



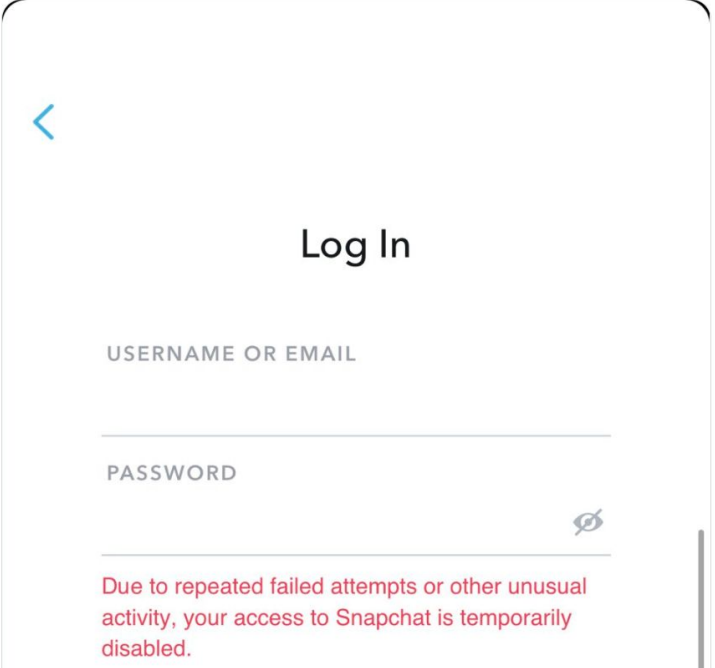
THE UNIVERSITY OF
AUCKLAND
Te Whare Wananga o Tamaki Makaurau
NEW ZEALAND

What's Inside a Cluster of Software User Feedback: A Study of Characterisation Methods

Peter Devine, James Tizard, Hechen Wang, Yun Sing Koh, Kelly Blincoe
The University of Auckland

One unhappy user...

Yo @snapchatsupport, can you fix your app? Literally just changed my password and there haven't been any failed login attempts. The f??



The screenshot shows a login interface with a blue back arrow in the top left corner. The title "Log In" is centered. Below it are two input fields: "USERNAME OR EMAIL" and "PASSWORD". The password field has a toggle icon on the right. At the bottom, a red error message reads: "Due to repeated failed attempts or other unusual activity, your access to Snapchat is temporarily disabled."

Or a bigger problem?

Is anyones snapchat permanently locked for no reason
!?!? @snapchatsupport @Snapchat FIX YOUR APP !!!!!
OMG

Snapchat is down like yesterday for me and some of
my friends @snapchatsupport

Why am I disabled from my Snapchat?
@snapchatsupport

cant log into my account, it says unusual activity and
failed attempts but i done nothing 😞
@snapchatsupport @Snapchat

Finding topics from user feedback

- Tools available (e.g. MERIT[1] and CLAP[2]) for multiple platforms
- Collect feedback into semantically coherent groups
- What is the best way to characterize these groups?



[1,2]



[1]



[3]

[1] Gao C, Zeng J, Wen Z, Lo D, Xia X, King I, Lyu MR. *Emerging app issue identification via online joint sentiment-topic tracing*. IEEE Transactions on Software Engineering. 2021 Apr 28.

[2] Scalabrino S, Bavota G, Russo B, Di Penta M, Oliveto R. *Listening to the crowd for the release planning of mobile apps*. IEEE Transactions on Software Engineering. 2017 Oct 4;45(1):68-86.

[3] Stanik C, Pietz T, Maalej W. *Unsupervised Topic Discovery in User Comments*. In: 2021 IEEE 29th International Requirements Engineering Conference (RE) 2021 Sep 20 (pp. 150-161). IEEE.

