Spam & Co.

Andrea Esuli

Spam!


Ingredients:
Pork with Ham,
Salt, Water,
Modified Potato
Starch, Sugar,
Sodium Nitrite
SPAM®
Classic
U.S.
INSPECTED
AND PASSED BY
DEPARTMENT OF
AGRICULTURE
Serving
Suggestion
NET WT
12 OZ
(340g)

Spamming

The act of **spamming** consists in sending **massive** amounts of **unsolicited** messages to a **large** number of recipients using a messaging platform (typically email).

Where does this meaning of the term spam originate?

Is this meaning somewhat related with Spam meat?

Spam!

Mail spam

Mail spam leverages on the virtually **zero cost** of sending an email to a huge numbers of mailboxes.

Messages usually are of **low quality**, with almost **no personalization** with respect to the recipient.

- Spam may just be advertising of legal/illegal products
- Spam may be an attempt at spreading a **virus**.
- Spam may be the starting point of a **scam**.
- Spam may be a **phishing** attempt.

Botnets are typically used to send emails.

Spam: history of a feature engineering war

Separating **spam** from **ham** is a binary classification problem.

First approaches used simple bayesian classifiers on the content of message.

Spammer counter attacked modifying the content of the message to make it hard to spot by filters, often paying more attention on fooling the spam-filter than the recipient. . .

`V1agra` `L0se W3ight`

`W0rk fr0M H0me` `fR33 iP0ds`

Spam: history of a feature engineering war

Spammers then started encapsulating their message into an attached image.

- Anti-spammers then used OCR to read text in images.

Spammers answer was to insert artifacts so as to make OCR fail.

- Anti-spammers dropped OCR and directly recognized these weird images.

Spammer currently gave up fighting spam filters, focusing more on user that do not use them (or even, convincing a user that message marked as spam is indeed a legit one).

Mail scam

Dear Beloved Friend,
I know this message will come to you as surprised but permit me of my desire to go into business relationship with you.
I am Miss Naomi Surugaba a daughter to late Al-badari Surugaba of Libya whom was murdered during the recent civil war in Libya in March 2011, before his death my late father was a strong supporter and a member of late Moammar Gadhafi Government in Tripoli.
Meanwhile before the incident, my late Father came to Cotonou Benin republic with the sum of USD4, 200,000.00 (US$4.2M) which he deposited in a Bank here in Cotonou Benin Republic West Africa for safe keeping.
I am here seeking for an avenue to transfer the fund to you in only you`re reliable and trustworthy person to Investment the fund. I am here in Benin Republic because of the death of my parent`s and I want you to help me transfer the fund into your bank account for investment purpose.
Please I will offer you 20% of the total sum of USD4.2M for your assistance. Please I wish to transfer the fund urgently without delay into your account and also wish to relocate to your country due to the poor condition in Benin, as to enable me continue my education as I was a medical student before the sudden death of my parent`s. Reply to my alternative email:missnaomisurugaba2@hotmail.com, Your immediate response would be appreciated.
Remain blessed,
Miss Naomi Surugaba.

Why Do Nigerian Scammers Say They are From Nigeria?

Cormac Herley - Microsoft

Mail scam

Why Do Nigerian Scammers Say They are From Nigeria?

Cormac Herley - Microsoft

The expected return of a scam attempt to a target user x can be modeled as:

$$P(viable|x)\cdot G - P(non\text{-}viable|x)\cdot C$$

where G is the gain from a successful scam and C is the cost of an unsuccessful attempt.

A scammer must be good at correctly classifying x as a viable target.

False positives produce a cost, false negatives are a missed gain but not a cost.

Mail scam

Why Do Nigerian Scammers Say They are From Nigeria?

Cormac Herley - Microsoft

A scammer must be good at correctly classifying *x* as a viable target.

*"Since **gullibility is unobservable**, the best strategy is to get those who possess this quality to **self-identify**. An email with tales of fabulous amounts of money and West African corruption will strike **all but the most gullible** as bizarre."* (section 4.1)

Phishing

Phishing is a type of attack that aims at stealing relevant personal information from the recipient.

Differently from scams, if a target falls in the trap set in the message there are little follow-up cost for the **phisher**.

For this reason, some resources may be spent on gathering and using contextual information to make the message appear realistic.

