1337 SP34K D3F3473D —

Gmail’s AI-powered spam detection is its biggest security upgrade in years

Gmail's spam filters can now understand "adversarial text manipulations."

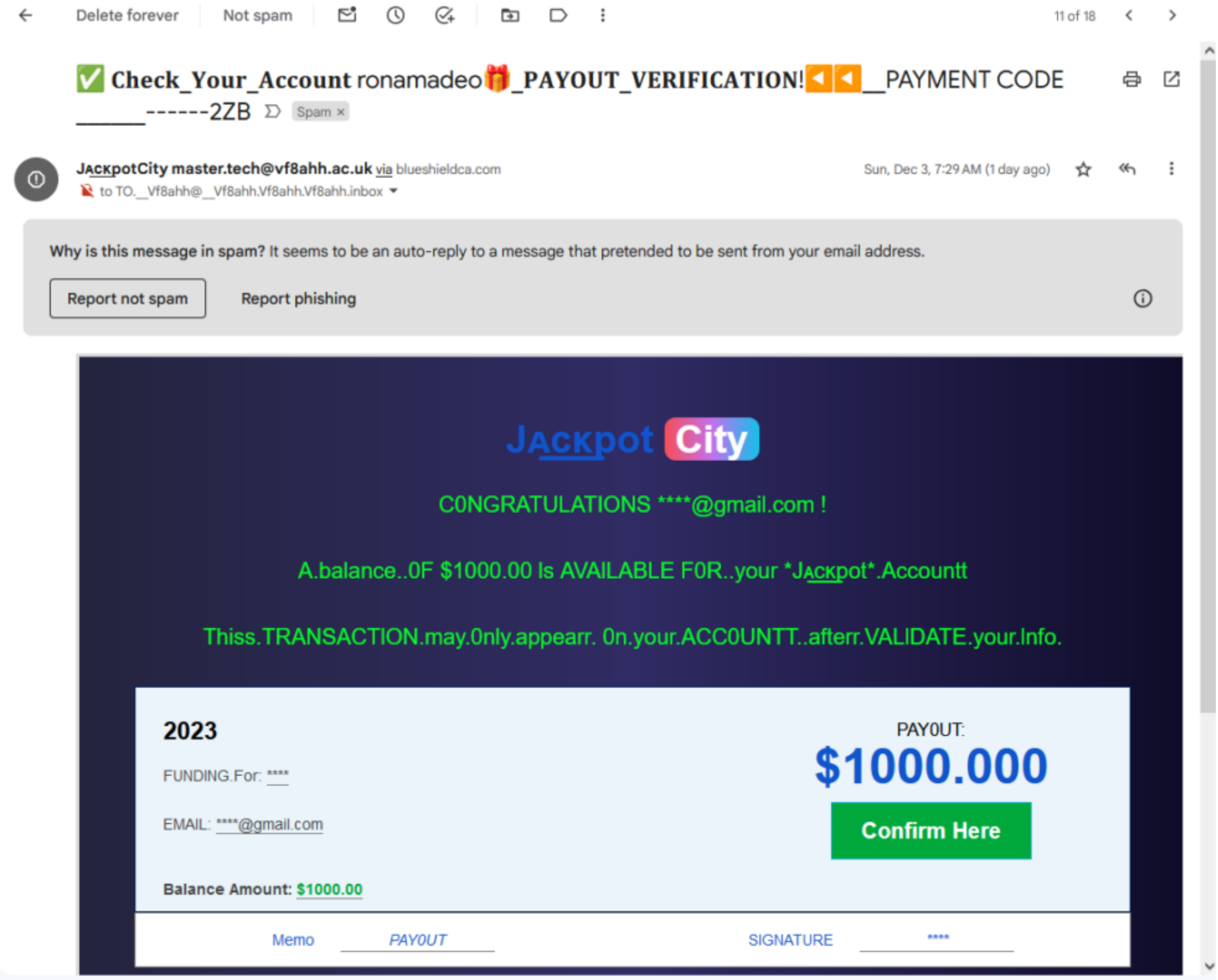
RON AMADEO - 12/4/2023, 11:04 AM

Getty Images | pagadesign

Enlarge

The latest post on the Google Security blog details a new upgrade to Gmail's spam filters that Google is calling "one of the largest defense upgrades in recent years." The upgrade comes in the form of a new text classification system called RETVec (Resilient & Efficient Text Vectorizer). Google says this can help understand "adversarial text manipulations"—these are emails full of special characters, emojis, typos, and other junk characters that previously were legible by humans but not easily understandable by machines. Previously, spam emails full of special characters made it through Gmail's defenses easily.

If you want an example of what "adversarial text manipulation" looks like, the below message is something from my spam folder. My personal Gmail experience with these emails is that they used to be a major problem during the first half of the year, with emails like this regularly landing in my inbox. It does seem like this RETVec tech upgrade works, though, because emails like this haven't been a problem at all for me in the last few months.