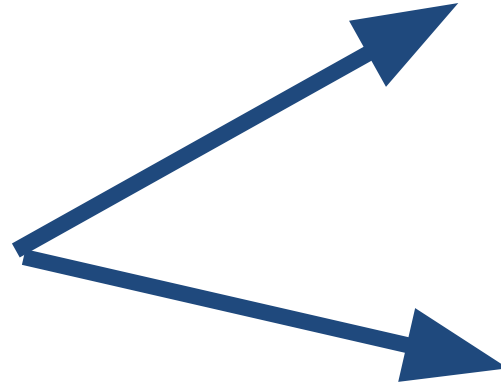
Ways to represent these tweets

Is anyones snapchat permanently locked for no reason
!?!? @snapchatsupport @Snapchat FIX YOUR APP !!!!!
OMG

Snapchat is down like yesterday for me and some of
my friends @snapchatsupport

Why am I disabled from my Snapchat?
@snapchatsupport

cant log into my account, it says unusual activity and
failed attempts but i done nothing 😞
@snapchatsupport @Snapchat



Uni-grams:

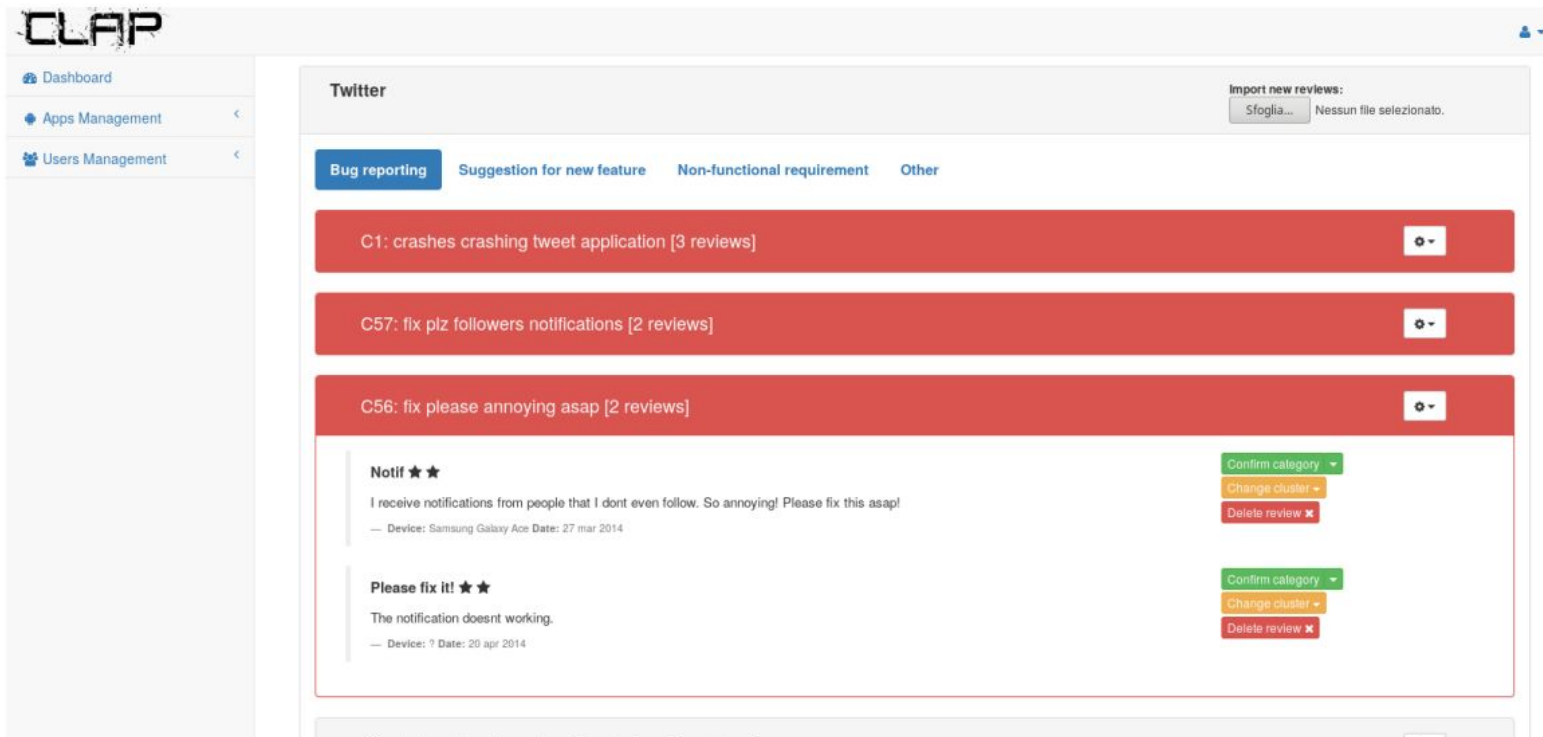
Account, locked,
unable, login

Sentence:

Is anyones snapchat
permanently locked
for no reason!?!?!?