Phishing

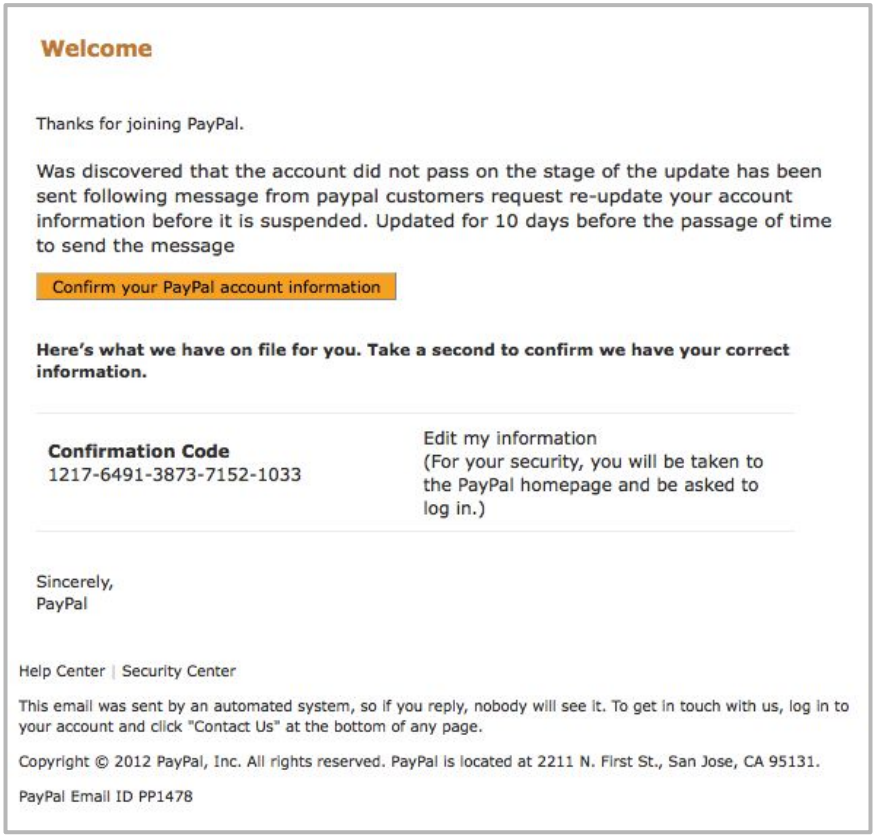
Gentile Andrea Esuli,

La cassetta postale ha superato il limite di archiviazione, che `e 20 GB come set del amministratore, si sta attualmente eseguendo il 20,9 GB, si potrebbe non essere in grado di inviare o ricevere nuovi messaggi fino a quando `e convalidare nuovamente la cassetta postale. A riconvalidare la cassetta postale, si prega di immettere e inviare a noi i tuoi dati qui sotto per verificare e aggiornare il tuo account:

(1) Posta elettronica: (2) Nome: (3) Password: (4) E-mail alternativo

Grazie
Consiglio Nazionale delle Ricerche

Phishing



Welcome

Thanks for joining PayPal.

Was discovered that the account did not pass on the stage of the update has been sent following message from paypal customers request re-update your account information before it is suspended. Updated for 10 days before the passage of time to send the message

Confirm your PayPal account information

Here's what we have on file for you. Take a second to confirm we have your correct information.

Confirmation Code
1217-6491-3873-7152-1033

Edit my information
(For your security, you will be taken to the PayPal homepage and be asked to log in.)

Sincerely,
PayPal

Help Center | Security Center

This email was sent by an automated system, so if you reply, nobody will see it. To get in touch with us, log in to your account and click "Contact Us" at the bottom of any page.

Copyright © 2012 PayPal, Inc. All rights reserved. PayPal is located at 2211 N. First St., San Jose, CA 95131.

