	81.34%	84.73%	84.56%	84.53%
Instagram	P	84.58%	83.82%	83.38%	82.63%
	R	81.49%	83.31%	85.31%	84.15%
	F1	82.89%	83.48%	84.23%	83.30%
Snapchat	P	88.33%	89.30%	90.59%	90.27%
	R	78.37%	83.43%	83.50%	84.16%
	F1	82.99%	86.16%	86.86%	87.09%
Gmail	P	75.92%	76.67%	78.23%	79.18%
	R	83.72%	83.72%	86.09%	87.37%
	F1	79.54%	79.94%	81.86%	83.00%
Yahoo Mail	P	84.87%	85.92%	85.52%	87.37%
	R	78.09%	84.94%	82.60%	85.07%
	F1	81.25%	85.32%	83.96%	86.13%
BPI Mobile	P	78.24%	80.26%	80.05%	80.32%
	R	77.59%	82.19%	83.74%	84.59%
	F1	77.73%	81.11%	81.76%	82.26%
Chase Mobile	P	82.59%	83.73%	83.95%	84.27%
	R	79.69%	83.88%*	84.31%*	85.06%*
	F1	81.10%	83.78%**	84.10%**	84.64%*

Compared to BERT-CRF, statistical significance p -value < 0.05 is denoted by **, and p -value < 0.01 is denoted by *.

RQ3

Metric \ Method		LDA	K-Means	SIRA
App	ARI	0.10	0.30	0.29
	NMI	0.72	0.78	0.84
Instagram	ARI	0.19	0.13	0.32
	NMI	0.80	0.72	0.85
Snapchat	ARI	0.18	0.07	0.45
	NMI	0.73	0.58	0.82
Gmail	ARI	0.42	0.47	0.41
	NMI	0.81	0.83	0.82
Yahoo Mail	ARI	0.44	0.10	0.59
	NMI	0.83	0.58	0.89
BPI Mobile	ARI	0.38	0.21	0.26
	NMI	0.81	0.79	0.82
Overall	ARI	0.21	0.14	0.38
	NMI	0.57	0.62	0.77



Visualization