Characterization examples

Scalabrino S, Bavota G, Russo B, Di Penta M, Oliveto R. *Listening to the crowd for the release planning of mobile apps*. IEEE Transactions on Software Engineering. 2017 Oct 4;45(1):68-86.



The screenshot displays the CLAP interface. On the left is a navigation sidebar with 'Dashboard', 'Apps Management', and 'Users Management'. The main content area is titled 'Twitter' and features a filter bar with 'Bug reporting' (selected), 'Suggestion for new feature', 'Non-functional requirement', and 'Other'. Below the filter bar, there are three red bars representing clusters of reviews:

- C1: crashes crashing tweet application [3 reviews]
- C57: fix plz followers notifications [2 reviews]
- C56: fix please annoying asap [2 reviews]

The detailed view for cluster C56 shows two reviews:

- Notif ★★**
I receive notifications from people that I dont even follow. So annoying! Please fix this asap!
— Device: Samsung Galaxy Ace Date: 27 mar 2014
- Please fix it! ★★**
The notification doesnt working.
— Device: ? Date: 20 apr 2014

For each review, there are three action buttons: 'Confirm category' (green), 'Change cluster' (orange), and 'Delete review' (red).

Characterization examples

Stanik C, Pietz T, Maalej W. *Unsupervised Topic Discovery in User Comments*. In 2021 IEEE 29th International Requirements Engineering Conference (RE) 2021 Sep 20 (pp. 150-161). IEEE.

Twitter Accounts: FitbitSupport PolarGlobal From: To:

Rising Trends

@FitbitSupport No need to be snarky. I'm just a customer expressing displeasure at a product
60 +7 (13%) ▲

@PolarGlobal Hi, I've attempted a reset to no avail. It's been eating batteries before then. Can I arrange a return to retail outlet through you? Thanks.
11 +5 (83%) ▲

@FitbitSupport Still vibrates when screen is blank. Screen doesn't turn on when plugged in to charge. Yes I have tried restarting that way.
6 +4 (200%) ▲

Falling Trends

@FitbitSupport please can you respond to my message
4 -12 (75%) ▼

@FitbitSupport I woke up to my Fitbit app updated ... I HATE THIS UPDATE 🤬 Not user friendly, waaaaaay too many changes all over the place, it's just a plain watch now 🤬🤬🤬!!
23 -6 (21%) ▼

@FitbitSupport My Fitbit skipped yesterday and now it's gone weird.
<https://t.co/PFW0rSjICB>
52 -5 (9%) ▼

Problem Reports

search

↓ Date	Tweet	Label required
09.02.2020	@FitbitSupport Hi so I am trying to reach you with a DM... don't want to explain the wifi problem im having with a brand new versa 2 on the public wall, would you be so kind as to respond so I know I might get some help?	
09.10.2019	@FitbitSupport hi i still havnt received an email to confirm you have my fitbit i sent back to you? Can someone please email me because im getting worried your not going to send me my replacement!	<input checked="" type="checkbox"/> <input type="checkbox"/>
	@FitbitSupport Hi, I've been sending DMs to you since Monday afternoon (one response). I have an issue with my Fitbit charge 3 -	

Challenges

- A cluster must be understandable (*characterized*) at a glance
- A cluster characterization should be both **descriptive** of the cluster content and **distinct** from other clusters' characterizations

Approach

- We study uni-, bi-, tri-grams and sentences as cluster representations.

Distinct

- **RQ1:** What characterization method is most easily able to distinguish a cluster compared to a random baseline?