✅ 𝐂𝐡𝐞𝐜𝐤_𝐘𝐨𝐮𝐫_𝐀𝐜𝐜𝐨𝐮𝐧𝐭 ronamadeo🎁_𝐏𝐀𝐘𝐎𝐔𝐓_𝐕𝐄𝐑𝐈𝐅𝐈𝐂𝐀𝐓𝐈𝐎𝐍!◀️◀️__PAYMENT CODE ______------2ZB

JᴀᴄᴋpotCity master.tech@vf8ahh.ac.uk via blueshieldca.com
to TO._Vf8ahh@_Vf8ahh.Vf8ahh.Vf8ahh.inbox

Sun, Dec 3, 7:29 AM (1 day ago)

Why is this message in spam? It seems to be an auto-reply to a message that pretended to be sent from your email address.

Jᴀᴄᴋpot City

C0NGRATULATIONS ****@gmail.com !

A.balance..0F $1000.00 Is AVAILABLE F0R..your *Jᴀᴄᴋpot*.Accountt

Thiss.TRANSACTION.may.0nly.appearr. 0n.your.ACC0UNTT..afterr.VALIDATE.your.Info.

2023

FUNDING.For: ****

EMAIL: ****@gmail.com

Balance Amount: $1000.00

PAY0UT:
$1000.000

Memo PAY0UT

SIGNATURE ****

Ron Amadeo

Enlarge / An example of "adversarial text manipulation" from my spam folder.

Emails like this have been so difficult to classify because, while any spam filter could probably swat down an email that says, "Congratulations! A balance of $1,000 is available for your jackpot account," that's not what this email actually says. A big portion of the letters here are "homoglyphs"—by diving into the endless depths of the Unicode standard, you can find obscure characters that look like they're part of the normal Latin alphabet but actually aren't.

For instance, the subject "**Check_Your_Account**" is weirdly bolded not because it has bolded styling but because it uses Unicode glyphs like the "Mathematical Bold Capital C." It's a math symbol that happens to look like the letter "C" to people, but the robot doing spam filtering accurately views it as a math symbol and doesn't understand the intended English meaning. The closer you look at an email like this, the worse it gets: "C0NGRATULATIONS" has a zero replacing one of the "O" characters, the underlined letters in "Jᴀᴄᴋpot" are so strange they don't even come up in Unicode searches, and a lot of spaces are swapped out for periods or underscores. The result is that a spam filter looks at this *hot mess* of an email and basically gives up. (I don't understand why illegible emails default to "inbox" instead of "spam," but I'm not in charge.)

Google says RETVec is here to save the day: "RETVec is trained to be resilient against character-level manipulations including insertion, deletion, typos, homoglyphs, LEET substitution, and more. The RETVec model is trained on top of a novel character encoder which can encode all UTF-8 characters and words efficiently. Thus, RETVec works out-of-the-box on over 100 languages without the need for a lookup table or fixed vocabulary size."

Google says the efficiency here is a big deal. Alternative approaches that used a "fixed vocabulary size" or "lookup table" for homoglyphs made them resource-intensive to run. Imagine a list of every possible spelling and misspelling of "congratulations" that swaps out one or more characters for numbers, math symbols, Cyrillic, Hebrew, or emojis, and you have a nearly endless list. Google says RETVec is only 200,000 "instead of millions of parameters," so while Google's spam-filtering cloud is probably big enough to run anything, this is small enough that it could even run on a local device. RETVec is open source, and Google hopes it will rid the world of homoglyph attacks, so even your local comment section could be running it someday.

RETVec appears to work a lot like how humans read: It's a machine-learning TensorFlow model that uses visual "similarity" to identify what words mean instead of their actual character content. Google's similarity demo uses the same technology to identify pictures of cats, so turning that into the world's fanciest optical character recognition system sounds pretty doable. Apparently, this approach has led to big improvements, with Google saying: "Replacing the Gmail spam classifier's previous text vectorizer with RETVec allowed us to improve the spam detection rate over the baseline by 38% and reduce the false positive rate by 19.4%. Additionally, using RETVec reduced the TPU usage of the model by 83%, making the RETVec deployment one of the largest defense upgrades in recent years."

Google says it has been testing RETVec internally "for the past year," and it has already rolled out to your Gmail account.

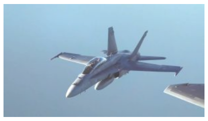
READER COMMENTS 129

RON AMADEO

Ron is the Reviews Editor at Ars Technica, where he specializes in Android OS and Google products. He is always on the hunt for a new gadget and loves to rip things apart to see how they work. He loves to tinker and always seems to be working on a new project.

SITREP: F-16 replacement search a signal of F-35 fail?

Footage courtesy of Dvids, Boeing, and The United States Navy.



SITREP: F-16 replacement search a signal of F-35 fail?

Sitrep: Boeing 707

Steve Burke of GamersNexus Reacts To Their Top 1000 Comments On YouTube

⊕ More videos

← PREVIOUS STORY NEXT STORY →

Related Stories

Today on Ars

STORE
SUBSCRIBE
ABOUT US
RSS FEEDS
VIEW MOBILE SITE

CONTACT US
STAFF
ADVERTISE WITH US
REPRINTS

NEWSLETTER SIGNUP

Join the Ars Orbital Transmission mailing list to get weekly updates delivered to your inbox. Sign me up →