

Jackson José de Souza

Identificação de alertas de segurança veiculados em redes sociais

São Paulo - Brasil

Setembro de 2013

Jackson José de Souza

Identificação de alertas de segurança veiculados em redes sociais

Universidade de São Paulo – USP

Instituto de Matemática e Estatística – IME-USP

Trabalho de formatura

Orientador: Daniel M. Batista

São Paulo - Brasil

Setembro de 2013

Agradecimentos

Resumo

Palavras-chaves: segurança computacional, redes sociais, twitter, aprendizado de máquina.

Lista de figuras

Lista de tabelas

Sumário

I	Parte objetiva	13
1	Introdução	15
1.1	Motivação do trabalho	16
1.2	Objetivo	17
1.3	Estrutura do trabalho	17
2	Atividades do trabalho	19
2.1	Atividades realizadas	19
2.1.1	Leitura de bibliografia	19
2.1.2	Desenvolvimento	20
2.2	O que falta fazer	20
2.2.1	Leitura de bibliografia	20
2.2.2	Desenvolvimento	21
2.2.3	Experimentos, discussões e resultados	21
3	Conceitos básicos	23
3.1	Redes sociais e segurança computacional	23
3.1.1	Segurança virtual	23
3.1.2	Redes sociais e Twitter	24
3.2	Aprendizagem Computacional	24
3.3	Estatística	24
3.4	Análise de dados	24
4	Desenvolvimento	25
4.1	Classes do problema	25
4.2	Coleta dos dados	25
4.3	Pré-processamento dos dados	25
4.4	Seleção de conjunto de teste e treinamento	25
4.5	Seleção das características do problema	25
5	Experimentos, discussões e resultados	27
6	Conclusão	29
II	Parte subjetiva	31
	Desafios e frustrações	33

Relação entre o trabalho e as disciplinas cursadas no BCC	35
Próximos passos	37
Referências	39

Parte I

Parte objetiva

1 Introdução

O alto grau de domínio tecnológico na fabricação de computadores e a diminuição do seu custo possibilitou a popularização do seu uso pela sociedade. Essa popularização tem provocado, nas últimas décadas, uma verdadeira revolução na sociedade. Tal revolução tem transformado a forma como as pessoas consomem, se comunicam, se relacionam com empresas, se entretêm etc. Como várias das atividades citadas acima necessitam da Internet para serem realizadas, a disseminação do seu uso acompanhou naturalmente o aumento do uso dos computadores. Dessa forma, ela tornou-se um ambiente ubíquo e que pode ser acessado também por dispositivos como vídeo-games, celulares, relógios, etc. Por sua vez, as redes sociais online atraíram milhões de usuários desde a sua introdução devido ao amplo uso da Internet. Elas se tornaram tão populares que entre as 10 páginas mais acessadas na internet 3 são redes sociais¹.

As redes sociais online possuem diferentes tipos de interesses como, a comunicação entre a comunidade de uma universidade, publicidade de empresas etc. Algumas delas tentam atrair os mais variados tipos de audiência e permitem aos usuários o compartilhamento de diversos tipos de informações, desde relatos sobre o cotidiano das pessoas, fofocas sobre celebridades, até notícias importantes e de última hora. O microblog Twitter é um exemplo. Ele pode ser utilizado tanto como rede social, quanto como fonte de notícias.

O fato de de 10 bilhões² de dispositivos estarem conectados à Internet mostra a abrangência do seu uso e a sua importância como meio de comunicação. Como em todo meio de comunicação, a segurança das informações, no caso dados, transmitidas é fundamental. Afinal, apenas em um meio de comunicação seguro é possível viabilizar e assegurar a disponibilidade, a integridade, a confidencialidade e a autenticidade das informações. Porém, não há um controle rígido sobre o tráfego de dados que passa pela Internet, o que facilita que informações sejam extraídas de computadores e boa parte dos usuários não tenha ciência disso. Isso demonstra que a liberdade de comunicação proporcionada pela Internet aliada às falhas de segurança presentes nos softwares que a usam revelam um risco a segurança de pessoas, empresas, instituições etc. Este risco é grave porque as brechas na segurança virtual podem provocar problemas no mundo real, como grandes prejuízos financeiros às vítimas. Existem alguns softwares desenvolvidos para proteger os computadores contra falhas de segurança como anti-vírus, firewalls, anti-adwares entre

¹ Informação fornecida no dia 13/09/2013 em <http://www.alexa.com/topsites>

² <http://gigaom.com/2011/10/13/internet-of-things-will-have-24-billion-devices-by-2020/>

outros. Porém, eles apenas funcionam contra softwares maliciosos já identificados e para os quais já foi encontrada uma forma de prevenção.

