

# 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



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 even when you turn off all the music and sound in the app. *Could we identify those defects automatically?* Then to them when I have my headset plugged in is a bit much.

# Existing solutions<sup>[1][2]</sup> focus on review level classification

## Defects reports

Just One improvement idea



Sep 15

Nikkie Z

Game is fun. Too many ads.



Aug 21

Excellent app for monitoring your scr...



Sep 29

terrilee\_nc

You can turn off the Emoji bar. I'm an...



Oct 24

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

[1] N. Chen, J. Lin, S. marketplace," in

[2] W. Maalej and H. RE 2015, 2015.

developers from mobile app

ifying app reviews," in IEEE

# Existing solutions focus on review level classification

## Defects reports

Autocorrect bua!!!

Oct 24

Just One improvement idea

Sep 15

Really fun but crashes to much (This i... Oct 8

★★★★☆

Pandu brah

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

if you go over your limit your bar is just the width of the screen so you don't see a visual of \*\*\*how far\*\*\* beyond the limit you are. I recommend make g the bars visually proportional to the hours you spend, making the scale from let's say 0-15 (for those who have a serious problem). I

***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  $R = \{s_1; s_2; \dots; s_n\}$ , consisting of  $n$  sentences, the sentence level defects classification generate a binary label for each sentence  $s_i = \mathbf{1}$ , indicating this sentence is reporting a defect, otherwise  $s_i = \mathbf{0}$ .

<sup>S1)</sup>All emails from my director won't open. <sup>S2)</sup>Gives me an error message. <sup>S3)</sup>Also, it would be nice if you could have a setting for what language alerts you receive. <sup>S4)</sup>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. <sup>S5)</sup>And since the newest update, I can't see anything under the my business tab. <sup>S6)</sup>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.

Game is fun. Too many ads.



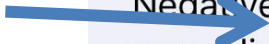
Aug 21

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.