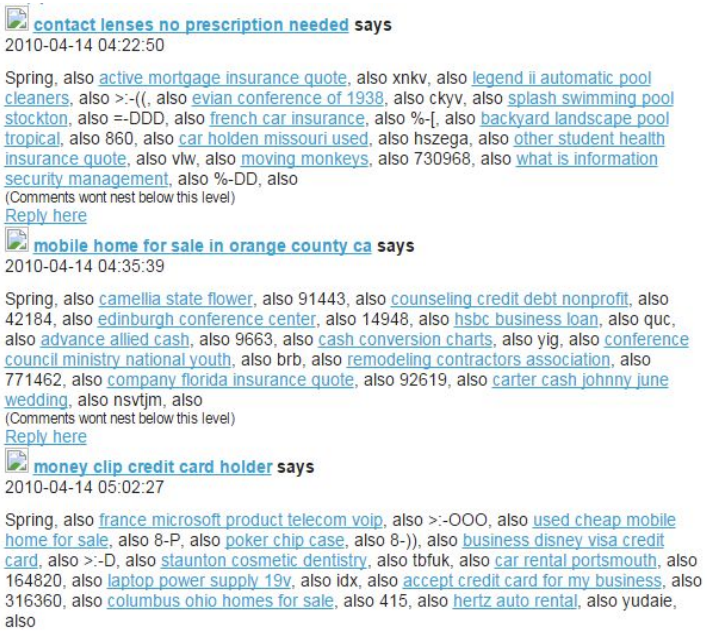
PayPal Email ID PP1478

Blog spam

Comments in blogs are spammed because they allow creating links to a target website, increasing the perceived relevance of the target in the Web graph.

- The actual comment is not relevant, because it is likely on a non-relevant website.
- Many websites have to be spammed in order to make it work.
- Blog spam is produced by botnets.

ML-based services like Askimet keep the problem confined.



contact lenses no prescription needed says
2010-04-14 04:22:50

Spring, also active mortgage insurance quote, also xnkv, also legend ii automatic pool cleaners, also >:-((, also evian conference of 1938, also ckyv, also splash swimming pool stockton, also =-DDD, also french car insurance, also %-[, also backyard landscape pool tropical, also 860, also car holden missouri used, also hszega, also other student health insurance quote, also vlw, also moving monkeys, also 730968, also what is information security management, also %-DD, also
(Comments wont nest below this level)
Reply here

mobile home for sale in orange county ca says
2010-04-14 04:35:39

Spring, also camellia state flower, also 91443, also counseling credit debt nonprofit, also 42184, also edinburgh conference center, also 14948, also hsbc business loan, also quc, also advance allied cash, also 9663, also cash conversion charts, also yig, also conference council ministry national youth, also brb, also remodeling contractors association, also 771462, also company florida insurance quote, also 92619, also carter cash johnny june wedding, also nsvtjm, also
(Comments wont nest below this level)
Reply here

money clip credit card holder says
2010-04-14 05:02:27

Spring, also france microsoft product telecom voip, also >:-OOO, also used cheap mobile home for sale, also 8-P, also poker chip case, also 8-)), also business disney visa credit card, also >:-D, also staunton cosmetic dentistry, also tbfuk, also car rental portsmouth, also 164820, also laptop power supply 19v, also idx, also accept credit card for my business, also 316360, also columbus ohio homes for sale, also 415, also hertz auto rental, also yudale, also

Fake reviews

Fake reviews are written to alter the generally perceived qualities of a product/service

- to promote the reviewed product,
- to criticize the reviewed product,
- to insert a reference to competing product and divert users to it.

These activities are obviously unfair and often illegal.

In the news

Chi scrive recensioni false su Tripadvisor rischia il carcere: nove mesi di reclusione a un truffatore - Owner of firm behind fake Tripadvisor reviews jailed in Italy

‘FAKE’ ONLINE REVIEWS HIT 85% OF HOTELS AND RESTAURANTS

Amazon's Fake Review Problem Is Now Worse Than Ever, Study Suggests

Researchers taught AI to write totally believable fake reviews, and the implications are terrifying

Italy fines TripAdvisor 500,000 Euros over false reviews

Companies to pay $350,000 fine over fake online reviews

Fake reviews

Spam/fake reviews are different from the previous cases:

- They do not aim at getting information from the user.
- They are strongly contextualized.
- They are typically hosted on a specific platform (Amazon, TripAdvisor, Booking, Google, Yelp).
- They have a high cost of production.
- Botnets can help, but not so much.

