



IDENTIFICAÇÃO DE ALERTAS DE SEGURANÇA VIRTUAL NO TWITTER

Jackson J. Souza

jackson@ime.usp.br

Orientador: Daniel M. Batista

Coorientadora: Elisabeti Kira

IME-USP

Departamento de Ciência da Computação, IME-USP

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, 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 eles o Twitter. Além disso, os usuários destes sites compartilham diversos tipos de informação neles [3].

^aInformação obtida 12/11/2014 em <http://www.alexa.com/topsites>

MOTIVAÇÃO

O problema de identificar alertas de segurança virtual (ASV) em redes sociais 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 identificá-los como um ASV ou não-ASV sem ter sido apresentada uma definição de ASV. Apenas 40% dos participantes da pesquisa conseguiram identificar corretamente mais de 80% dos tuítes.

Além do fato de a maioria das pessoas com conhecimento sobre computação não terem uma noção clara do que caracteriza um ASV, um ser humano é incapaz de fazer essa identificação manualmente utilizando redes sociais como fontes, pois atualmente são postados 9.100 tuítes por segundo no Twitter^a.

^aInformação obtida 12/11/2014 em <http://www.statisticbrain.com/twitter-statistics/>

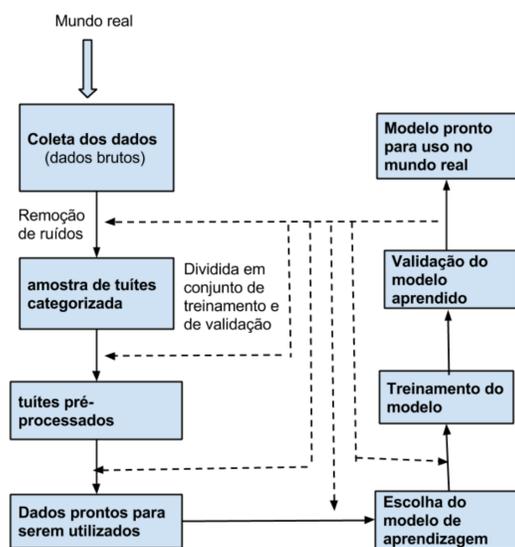
OBJETIVO

Neste trabalho é feito um estudo empírico das mensagens de segurança veiculadas no Twitter escritas em língua inglesa para detectar ASVs. Para tal, é feita uma comparação de desempenho entre os classificadores *Support vector machines* (SVM) e *Naive Bayes* na detecção de ASVs usando um software de mineração de dados, a Weka. Este estudo é derivado da tese de doutorado de Rodrigo Campiolo que busca detectar antecipadamente incidentes de segurança usando fontes de dados heterogêneas e abertas.

CONTRIBUIÇÕES

Os resultados da classificação de tuítes mostram que este tipo de abordagem para identificar ASVs possui um bom desempenho mesmo com documentos (tuítes) de tamanho pequeno e abrem espaço para a investigação de outras redes sociais como fontes de ASVs, por exemplo o Facebook.

PROCESSO DE CLASSIFICAÇÃO



Fluxograma da classificação dos tuítes

CONCEITOS E DEFINIÇÕES

Matriz de confusão

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

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: $\mathbb{C} = \{c_1, c_2, \dots, c_n\}$

Conjunto de documentos: $\mathbb{D} = \{d_1, d_2, \dots, d_n\}$

$|V|$: Tamanho do dicionário de tokens

L_a : Número de tokens de um documento

M_a : Número de termos em um documento

CLASSIFICADOR *Naive Bayes*

$$P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c)$$

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	Consumo de tempo
Treinamento	$\Theta(\mathbb{D} L_{medio} + \mathbb{C} V)$
Teste	$\Theta(L_a + \mathbb{C} M_a) = \Theta(\mathbb{C} M_a)$

CLASSIFICADOR *Support Vector Machines*

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

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

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

Modo	Consumo de tempo
Treinamento	$O(\mathbb{C} \mathbb{D} ^3 M_{medio})$
Teste	$O(T_a + \mathbb{C} M_a) = \Theta(\mathbb{C} M_a)$

TEORIA DA INFORMAÇÃO

Entropia

$$H(V) = - \sum_{t \in |V|} p(t) \log p(t), \quad t : \text{token}$$

Information Gain

$$H(c|V) = \sum_{k=0}^{|V|} P(V = t_k) H(c|V = t_k)$$

$$IG(c|V) = H(c) - H(c|V)$$

RESULTADOS

Naive Bayes

Tuítes corretamente classificados	2507	79.16%
Tuítes incorretamente classificados	660	20.84%

Taxa VP	Taxa FP	Precision	Recall	F-Measure	Classe
0.456	0.054	0.569	0.456	0.507	Notícia de segurança virtual
0.858	0.169	0.816	0.858	0.837	Alerta de segurança virtual
0.883	0.059	0.887	0.883	0.885	Notícia de segurança geral
0.474	0.035	0.439	0.474	0.456	Spam
0.792	0.108	0.787	0.792	0.788	Média ponderada das classes

a	b	c	d	
194	180	19	32	a = Notícia de segurança virtual
98	1267	68	44	c = Alerta de segurança virtual
18	81	964	29	d = Notícia de segurança geral
31	24	36	82	g = Spam

Support Vector Machines (SVM)

Tuítes corretamente classificados	2542	80.26%
Tuítes incorretamente classificados	625	19.73%

Taxa VP	Taxa FP	Precision	Recall	F-Measure	Classe
0.374	0.034	0.628	0.374	0.469	Notícia de segurança virtual
0.907	0.202	0.797	0.907	0.848	Alerta de segurança virtual
0.897	0.064	0.881	0.897	0.889	Notícia de segurança geral
0.37	0.019	0.525	0.37	0.434	Spam
0.803	0.122	0.789	0.803	0.789	Média ponderada das classes

a	b	c	d	
159	222	26	18	a = Notícia de segurança virtual
51	1339	62	25	c = Alerta de segurança virtual
14	83	980	15	d = Notícia de segurança geral
29	36	44	64	g = Spam

Os 9403 tuítes utilizados na criação do modelo foram coletados entre outubro de 2013 e 09/06/2014.

REFERÊNCIAS

- [1] 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.
- [2] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *An Introduction to Information Retrieval*. Cambridge University Press, draft edition, April 2009.
- [3] 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.