*Functional suggestion*



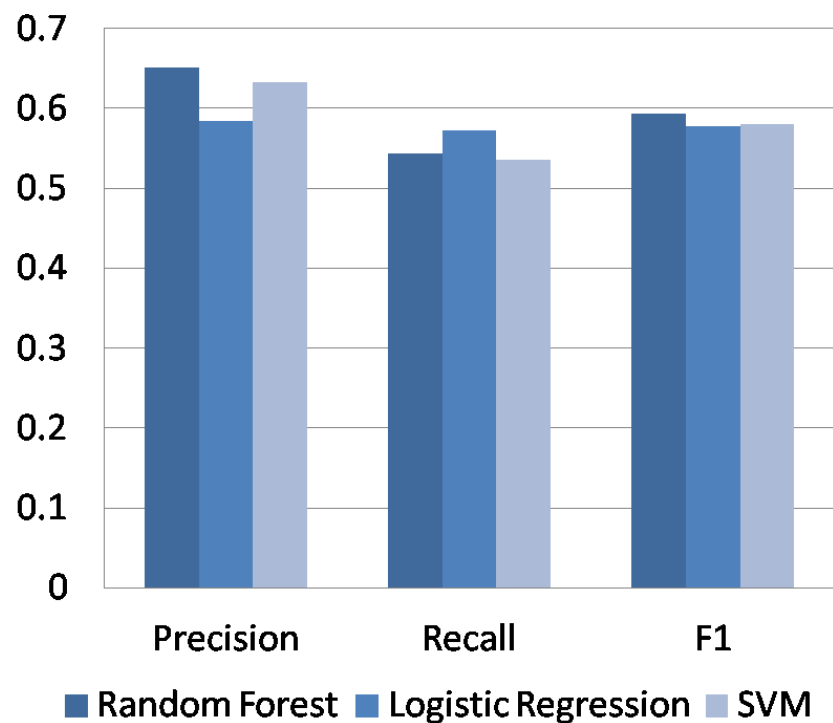
*Defects report*





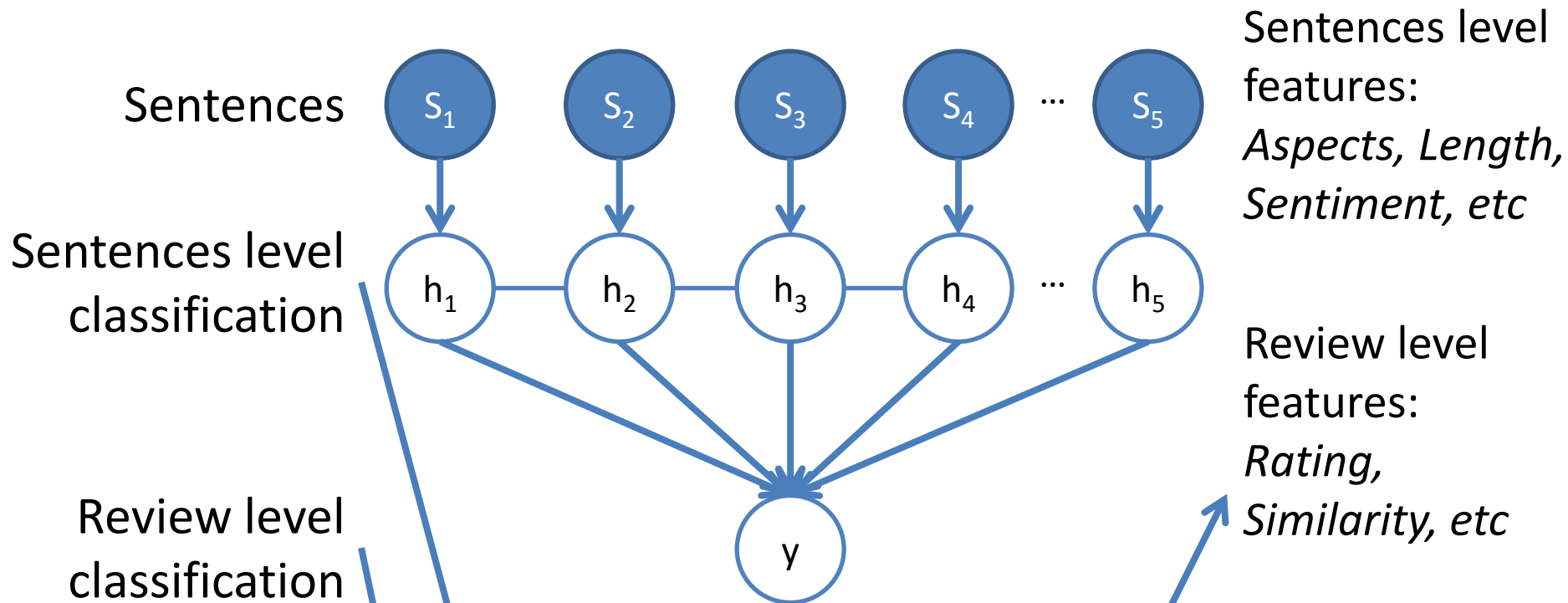
# A straightforward solution

- 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)



$$(\hat{y}, \hat{H}) = \arg \max_{(y, H) \in \mathcal{Y} \times \mathcal{H}} \omega^\top \Phi(R, H, y)$$

# The constraints between sentences and review

$$\text{if } \hat{y} = 1 \quad \rightarrow \quad \sum_{\hat{h} \in \hat{H}} \hat{h} \neq 0$$

Defect reporting review must contain at least one defect reporting sentence

$$\text{if } \hat{y} = 0 \quad \rightarrow \quad \sum_{\hat{h} \in \hat{H}} \hat{h} = 0$$

Non defect reporting review cannot contain any defect reporting sentence

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

# Model training

- Estimate  $w$  with structured SVM framework<sup>[3]</sup>

$$\begin{aligned} & \min_{\omega, \xi} \frac{1}{2} \|\omega\|^2 + C \sum_{m=1}^M \xi_m \\ \text{s.t. } & \forall m, \max_{H \in \mathcal{H}} \omega^\top \Phi(R_m, H, y_m) \geq \\ & \max_{(\hat{y}, \hat{H}) \in \mathcal{Y} \times \mathcal{H}} \omega^\top \Phi(R_m, \hat{H}, \hat{y}) + \underbrace{\Delta(y_m, \hat{y}, H_m, \hat{H})}_{\downarrow} - \xi_m \\ & \xi_m \geq 0 \end{aligned}$$

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.

<p><b>Just needs a few more fixes</b> ★★★★☆</p> <p>App is great for typing, but it CAI irritating when auto correct tries</p> <p>If I were to request any big change would say they need to revise the <u>correcting</u> system and get it more standard keyboard. The keyboard already know most of the words</p>	<p><b>Autocorrect bug!!!</b> ★★★★☆</p> <p>Could you please fix the bug where y get <u>autocorrected</u> when you press th "punctuation" key. Even with the autoc feature turned off, it still decides to s your typed to some random word tha with the first letter if your word. This to get very annoying because of this very long time it's been since this prc started or addressed. People have be</p>	<p><b>Used to be a lot better</b> ★★★★☆</p> <p>Oct 26 Ericdadamson</p> <p>I'm not sure what happened, but the auto predict and <u>autocorrect</u> 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.</p> <p>Example - re swyping the above without</p>
--	--	--

*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 you 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

Ericdadamson

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!!!

