Extracting User-Reported Mobile Application Defects from Online Reviews

Yue Wang
University of Delaware
U.S.A

Hongning Wang
University of Virginia
U.S.A

Hui Fang
University of Delaware
U.S.A
Defects identification is important for developers

- Know user need
- Fix the defects quickly

Developers
Verbose reviews are hard to digest in a short time

Fun but too many ads ★★★★★ by Azalard

This is a fantastic game for both children and adults. It's challenging, entertaining, and fairly easy to play. And unlike other games, you don't have to spend real money in order to purchase cheats or power-ups. You can progress on your own skill. It's an enjoyable way to pass the time on a road trip, a lazy afternoon, or when you just need a break and want something mindless to do.

The game does have a few issues. For me, it crashes quite frequently. It's a regular occurrence. The game also has these annoying ads and videos that pop-up at random. I realize that these ads are what gives us all this free content, but they often appear right when you're trying to move from puzzle to puzzle. It's like a kid running right out in the road while you're driving down the street. It's unpredictable and unexpected. What's worse is that the ads will still play sound even when you turn off all the music and sound in the app's settings. I don't mind watching the ads if I have to, but being forced to listen to them when I have my headset plugged in is a bit much.

Could we identify those defects automatically?
Existing solutions\[1\][2] focus on review level classification.

Defects reports

Game is fun. Too many ads.
Rating: ★★★★★
Posting date: Aug 21
Poster: Nikkie Z

Excellent app for monitoring your scr...
Rating: ★★★★★
Posting date: Sep 29
Poster: terrilee_nc

You can turn off the Emoji bar. I'm an...
Rating: ★★★★★
Posting date: Oct 24
Poster: LackAnAlibi

Updated from my earlier angsty review below.
You can turn off the emoji bar I was complaining about. I asked tech support, and they responded right away--it's under settings when you hold the Swype button. So back to five stars it goes! Love this app.

This was fab, and I was spreading the Swype gospel until this update. The emoji bar has to go or at least be optional. Who wanted this? Aside from the thing just being annoying to look at, I keep thinking I'll accidentally send a stupid emoji in a professional email, which is frustrating since email is why I got Swype in the first place. Please help us out, folks.

Non defects reports

Just One improvement idea
Rating: ★★★★☆
Posting date: Sep 15
Poster: Nikkie Z

Existing solutions focus on review level classification

Defects reports

<table>
<thead>
<tr>
<th>Title</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autocorrect baa!!</td>
<td>Oct 24</td>
</tr>
<tr>
<td>Just One improvement idea</td>
<td>Sep 15</td>
</tr>
<tr>
<td>Really fun but crashes to much (This i...</td>
<td>Oct 8</td>
</tr>
<tr>
<td>★★★★★☆</td>
<td>Pandu brah</td>
</tr>
</tbody>
</table>

Like it’s fun but it crashes like every 10 minutes
very annoying but it’s still fun there is not many
ads it has ads every like 20-30 minutes but it’s
fine they last less than 25 seconds

It’s great
and fun

Thanks for reading

Still time consuming and labor intensive!
Sentence level classification is necessary!
Contribution

• We developed a structured learning solution based on hidden structural support vector machines.

• The trained model utilize the dependency between sentences and reviews to predict the defects status at sentence level.

• The proposed method could be trained with partially annotated data.
Problem formulation

• Sentence level defects reporting classification
  – Given a user-generated review $R = \{s_1; s_2; \ldots; s_n\}$, consisting of $n$ sentences, the sentence level defects classification generate a binary label for each sentence $s_i = 1$, indicating this sentence is reporting a defect, otherwise $s_i = 0$.

S1) All emails from my director won’t open. S2) Gives me an error message. S3) Also, it would be nice if you could have a setting for what language alerts you receive. S4) Although it is nice to have 10 new alerts on my phone all at once, I don’t need the same one in 3 languages. S5) And since the newest update, I can’t see anything under the my business tab. S6) It's just white.
Problem formulation

• Defects:
  – The abnormal behavior of an app which produces an incorrect or unexpected result, or behaves in unintended ways.

Functional suggestion

Game is fun. Too many ads.
★★★☆☆
Aug 21
Musehobo

This is a solid puzzle-based game. Unlike the original, there is a sufficient amount of levels to play. It works well for either me or my five year old. This sequel also offers variations of stages to play, providing different types of challenges.

