

# Identificação de alertas de segurança virtual veiculados no Twitter

Jackson J. de Souza

Orientador: Daniel M. Batista

Instituto de Matemática e Estatística

IME-USP

11 de Novembro de 2013



IME - Instituto de  
Matemática e Estatística

# Introdução

- Computadores e internet bastante popularizados
- Redes sociais online também
- Nova gama de crimes possíveis =D
- Formas atuais de combate e prevenção são insuficientes

# Objetivo

- Identificar o surgimento de alertas por meio de postagem de mensagens
- Utilização do Twitter como fonte de informação
- Aprendizagem de máquina

# Proposta

Desenvolver um classificador de tuítes na língua inglesa que identifica alertas de segurança virtual utilizando a Weka

Para isso será feita uma comparação de desempenho entre os classificadores *SVM* e *Naive Bayes*

Este trabalho é derivado do estudo realizado pelo Rodrigo Campiolo (UTFPR) e o Luiz Artur (UTFPR).

Eles confirmaram a hipótese de que alertas de segurança virtual se espalham de forma confiável rapidamente no Twitter e informam sobre ameaças de segurança antes de alguns sites de mídia especializada



# Aprendizagem de máquina

Aprendizagem de Máquina é uma técnica em que se busca encontrar padrões nos dados relacionados à tarefa de interesse e são definidas regras ou maneiras de utilizar tais dados extraídos para que o sistema possa executar a tarefa de forma satisfatória.

O aprendizado é bem sucedido se ele se aperfeiçoa conforme aumenta a exposição aos dados relacionados ao problema que deve ser resolvido.

# Exemplo de aprendizagem



## Características

- Tamanho
- Cor
- Formato
- etc.

- O problema foi abordado com duas modelagens
  - 7 classes
  - 2 classes
  - As classes são mutuamente excludentes em ambos os casos

# Classes do problema

- Notícia sobre segurança virtual

“Antivirus News: McAfee mocks McAfee anti-virus in video rant - WA today: McAfee mocks McAfee anti-virus in vid...  
<http://t.co/nuhVDMpU48>

- Potencial Alerta de segurança virtual

“McAfee Threats Report Q2 2012: Malware spread fastest in last 4 years: The McAfee Security Threats report revea...  
<http://t.co/kb3BvyGP>



# Classes do problema

- Alerta de segurança virtual

“Malware spread on Skype taps victim PCs to mint bitcoin  
<http://t.co/snc8yqGIL9> by @ArsTechnica's @dangoodin001
- Alerta de segurança (com impacto direto sobre segurança não-virtual)

“RT @tweetjournal83: Hackers post personal data for Biden, Beyonce, Britney: The site includes credits reports, social security num... <http://t.co/Tter9uYroM>

# Classes do problema

- Segurança não-virtual (com influência direta sobre segurança virtual)

"Obama: Cyber attack serious threat to economy, national security:  
U.S. President Barack Obama is urging the Sena...

<http://t.co/igRpGRwb>

- Notícia sobre segurança não-virtual

"RT @Timodc: RT @Timodc: As Biden makes a absurdly false  
attack on Mitt, guess who has supported social security taxes?

Biden. <http://t.co/5YCI3etf>

- Spam

"Choose award-winning security for your PC. Choose @McAfee to  
protect against viruses, malware, and other threats.

<http://t.co/xcDrIMkl>

Para construir cada um dos modelos foram utilizados 1256 tuítes divididos em 66% (conjunto de treinamento) e 33% (conjunto de testes)

## Naive Bayes

Tuítes corretamente classificados	283	66.27%
Tuítes incorretamente classificados	144	33.72%

Taxa VP	Taxa FP	Precision	Recall	F-Measure	Classe
0.241	0.07	0.35	0.241	0.286	Notícia de segurança virtual
0.529	0.024	0.474	0.529	0.5	Potencial alerta de seg. virtual
0.695	0.17	0.691	0.695	0.693	Alerta de segurança virtual
0.606	0.046	0.526	0.606	0.563	Alerta de seg. virt. e seg. não virt.
0.385	0.022	0.357	0.385	0.37	Notícia de seg. geral e seg. virtual
0.923	0.054	0.882	0.923	0.902	Notícia de segurança geral
0.4	0.045	0.357	0.4	0.377	Spam
0.663	0.094	0.652	0.663	0.655	Média ponderada das classes

$$\text{Precisão: } \frac{pv}{pv+fp}$$

$$\text{Recall: } \frac{pv}{pv+fn}$$

$$F\text{-measure: } \frac{1}{\alpha \frac{1}{P} + (1-\alpha) \frac{1}{R}} \quad \alpha = \frac{1}{2}$$

## Naive Bayes

---

a	b	c	d	e	f	g	
14	2	24	3	4	1	10	a = Notícia de segurança virtual
1	20	7	2	0	2	1	b = Alerta de seg. virt. e seg. não virt.
19	14	105	4	4	1	4	c = Alerta de segurança virtual
0	0	4	120	2	2	2	d = Notícia de segurança geral
1	0	5	1	9	1	0	e = Potencial alerta de seg. virtual
0	1	5	1	0	5	1	f = Notícia de seg. geral e seg. virtual
5	1	2	5	0	2	10	g = Spam

---

## Matriz de confusão

## Support Vector Machines (SVM)

Tuítes corretamente classificados	335	78.45%
Tuítes incorretamente classificados	92	21.54%

Taxa VP	Taxa FP	Precision	Recall	F-Measure	Classe
0.754	0.185	0.799	0.754	0.776	Alerta de segurança virtual
0.815	0.246	0.772	0.815	0.793	Spam
0.785	0.216	0.785	0.785	0.784	Média ponderada das classes

$$\text{Precisão: } \frac{pv}{pv+fp}$$

$$\text{Recall: } \frac{pv}{pv+fn}$$

$$F\text{-measure: } \frac{1}{\alpha \frac{1}{p} + (1-\alpha) \frac{1}{R}} \quad \alpha = \frac{1}{2}$$

a	b	
159	52	a = Alerta de segurança virtual
40	176	b = Spam

Proposta de 2 modelos de classificação de tuítes que podem ser usados na construção de um sistema que identifica alertas de segurança

# Trabalhos futuros

- Coletar mais tuítes para aperfeiçoar o modelo de classificação
- Desenvolver um software que classifica tuítes logo após a postagem deles



# Agradecimentos

jackson@ime.usp.br