HOW WHATSAPP REDUCED SPAM WHILE LAUNCHING END-TO-END ENCRYPTION
HOW WHATSAPP REDUCED SPAM WHILE LAUNCHING END-TO-END ENCRYPTION
Multiple reasons *not* to go E2E

- Spam & abuse
- Synching, searching
- Backwards compatibility
- “Smart assistant”
- Advertising?
Work from home & earn up to 1180 GBP Daily with 14Profit.com Program.
Learn more with the video at: www.14Profit.com

Reply back ‘OK’ to activate link and unlock the offer.
HOW WHATSAPP REDUCED SPAM
WHILE LAUNCHING END-TO-END ENCRYPTION

MATT JONES
HOW WHATSAPP REDUCED SPAM WHILE LAUNCHING E2E

SPAM ON WHATSAPP

- Unwanted
- Commercial
- High-volume
- Automated
- From Malicious Accounts

[Image showing a screenshot of a WhatsApp conversation with a spam message and options to block, report spam, or add to contacts]
HOW WHATSAPP REDUCED SPAM WHILE LAUNCHING E2E

SPAM ON WHATSAPP

- Unwanted
- Commercial
- High-volume
- Automated
- From Malicious Accounts
HOW WHATSAPP REDUCED SPAM WHILE LAUNCHING E2E

SPAM ON WHATSAPP

▸ Unwanted
▸ Commercial
▸ High-volume
▸ Automated
▸ From Malicious Accounts
HOW WHATSAPP REDUCED SPAM WHILE LAUNCHING E2E

SPAM ON WHATSAPP

▸ Unwanted
▸ Commercial
▸ High-volume
▸ Automated
▸ From Malicious Accounts
HOW WHATSAPP REDUCED SPAM WHILE LAUNCHING E2E

SPAM ON WHATSAPP

- Unwanted
- Commercial
- High-volume
- Automated
- From Malicious Accounts
HOW WHATSAPP REDUCED SPAM WHILE LAUNCHING E2E

REDUCE TO EASIER PROBLEM

- Spam?
  - Spam account? (No passwords = no ATO)
    - Automated? (High-volume = automated)
      - Spam client? (automated = script or emulator)

Faster
Less effective
HOW WHATSAPP REDUCED SPAM WHILE LAUNCHING E2E

GENERAL WORKFLOW

Action

user age: 3 hours
# messages 30sec: 2
...

Allow?

Ban?

Classify
HOW WHATSAPP REDUCED SPAM WHILE LAUNCHING E2E

GENERAL WORKFLOW

- Action
- Allow?
- Ban?
- Classify

- user age: 3 hours
- # messages 30sec: 2
- ...
HOW WHATSAPP REDUCED SPAM WHILE LAUNCHING E2E

TRAINING DATA

<table>
<thead>
<tr>
<th>Features</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>user age=5280 days, # messages 30sec=0...</td>
<td>Ham</td>
</tr>
<tr>
<td>user age=2 minutes, # messages 30sec=10...</td>
<td>Spam</td>
</tr>
<tr>
<td>user age=3 months, # messages 30sec=2...</td>
<td>Ham</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
## HOW WHATSAPP REDUCED SPAM WHILE LAUNCHING E2E

### TRAINING DATA

<table>
<thead>
<tr>
<th>Features</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>user age=5280 days, # messages 30sec=0...</td>
<td>???</td>
</tr>
<tr>
<td>user age=2 minutes, # messages 30sec=10...</td>
<td>???</td>
</tr>
<tr>
<td>user age=3 months, # messages 30sec=2...</td>
<td>???</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
HOW WHATSAPP REDUCED SPAM WHILE LAUNCHING E2E

TRAINING LABELS

- Need source of truth for labels in classifier training
- 2 mechanisms:
  - People we eventually caught
  - “Spam App” detection
TRAINING LABELS

- Need source of truth for labels in classifier training
- 2 mechanisms:
  - People we eventually caught
  - “Spam App” detection
HOW WHATSAPP REDUCED SPAM WHILE LAUNCHING E2E

TRAINING LABELS

- Need source of truth for labels in classifier training
- 2 mechanisms:
  - People we eventually caught
  - “Spam App” detection
The php WhatsApp library