★★★★★

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

Ericdadamson

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 review to review similarity is used as feature*

# 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

	Fully labeled reviews	Partially labeled reviews
Number of <b>reviews</b> (defect / non-defect)	274/572	400/818
Number of <b>sentences</b> (defect / non-defect)	557/3917	84/441
Unlabeled sentences	0	8968

# Performance on Fully annotated reviews

- hSVM model outperformed most of the baselines

	Prec	Recall	F1
J48-S <sup>[4]</sup>	0.716	0.863	0.783
LR-S	0.799	0.811	0.805
SVM-S	0.785	0.832	0.808
RF-S	0.882	<b>0.947</b>	<b>0.913</b>
<b>hSVM</b>	<b>0.919</b>	0.898	0.908

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

# Performance on Fully annotated reviews

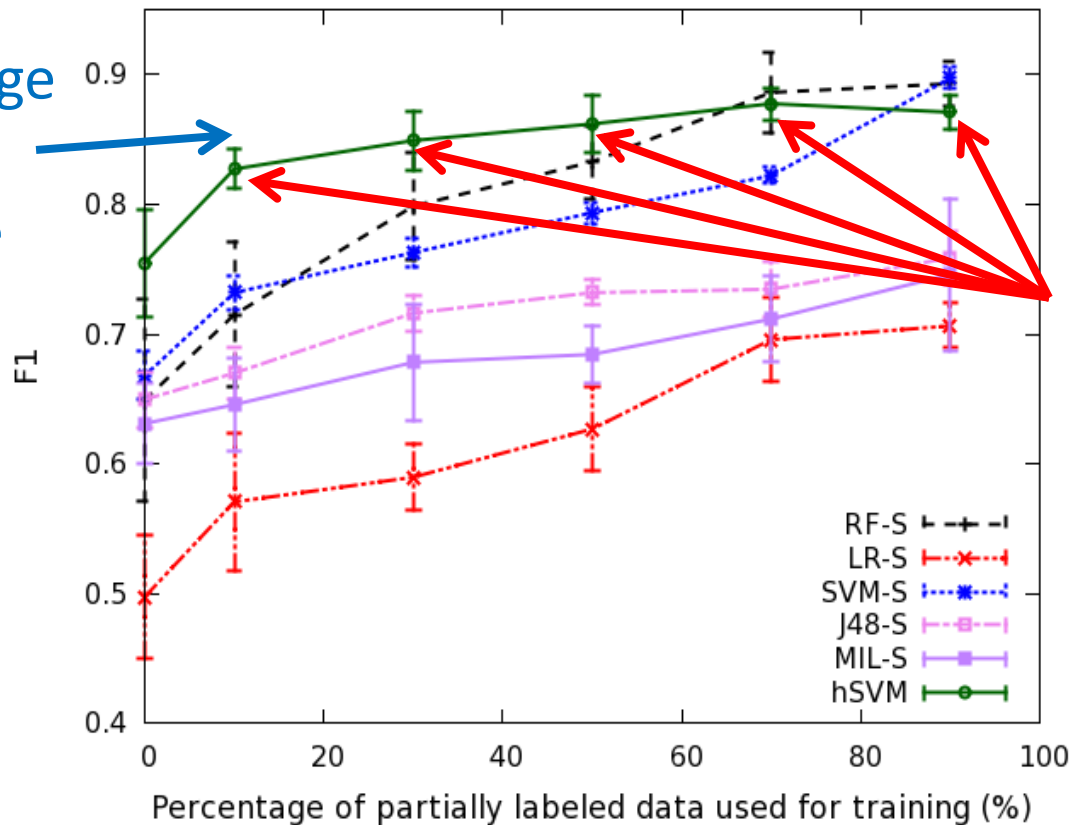
- hSVM model outperformed all the baselines on *short* sentences

	Prec	Recall	F1
J48-S	0.672	0.826	0.741
LR-S	0.761	0.792	0.776
SVM-S	0.763	0.814	0.781
RF-S	0.854	0.893	0.873
<b>hSVM</b>	<b>0.912</b>	<b>0.931</b>	<b>0.921</b>

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

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



Smaller  
variance  
(More robust)

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

Dec. 5<sup>th</sup>, 2011

Version 5.1  
released

24  
reviews

Dec. 20<sup>th</sup>, 2011

Version 5.5  
released

249  
reviews

Oct. 14<sup>th</sup>, 2012

Version 5.6  
released

29  
reviews

### Description of version 5.5

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.**

### Identified defect reporting sentences

#### 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)