Negatives: Ads are too common, too long, and unpredictable. Current game version crashes VERY often. The game is just enjoyable enough to overcome these frustrations.
A straightforward solution

- Review level classifier on sentence level classification

Unsatisfying performance due to the lack of connections between sentence and reviews
Sentence level defects classification with latent structures (hSVM)

\[
\begin{align*}
(\hat{y}, \hat{H}) &= \arg \max_{(y, H) \in \mathcal{Y} \times \mathcal{H}} \omega^T \Phi(R, H, y) \\
\end{align*}
\]
The constraints between sentences and review

\[ \text{if } \hat{y} = 1 \implies \sum_{\hat{h} \in \hat{H}} \neq 0 \]

Defect reporting review must contain at least one defect reporting sentence

\[ \text{if } \hat{y} = 0 \implies \sum_{\hat{h} \in \hat{H}} = 0 \]

Non defect reporting review cannot contain any defect reporting sentence

\[ \sum_{i=1}^{n} \hat{h}_i \geq \hat{y} \]
Model training

- Estimate $\omega$ with structured SVM framework\cite{3}

\[
\min_{\omega, \xi} \frac{1}{2} \|\omega\|^2 + C \sum_{m=1}^{M} \xi_m \\
\text{s.t. } \forall m, \max_{H \in \mathcal{H}} \omega^T \Phi(R_m, H, y_m) \geq \max_{(\hat{y}, \hat{H}) \in \mathcal{Y} \times \mathcal{H}} \omega^T \Phi(R_m, \hat{H}, \hat{y}) + \Delta(y_m, \hat{y}, H_m, \hat{H}) - \xi_m \\
\xi_m \geq 0
\]

Measure error between prediction and ground-truth

\[
\sum_{i=1}^{n} \hat{h}_i \geq \hat{y}
\]

Sentence level features

- Aspects
  - User usually mention one or more specific aspects of the app if it is defective.

Trained PLSA topic model to recognize the aspects from the reviews
Sentence level features

• Indicator words
  – A list of words that shared by the users when they tend to report the defects.
  – *Freeze, crush, idle, etc.*

```
Autocorrect bug!!! Oct 24
★★★★★
RoBuJr

Could you please fix the bug where your words get autocorrected when you press the "punctuation" key. Even with the autocorrect feature turned off, it still decides to switch what your typed to some random word that may begin with the first letter if your word. This has started to get very annoying because of this and the very long time it's been since this problem started or addressed. People have been
```

```
Game is fun. Too many ads. Aug 21
★★★★☆☆
Musehobo

This is a solid puzzle-based game. Unlike the original, there is a sufficient amount of levels to play. It works well for either me or my five year old. This sequel also offers variations of stages to play, providing different types of challenges.

Negatives: Ads are too common, too long, and unpredictable. Current game version crashes VERY often. The game is just enjoyable enough to overcome these frustrations.
```

*Created an indicator word list based on the training data*
Sentence level features

• Similarity with sentences in the other reviews
  – Users would report the same defects in a similar way

Autocorrect bug!!!  Oct 24  RoBuJr
★★★★★
Could you please fix the bug where your words get autocorrected when you press the "punctuation" key. Even with the autocorrect feature turned off, it still decides to switch what your typed to some random word that may begin with the first letter if your word. This has started to get very annoying because of this and the very long time it's been since this problem started or addressed. People have been complaining about this problem for years! I'm having to give just 1 star.

Used to be a lot better  Oct 26  Ericdadamon
★★★★★
I'm not sure what happened, but the auto predict and autocorrect has been terrible last year or so... It seems like they added some predictive neural net layer that is God awful. And app assumes too often that I'm trying to type some random name that I am definitely not.

Example - re swyping the above without fixing errors:

I'm not due Wendy happened, but the situ

Highest sentence to review similarity is used as feature
Review level features

• Similarity with the other reviews

**Autocorrect bug!!!**

Could you please fix the bug where your words get autocorrected when you press the "punctuation" key. Even with the autocorrect feature turned off, it still decides to switch what your typed to some random word that may begin with the first letter if your word. This has started to get very annoying because of this and the very long time it's been since this problem started or addressed. People have been complaining about this problem for years! I'm having to give just 1 star.