Producing a review has a cost

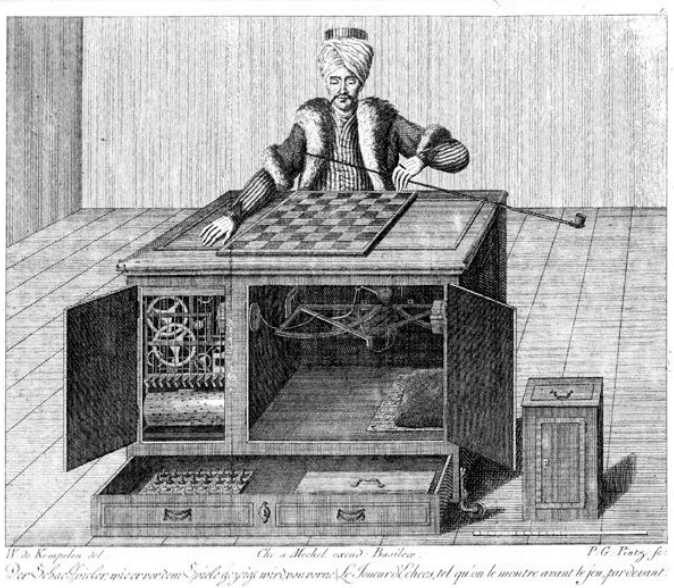
In order to submit a review:

- the user must be registered.
- the user must use the website interface or dedicated app (no API).
- the user can be asked to solve a **captcha**.
- the review must be of a minimum length.
- the review must be written correctly.
- the review must be on topic.

Most of these are hard/impossible tasks for bots.

Mechanical turk

Amazon Mechanical Turk is an example of platform used by who wants to **produce and distribute** fake reviews to hire **low cost workers** that perform the **tasks that bots cannot do**.



Help with outlook , hotmail email deliverability test

Requester: Kaan Aksoy

Description: Register for hotmail , outlook email addresses

Keywords: deliverability, register

Qualifications Required:	Your Value	
Masters has been granted	none	Request Qualification

Write 100-word restaurant reviews

Requester: Kelley Hornyak

Description: Write 100-word restaurant reviews using information from online review sites.

Keywords: writing, summary, reviews

Qualifications Required:	Your Value	
Review Writer 2 is not less than 100	none	Request Qualification

Fake reviews detection

Detection of fake reviews can be tackled as a binary classification problem.

Reviews are represented using two kinds of features:

- Internal features extracted from the text of the reviews.
- External features extracted from the review metadata and from more complex data structures related to the review context.

Internal features

The text of a fake review may exhibit differences with respect to text from real reviews:

- use of templates to produce reviews may produce unrealistic mentions of products:

  “I always used $PRODUCT NAME’’

  “I always used Sandisk SDMX26-008G-G46K Clip Jam MP3 Player’’

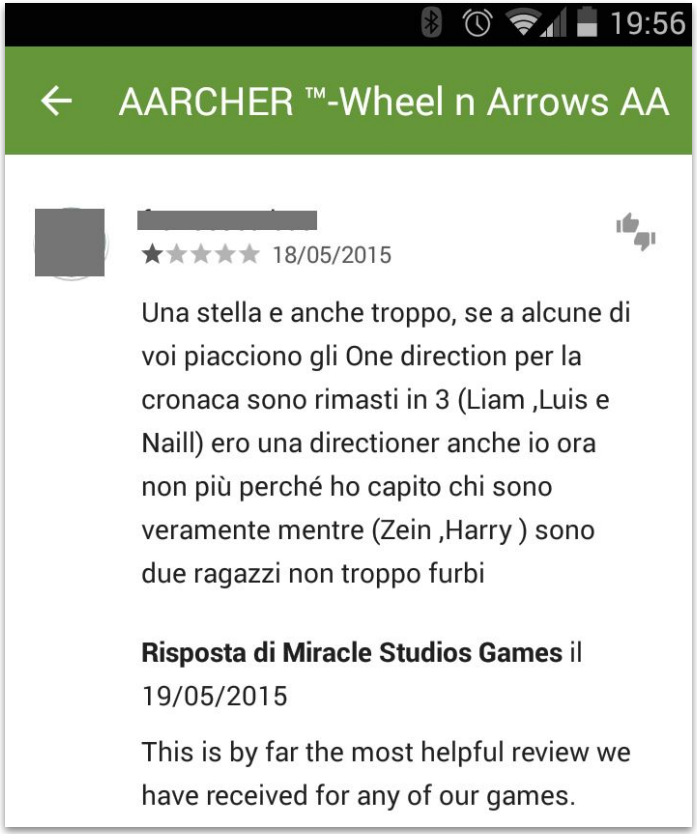
  features: exact matches of names and attributes of products as they are written in the product specifications

Internal features

- no mention of specific attributes of the product:

"Great value for the money! One of the best products on the market. Top quality and durability."

- text unrelated to the domain (just to fill up space and give more prominence to the star rating)



Internal features

Ott et al. created a dataset of true and fake reviews.