Consequentemente, há falhas para as quais não foi encontrada uma solução. Seja por não terem sido identificadas ainda ou por serem muito recentes. Tais buracos na segurança dos softwares são críticos, pois podem ser explorados por pessoas mal intencionadas. Logo, é necessário corrigir tais brechas o quanto antes, mas para isso tais falhas devem ser identificadas. Entre as várias formas de se descobrir falhas de segurança em um software existe a identificação de alertas de segurança virtual (ASV) veiculados pela rede em sites de segurança, fóruns etc. Em (SANTOS et al., 2012) foi mostrado que é possível utilizar redes sociais para detecção de ASVs, como no próprio Twitter ³. Contudo, os ASVs não são separados em uma categoria específica e não é fácil encontrá-los usando as ferramentas de busca disponibilizadas pelas redes sociais. Logo, percebeu-se que esse é um problema interessante de se abordar e é dele que este trabalho trata.

1.1 Motivação do trabalho

O problema da identificação de ASVs ainda não foi muito explorado e várias pessoas dentro da própria comunidade da área da computação não sabem definir bem o que caracteriza um ASV. Essa constatação foi feita após a análise de uma pesquisa lançada no IME na qual se desejava perceber qual a percepção que as pessoas tinham de ASVs. Dentre as questões havia 10 tweets para serem classificados. Os participantes da pesquisa deveriam dizer se cada tweet continha ou não um ASV ou se ela não sabia. Na questão da pesquisa aberta a comentários alguns participantes manifestaram ter sentido dificuldade em fazer a classificação dos tweets. Para exemplificar a dificuldade seguem abaixo as opiniões enviadas por 2 participantes da pesquisa:

“Muitas possibilidades para definir o que é segurança virtual. Em um universo de expressões infinitas. Virus Definitions Update Download -> Definitions Update Download is a Virus. São muito próximas as expressões, mas diferentes. A questão é o que muda nas duas?”

“É meio difícil separar o que é “alerta” mesmo (urgente, corra para se proteger/atualizar algo específico) do que é notícia relacionada com segurança (algo mais genérico, como a história dos plugins de browser), mas todos são relevantes no aspecto de segurança digital. Claro que tem que separar notícias que realmente falam de segurança daquelas que não tem nenhum conteúdo relevante nesse aspecto (ex: a da venda do exploit).”

A dificuldade de identificar alertas de segurança também se revelou nas classificações. Alguns tweets foram classificados como alerta de segurança virtual por aproximadamente 50% dos participantes enquanto os outros cerca de 50% os classificaram como não sendo alertas de segurança virtual. Além da tarefa de identificar um AVS em publicações de

³ <https://www.twitter.com>

redes sociais não ser simples, é inviável a um ser humano olhar cada possível alerta e classificá-lo manualmente, dado que 9.100 tweets por segundo são postados no twitter⁴. Por isso, uma solução pra o problema é treinar um sistema para identificar os alertas de segurança automaticamente.

1.2 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 será feita uma comparação de desempenho entre os classificadores *Support vector machines* (SVM) e *Naive Bayes* para a detecção dos alertas de segurança computacional. Este estudo serve de apoio às teses de doutorado do Luiz A. F. Santos e do Rodrigo Campiolo que estão relacionadas com a detecção antecipada de anomalias em redes de computadores.

1.3 Estrutura do trabalho

O trabalho está dividido da seguinte forma:

- **Atividades do trabalho** Breve descrição das atividades realizadas até agora.
Este capítulo estará presente apenas na versão preliminar da monografia.
- **Conceitos básicos**
Introdução teórica a conceitos importantes para a compreensão do desenvolvimento do trabalho
- **Desenvolvimento**
Possui a formulação do problema e o desenvolvimento teórico e prático do trabalho
- **Experimentos, discussões e resultados**
Aplicação do problema, discussão do método utilizado e apresentação dos resultados
- **Conclusão**
Conclusão do trabalho desenvolvido

⁴ <http://www.statisticbrain.com/twitter-statistics/>

2 Atividades do trabalho

Este capítulo faz uma breve descrição das atividades realizadas e as tarefas que ainda necessitam ser executadas.

Vale observar que nem tudo que já foi feito está documentado na monografia. Isso se deve a proximidade da entrega da monografia e de compromissos particulares e acadêmicos que disputaram com a monografia o tempo disponível para a realização dos compromissos.

2.1 Atividades realizadas

2.1.1 Leitura de bibliografia

Aqui fala-se sobre a bibliografia lida ou pesquisada para a realização do trabalho. A parte da bibliografia consultada cujo conhecimento foi utilizado no trabalho está presente na seção de referências

- Pesquisa sobre conceitos básicos necessários para a compreensão do trabalho.
- Estudei os capítulos 1, 2 e 3 do livro “An introduction to information retrieval”.
- Também fiz a leitura parcial de alguns artigos sobre segurança computacional e rede sociais, para buscar definições e explicações de alguns conceitos visando dar uma maior sustentação teórica ao trabalho.
- Li alguns artigos sobre classificação de texto envolvendo bloom filters e uma abordagem clássica que aborda n-grams. Inclusive testei duas gems (bibliotecas) Ruby para verificar a precisão de identificação da língua de um texto. Porém, ambos foram considerados com acurácia não muito boa ($< 90\%$). Por fim, li um artigo que comparava algumas abordagens de identificação de língua em documentos e, assim, consegui encontrar um classificador bastante preciso e que funciona não apenas com línguas ocidentais, mas orientais também.
- Para aprender mais sobre classificação de textos tenho me mantido atento a palestras e seminários oferecidos no IME sobre o tema e acabei por assistir um seminário oferecido pelo Altigran Soares da Silva da UFAM cujo título é “Explorando Dados Estruturados em Conteúdo Textual da Web: Métodos, Técnicas e Aplicações” que foi bastante útil para conhecer mais sobre a área de classificação de textos.