Descriptive

- **RQ2:** What characterization method has the highest amount of requirements relevant information about a cluster?

Evaluation datasets

- 6 literature datasets + supplementary data

Dataset ID	Citation	Feedback source	Downloaded data size	Post-cleaning size	Number of apps	Number of apps with clusters	Number of clusters
A	[1]	App reviews	181,081	141,877	4	4	40
B	[2]	App reviews	2,325	1,415	27	3	8
C	[3]	App reviews	6,159	3,863	7	7	28
D	[4]	App reviews	3,000	2,867	694	1	2
E	[5]	Forums	3,806	3,293	3	3	16
F	[6]	Twitter	4,000	3,795	10	10	24
G	Supplementary data for Spotify Twitter	Twitter	6,115	1,653	1	1	10
H	Supplementary data for Spotify Apple App Store	App reviews	5,670	2,889	1	1	10
I	Supplementary data for Spotify Play Store	App reviews	24,346	16,394	1	1	10
J	Supplementary data for Spotify Reddit	Reddit	1,536	840	1	1	4
K	Supplementary data for Spotify Forum	Forums	1,061	1,032	1	1	8

[1] Chen N, Lin J, Hoi SC, Xiao X, Zhang B. *AR-miner: mining informative reviews for developers from mobile app marketplace*. In Proceedings of the 36th international conference on software engineering 2014 May 31 (pp. 767-778).

[2] Ciurumelea A, Schaufelbühl A, Panichella S, Gall HC. *Analyzing reviews and code of mobile apps for better release planning*. In 2017 IEEE 24th International Conference on Software Analysis, Evolution and Reengineering (SANER) 2017 Feb 20 (pp. 91-102). IEEE.

[3] Guzman E, El-Haliby M, Bruegge B. *Ensemble methods for app review classification: An approach for software evolution (n)*. In 2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE) 2015 Nov 9 (pp. 771-776). IEEE.

[4] Scalabrino S, Bavota G, Russo B, Di Penta M, Oliveto R. *Listening to the crowd for the release planning of mobile apps*. IEEE Transactions on Software Engineering. 2017 Oct 4;45(1):68-86.

[5] Tizard J, Wang H, Yohannes L, Blincoe K. *Can a conversation paint a picture? mining requirements in software forums*. In 2019 IEEE 27th International Requirements Engineering Conference (RE) 2019 Sep 23 (pp. 17-27). IEEE.

[6] Williams G, Mahmoud A. *Mining twitter feeds for software user requirements*. In 2017 IEEE 25th International Requirements Engineering Conference (RE) 2017 Sep 4 (pp. 1-10). IEEE.

Clustering

Embed feedback using Universal Sentence Encoder (USE) [1]



Cluster embeddings using UMAP[2] + HDBSCAN[3]



Randomly sample 1 piece of feedback from each cluster

[1] Cer D, Yang Y, Kong SY, Hua N, Limtiaco N, John RS, Constant N, Guajardo-Cespedes M, Yuan S, Tar C, Sung YH. *Universal sentence encoder*. arXiv preprint arXiv:1803.11175. 2018 Mar 29. **11**

[2] McInnes L, Healy J, Melville J. *Umap: Uniform manifold approximation and projection for dimension reduction*. arXiv preprint arXiv:1802.03426. 2018 Feb 9.

[3] Campello RJ, Moulavi D, Sander J. *Density-based clustering based on hierarchical density estimates*. In Pacific-Asia conference on knowledge discovery and data mining 2013 Apr 14 (pp. 160-172). Springer, Berlin, Heidelberg.