**Used to be a lot better**

I'm not sure what happened, but the auto predict and autocorrect has been terrible last year or so... It seems like they added some predictive neural net layer that is God awful. And app assumes too often that I'm trying to type some random name that I am definitely not.

Example - re swyping the above without fixing errors:

I'm not due Wendy happened, but the situ
All features

• 14 sentence level features
  – Contains negation words?
  – Overall sentiment orientation
  – Sentence length
  – etc...

• 8 review level features
  – Aspect coverage
  – Aspect consistence
  – Indicator consistence
  – etc...
Data sets

- Apps crawled from Apple App Store
  - Fields being crawled: App name, description, reviews, etc
  - Fully labeled reviews and Partially labeled reviews

<table>
<thead>
<tr>
<th></th>
<th>Fully labeled reviews</th>
<th>Partially labeled reviews</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of reviews</td>
<td>274/572</td>
<td>400/818</td>
</tr>
<tr>
<td>(defect / non-defect)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of sentences</td>
<td>557/3917</td>
<td>84/441</td>
</tr>
<tr>
<td>(defect / non-defect)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unlabeled sentences</td>
<td>0</td>
<td>8968</td>
</tr>
</tbody>
</table>
Performance on Fully annotated reviews

- hSVM model outperformed most of the baselines

<table>
<thead>
<tr>
<th></th>
<th>Prec</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48-S</td>
<td>0.716</td>
<td>0.863</td>
<td>0.783</td>
</tr>
<tr>
<td>LR-S</td>
<td>0.799</td>
<td>0.811</td>
<td>0.805</td>
</tr>
<tr>
<td>SVM-S</td>
<td>0.785</td>
<td>0.832</td>
<td>0.808</td>
</tr>
<tr>
<td>RF-S</td>
<td>0.882</td>
<td>0.947</td>
<td>0.913</td>
</tr>
<tr>
<td>hSVM</td>
<td>0.919</td>
<td>0.898</td>
<td>0.908</td>
</tr>
</tbody>
</table>

Although Random Forest outperforms hSVM on recall, it requires much more training data.

Performance onFully annotated reviews

- hSVM model outperformed all the baselines on short sentences

<table>
<thead>
<tr>
<th>Model</th>
<th>Prec</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48-S</td>
<td>0.672</td>
<td>0.826</td>
<td>0.741</td>
</tr>
<tr>
<td>LR-S</td>
<td>0.761</td>
<td>0.792</td>
<td>0.776</td>
</tr>
<tr>
<td>SVM-S</td>
<td>0.763</td>
<td>0.814</td>
<td>0.781</td>
</tr>
<tr>
<td>RF-S</td>
<td>0.854</td>
<td>0.893</td>
<td>0.873</td>
</tr>
<tr>
<td>hSVM</td>
<td>0.912</td>
<td>0.931</td>
<td>0.921</td>
</tr>
</tbody>
</table>

The improvement of our method comes from utilizing the dependency between sentences and review to predict the sentence level labels.
Performance on Partially annotated data

- hSVM model could learn from the partially annotated data and less likely to overfit to a particular training set

- Quickly converge to optimal performance
  - (Require less training data)

- Smaller variance
  - (More robust)
The Microsoft Tag app lets you instantly connect to a whole new world of information and entertainment. This latest version now reads Microsoft Tag barcodes and QR Codes. No need to type long URLs or text short codes; simply scan a 2D barcode wherever you see one. Tag recognition technology transforms traditional marketing, as well as view content in an embedded browser and share items via email, Facebook and Twitter. Plus, it allows you to save and edit contact information.

**69 sentences mentioning “QR code”:**
Close when read QR code.
Can’t read the code because it closes every time!

**137 sentences mentioning “scan”:**
This crashes before I can scan, it has a bug please fix.
It crashes after I tapped the scan button

**2 sentences mentions “save”:**
I can not save the information from the link!
The save and edit is a crap!
Conclusion and Future Work

• Conclusion
  – We proposed a structured learning solution to address the problem of identifying sentence-level defects reports.
  – The proposed method utilize the dependency between sentences and reviews to better predict the sentence level label

• Future work
  – Study how to include unlabeled data for training
  – Summarize the extracted sentences and map to aspects
  – Investigate the usage of update history
Thank you!
Q&A

Corresponding Author
Yue Wang (wangyue@udel.edu)