They trained an automatic classifier to tell fake/true reviews both for positive and negative polarity.

Linguistic comparison of true and fake reviews highlighted a number of distinct stylistic features that help telling fake and real reviews apart:

- different distribution of POS, fake reviews usually have more verbs, adverbs, and superlatives than real ones.
- real reviews have more sensorial (e.g., spatial) information.
- fake reviews exaggerate the polarized expressions.
- fake reviews use more frequently the first person.

External features

When a review is written it is a waste to use it only once.

- Check for recycled reviews across accounts.

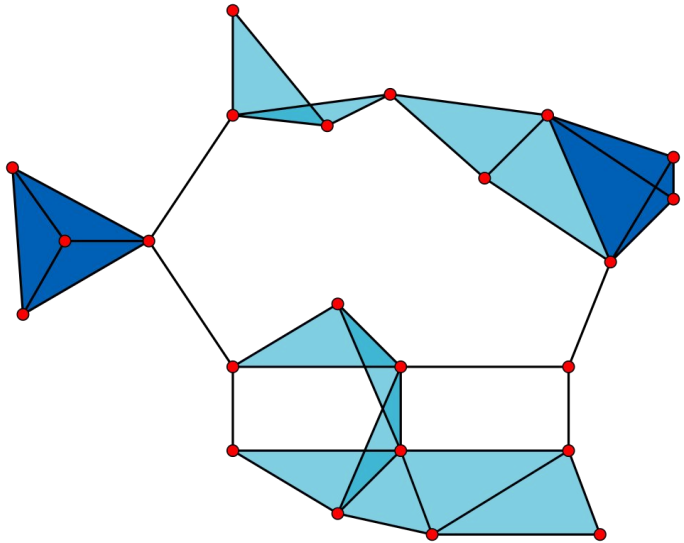
Once an account is created it is used multiple times.

- Check for accounts with a "strange" behavior in assigning stars.
- Check for inconsistencies across reviews from the same account.
- Check for time distribution of reviews.

External features

Multiple accounts are used to create a critical mass of reviews.

Check for cliques of users reviewing the same products.



External features

Information that is available only to the website owner:

- Detailed history of interaction with the website
- IP address
- Browser signature [check also this one]
- Geographic location

A typical case

http://www.amazon.com/JAMBALAYA-Audio-Xtract-Pro/dp/B0002VRPBO/

http://www.amazon.com/gp/pdp/profile/A3URRTIZEE8R7W
http://www.amazon.com/gp/pdp/profile/A254LYRIZUYXZG
http://www.amazon.com/gp/pdp/profile/A4XRKSD7CCPSH
http://www.amazon.com/gp/pdp/profile/A1RWJ387BL0FEK

The two sides

Writing fake reviews is a job.

Detecting fake reviews is a job too.

Clickbaiting

Website of a workshop on clickbaiting

10 things that will change your mind about clickbaiting

Clickbaiting aims at making as many people as possible clicking on links.

The techniques exploits natural curiosity of people with respect to some kind of messages.

A clickbait:

- gives to readers some interesting information, but not enough to satisfy their curiosity, urging them to follow the link
- leads to content that is not really as interesting as the initial link, and often it could be even unrelated.

Clickbaiting

What's the deal?

- Driving traffic to a specific website/account increases its relevance metrics.
- Visualizations may include advertising, which directly create a revenue for the clickbaiter.

Clickbaiting is another topic in which there are two sides:

- Recognizing clickbaits.
- Generating clickbaits.

Fake news

A fake news is a piece of text that states a fact that is not true.

Fake news may have different aims:

- humour, satire, sarcasm
- clickbaiting leveraging on hot and debated topics
- advertising
- supporting a political stance
- hurting someone's reputation
- spreading misinformation to disrupt social order

Fake news

Research on automatic fake news recognition has exploited image content to identify inconsistencies between textual and visual content.

- Use of fake images in tweets about hurricane Sandy
- Spotting fake tweets by using images similarity search

Fake news

Recognizing that a piece of text reports some false information is a very hard task that goes beyond simple language processing.

“US Unemployment went up during the Obama years"

“The Russians under Putin interfered with the US Presidential Election”

Text-only based automatic analysis is still in an early stage, starting from the simpler problem recognizing inconsistencies between documents.

Fake news challenge: stance detection

The long term goal is the ability to perform automatic fact checking.