Characterization

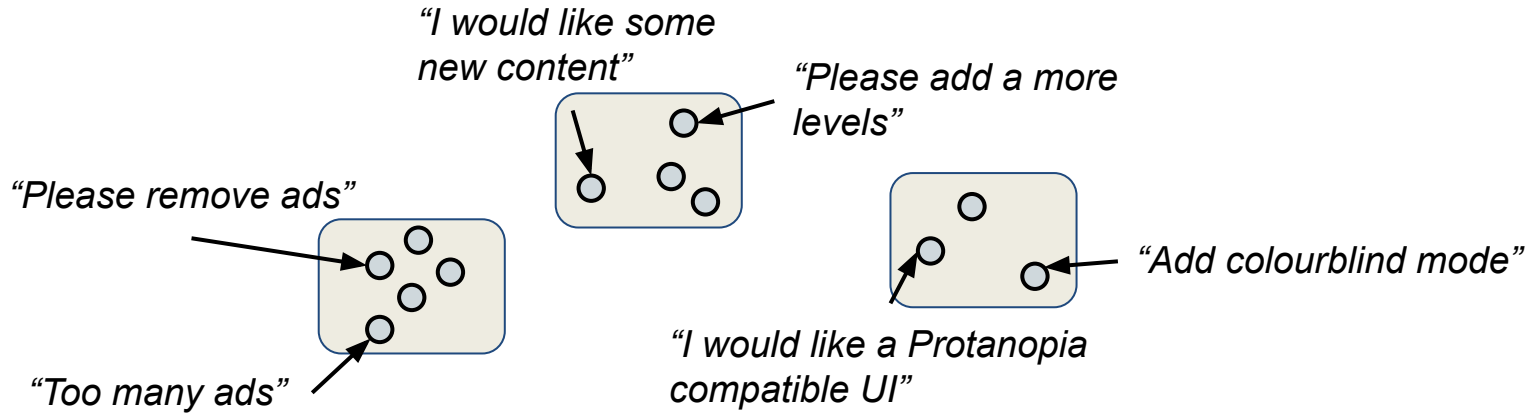
Find all uni-, bi-, tri-grams and sentences from feedback

Embed all n-grams and sentences

Find cosine similarity between n-gram/sentence embeddings and cluster embeddings

Choose k most similar n-grams/sentences for each cluster

Clustering



Example characterizations

Unigram characterization	login, offline, authentication, unable, connect, apk, password, reconnect, unavailable, unusable, unresponsive, connectivity
Bigram characterization	unable sign, app unresponsive, offline issue, account error, bug login, unable access
Trigram characterization	unable login app, unable login offline, unable login facebook, app bug login
Sentence characterization	Won't let me log in just keeps saying I have no Internet connection when every other app works fine?

Characterization labelling

- 2 tasks done by three labellers

Random characterization comparison task

Sample feedback	A	B
please add "move to sd card"	memory moved card app memory card phone memory moved memory movable card ✓	deserves star rating rating star gladly givit stars thx wil giv star

Requirements relevance labelling

Sample feedback	Correct characterization
please add "move to sd card"	memory moved card app memory card phone memory moved memory movable card
✓ Requirements relevant	✓ Requirements relevant

Outcomes from clustering and labelling

- Number of clusters:
 - 160 clusters taken from 33 apps across all datasets
- Labelling inter-coder reliability Cohens κ :
 - Random characterization comparison task - 0.677 (84.2% same label)
 - Requirements relevance labelling - 0.555 (78.3% same label)

Random characterization comparison

- Labelling accuracy (%) of per-coder, either coder, and both:

	Per-coder	Either coder	Both coder
Unigrams	0.838	0.900	0.775
Bigrams	0.853	0.906	0.800
Trigrams	0.847	0.944	0.750
Sentences	0.888	0.956	0.819

Requirements relevance labelling

	% Characterizations requirements relevant
Unigrams	6.98%
Bigrams	31.91%
Trigrams	29.55%
Sentences	45.65%

In a nutshell

- All characterization methods tested (uni, bi, tri-grams and sentences) were similarly distinguishable from other clusters
- However, even when number of words are kept constant, sentences are far superior at conveying requirements
- We find that longer, expressive single characterizations are more appropriate for requirements summarization than more but shorter ones

Implications

- Future feedback clustering tools will benefit from sentence level characterizations over unigram ones
- Opens questions for future research:
 - Does using extractive (e.g. HAHSum) or abstractive summarization models (e.g. PEGASUS) improve characterizations further?

Thanks

- Feel free to contact at:
 - pdev438@aucklanduni.ac.nz
 - [@p_d_research](https://twitter.com/p_d_research) on Twitter
- Questions?