- 679 commits
- 1 branch
- 14 releases
- 55 contributors
- GPL-3.0

mgp25 committed on GitHub Merge pull request #2132 from arnoldnlboer/master

- .github
- examples
- src
- tests
- .gitignore
- LICENSE
- README.md
- composer.json

Latest commit 9af443f on Sep 6

8 months ago
3 months ago
a month ago
3 months ago
a year ago
9 months ago
8 months ago

Chat API stable v3.2.0.1 downloads 17.68 k License GPL-3.0+

Interface to WhatsApp Messenger

Read the wiki and previous issues before opening a new one! Maybe your issue is already answered.

For new WhatsApp updates check WhatsApp incoming updates log
whatsapprus commented on Jul 13 • edited

Need a developer program (skript) to send whatsapp messages. Requirements: send image and text, sending more than 50 messages with a single hash channel, sending more than 50K messages per day, work on the local PC under ubuntu. My skype: ***

0xTryCatch closed this on Jul 13

0xTryCatch locked and limited conversation to collaborators on Jul 13

0xTryCatch added the Spam label on Jul 13
HOW WHATSAPP REDUCED SPAM WHILE LAUNCHING E2E

SPAM VENDORS STOPPED

ashishnetworks commented on Apr 28

Now it has become severe issue. No use of anything. Got frustrated! Nothing working. Getting banned every single time. We must do something about it and give this issue higher priority, otherwise there is no use of any feature if we keep getting banned.

adelsjunior commented on Aug 8

I can receive and send message, and EVEN synchronizing numbers, no day Next 'm blocked, someone KNOWS Why?
HOW WHATSAPP REDUCED SPAM WHILE LAUNCHING E2E

### TRAINING DATA

<table>
<thead>
<tr>
<th>Features</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>user age=5280 days, # messages 30sec=0...</td>
<td><strong>Ham</strong></td>
</tr>
<tr>
<td>user age=2 minutes, # messages 30sec=10...</td>
<td><strong>Spam</strong></td>
</tr>
<tr>
<td>user age=3 months, # messages 30sec=2...</td>
<td><strong>Ham</strong></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
### Training Data

<table>
<thead>
<tr>
<th>Features</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>???</td>
<td>Ham</td>
</tr>
<tr>
<td>???</td>
<td>Spam</td>
</tr>
<tr>
<td>???</td>
<td>Ham</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
BEHAVIORAL FEATURES

- How do spammers generate actions?
- How do others respond to those actions?
- Classify this behavior - no access to content is fine
FEATURES: REPUTATIONAL

- "How many people have been seen on your ASN? What % are banned?"
"How many people have been seen on your ASN? What % are banned?"

ASN:
20 total
12 banned
HOW WHATSAPP REDUCED SPAM WHILE LAUNCHING E2E

FEATURES: REPUTATIONAL

"How many people have been seen on your Netblock? What % are banned?"

ASN:
20 total
12 banned

Netblock:
8 total
5 banned
"How many people have been seen on your Phone Prefix? What % are banned?"

**ASN:**
- 20 total
- 12 banned

**Netblock:**
- 8 total
- 5 banned

**Phone Prefix:**
- 30 total
- 23 banned
How WhatsApp reduced spam while launching E2E

**Features: Reputational**

- "How many people have been seen on your ASN? What % are banned?"
- Generalize "ASN": "netblock", "phone number prefix", etc
- Generalize "banned": "reported as spam", etc
- Forces attacker to buy more things
HOW WHATSAPP REDUCED SPAM WHILE LAUNCHING E2E

FEATURES: PROBABILITY

▸ “Probability that someone with phone country code X is seen from network country Y”

▸ Generalize each (carrier, language, etc)
HOW WHATSAPP REDUCED SPAM WHILE LAUNCHING E2E

SPAM REPORTS DECREASED

Announced e2e

0.5x / day

0.25x / day
HOW WHATSAPP REDUCED SPAM WHILE LAUNCHING E2E

BAN APPEALS DECREASED

# Appeals

Time (3 months)
YOU DON’T NEED CONTENT FOR SPAM DETECTION

- 2 keys to success:
  - Good behavioral features
  - Good labeling: Spam Reports + Spam App detection