IDENTIFICAÇÃO DE ALERTAS DE SEGURANÇA VIRTUAL NO TWITTER

Jackson J. Souza — Orientador: Daniel M. Batista

jackson@ime.usp.br

Departamento de Ciência da Computação, Instituto de Matemática e Estatística, Universidade de São Paulo

INTRODUÇÃO

A popularização de dispositivos eletrônicos como computadores, *laptops*, *smartphones* entre outros, além da Internet abre espaço para uma grande variedade de crimes virtuais como *phishing*, invasão e furto de contas de usuários, roubo de dinheiro via *internet banking*, espionagem entre outros. Um dos maiores usos que as pessoas fazem atualmente da Internet é a navegação em redes sociais online, que estão entre os sites mais visitados na internet[a]. Entre elas o Twitter. Além disso, os usuários destes sites compartilham neles diversos tipos de informação [**?**].

[a]Informação obtida no dia 13/09/2013 em http://www.alexa.com/topsites

MOTIVAÇÃO

O problema de identificar alertas de segurança virtual (ASV) ainda não foi muito explorado e mesmo as pessoas que estudam ou trabalham com computação têm dificuldade de identificar ASVs. Isso foi constatado em uma pesquisa realizada com pessoas ligadas à área da computação na qual as pessoas precisavam ler 10 tuítes e responder para cada um deles se tratava-se de um ASV ou não sem ter sido apresentada uma definição de ASV.

Não obstante o fato de pessoas com conhecimento sobre computação não terem uma noção clara do que caracteriza um ASV, um ser humano é incapaz de de fazer essa identificação manualmente usando redes sociais como fontes, pois, por exemplo, são postados 9.100 tuítes por segundo no Twitter[a].

[a]http://www.statisticbrain.com/twitter-statistics/

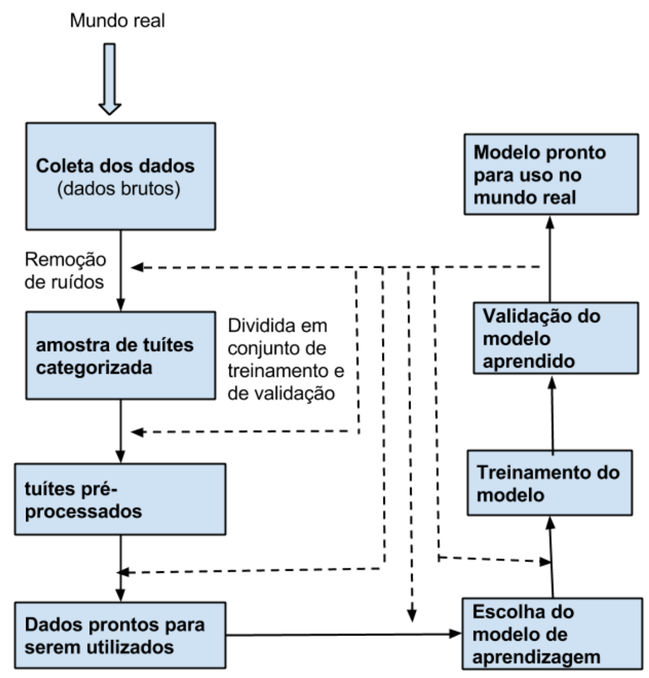
OBJETIVO

Neste trabalho é feito um estudo empírico das mensagens de segurança no Twitter escritas na língua inglesa para detectar alertas de segurança virtual. Para tal é feita uma comparação de desempenho entre os classificadores *Support vector machines* (SVM) e *Naive Bayes* na tarefa de detecção dos alertas de segurança computacional usando um software de mineração de dados, a Weka. Este estudo é derivado das teses de doutorado de Luiz A. F. Santos e Rodrigo Campiolo que estão relacionadas com a detecção antecipada de anomalias em redes de computadores.

CONTRIBUIÇÕES

Os resultados da classificação de tuítes mostram que este tipo de abordagem para identificar ASVs é bastante efetiva apesar do tamanho dos documentos (tuítes) serem pequenos. Inclusive, a efetividade da classificação serve de estímulo para realizar o mesmo tipo de estudo com postagens de outras redes sociais como o Facebook.

PROCESSO DE CLASSIFICAÇÃO



Fluxograma da classificação dos tuítes

CONCEITOS E DEFINIÇÕES

		Classe inferida	
		Sim	**Não**
Classe do tuíte	**Sim**	**verdadeiro positivo - vp**	**falso negativo - fn**
	Não	**falso positivo - fp**	**verdadeiro negativo - vn**

Matriz de confusão

Precisão (P): $\frac{vp}{vp+fp}$ *Recall* (R): $\frac{vp}{vp+fn}$

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

Conjunto de classes: $\mathbb{C} = \{c_1, c_2, \ldots c_n\}$
Conjunto de documentos: $\mathbb{D} = \{d_1, d_2, \ldots d_n\}$
$|V|$: Tamanho do dicionário
L_a: Número de tokens de um documento
M_a: Número de termos em um documento

CLASSIFICADOR *Naive Bayes*

$$P(c_i|d) \propto P(c_i) \prod_{1 \leqslant k \leqslant n_d} P(t_k|c_i)$$

Suavização de Laplace

$$\hat{P}(t|c) = \frac{T_{ct}+1}{\sum_{t' \in V}(T_{ct'}+1)} = \frac{T_{ct}+1}{(\sum_{t' \in V} T_{ct'}) + B'}$$

$$B = |V|$$

Modo	Complexidade do tempo
Treinamento	$\Theta(\lvert\mathbb{D}\rvert L_{medio} + \lvert\mathbb{C}\rvert\lvert V\rvert)$
Teste	$\Theta(L_a + \lvert\mathbb{C}\rvert M_a) = \Theta(\lvert\mathbb{C}\rvert M_a)$

CLASSIFICADOR *Support Vector Machines (SVM)*

max. $2/|\vec{w}|$
$\forall\ (\vec{x}_i, y_i) \in \mathbb{D},\ y_i(\vec{w}^T\vec{x}_i + b) \geqslant 1$

min. $\frac{1}{2}\vec{w}^T\vec{w}$
$b = y_k - \vec{w}^T\vec{x}_k \quad \forall \vec{x}_k \mid \alpha_k \neq 0$

$$f(\vec{x}) = \text{sgn}(\sum\nolimits_i \alpha_i y_i \vec{x}_i^T \vec{x}_i + b)$$

Modo	Complexidade do tempo
Treinamento	$O(\lvert\mathbb{C}\rvert\lvert\mathbb{D}\rvert^3 M_{medio})$
Teste	$O(T_a + \lvert\mathbb{C}\rvert M_a) = \Theta(\lvert\mathbb{C}\rvert M_a)$

RESULTADOS

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

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

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

Matriz de confusão

REFERÊNCIAS

[1] Danah M. Boyd and Nicole B. Ellison. Social network sites: Definition, history, and scholarship. *J. Computer-Mediated Communication*, 13(1):210–230, 2007.

[2] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *SIGKDD Explorations*, 11, 2009. Issue 1.

[3] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *An Introduction to Information Retrieval*. Cambridge University Press, draft edition, April 2009.

[4] LUIZ ARTHUR F. SANTOS, Rodrigo CAMPIOLO, MARCO AURELIO GEROSA, and DANIEL MACEDO BATISTA. Análise de mensagens de segurança postadas no twitter. *Anais do simpósio brasileiro de sistemas colaborativos (SBSC)*, (3):20–28, 2012.

[5] I.H. Witten, E. Frank, and M.A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science, 2011.