2.1.2 Desenvolvimento

- Foi feita uma pesquisa para verificar a percepção das pessoas sobre alertas de segurança computacional.
- Elaboração de um script que me auxilia na coleta dos dados que irão compôr o conjunto de testes e o de treinamento.
- Foram classificados 4000 tweets para compôr o conjunto de testes e treinamento para validar os classificadores que serão usados no trabalho.
- Os dados estão sendo estudados durante a coleta dos dados para que se pudesse conhecê-los melhor, extrair métricas e num futuro próximo escolher boas características (features) que ajudam a diferenciar os dados na sua classificação.
- Também foram escritos os scripts que fazem o pré-processamento dos dados e que estão disponíveis em
- Também li alguns exemplos de uso da Weka, o programa que será utilizado pra fazer a classificação dos tweets.
- Foram definidas as classes do problema. Elas são:
 - Spam
 - Alerta de segurança definido
 - Alerta de segurança indefinido
 - Informação sobre segurança não-virtual
 - Informação sobre segurança virtual

virtual = relativo a computadores ou que se realiza neles.

Mais detalhes sobre as classes estarão disponíveis na versão final da monografia.

2.2 O que falta fazer

2.2.1 Leitura de bibliografia

- Estudar os capítulos 13, 14 e 15 do livro “An introduction to information retrieval”.
- Preciso escolher as características (features), estudar e implementar os classificadores que serão adotados e aplicá-los ao problema.

2.2.2 Desenvolvimento

- Testes de uso da Weka. Inclusive, vou fazer uma demonstração dela pro meu orientador até a última semana de setembro.
- Falta fazer testes para os scripts responsáveis pelo pré-processamento dos tweets. Em especial o script que identifica a língua de um texto, pois é necessário mostrar a sua acurácia para que seja possível avaliar a sua validade ou utilidade.
- Falta coletar cerca de 5000 tweets para construir o conjunto de testes e treinamento. A ideia é usar 9000 tweets ao todo para construir o classificador de alertas de segurança computacional. Considera-se 9000 um bom número para obter uma boa acurácia para o problema que o trabalho se propõe a tratar.
- É necessário implementar os classificadores e fazer os devidos ajustes nas features dos tweets.

2.2.3 Experimentos, discussões e resultados

- Após fazer a validação da classificação dos tweets é necessário fazer análises estatísticas e fazer ajustes nas features e classificadores de forma a otimizar a classificação e evitar overfitting. O overfitting consiste no exagero de precisão na escolha das features para fazer com que o classificador alcance alta precisão. Porém, o risco do classificador se tornar pouco acurado torna-se muito alto.

3 Conceitos básicos

3.1 Redes sociais e segurança computacional

Esta seção aborda os conceitos de segurança utilizados para entender melhor o conteúdo dos dados (tweets) e, assim, classificá-los adequadamente. Também é definido o que é uma rede social e como os alertas são gerados e propagados na rede social estudada, o Twitter.

3.1.1 Segurança virtual

Segurança virtual consiste na prevenção de dano, uso não autorizado, exploração e, se necessário, na restauração de dados, sistemas de comunicação e de informação contida nos dados para garantir confidencialidade, integridade e acessibilidade de dados e programas de computador. Tradução livre de ([FUTURE, 2011](#)) pág. 43.

Uma ameaça de segurança é qualquer esforço identificado cujo objetivo seja acessar, extrair, manipular ou danificar a integridade, confidencialidade, segurança ou acessibilidade de dados ou programas de computador, sejam públicos ou privados, sem autorização legal. Tradução livre de ([SECURITY, 2010](#)) pág. 3.

Uma ameaça pode ser intencional ou não intencional e ter um alvo específico ou vago. Ela pode surgir a partir de diferentes fontes, incluindo um Estado fazendo espionagem ou guerra de informações contra outros, criminosos, crackers, programadores de vírus e funcionários e prestadores de serviços descontentes trabalhando em uma empresa.

Ameaças não intencionais podem ser causadas por funcionários displicentes ou sem treinamento, atualizações de programa de computador, realização de manutenções e falhas de equipamentos. Todas essas fontes podem danificar dados ou provocar interrupções ou mal funcionamento de sistemas.

Ameaças intencionais incluem ataques com alvo específico e alvo vago. Um ataque com alvo específico ocorre quando um grupo de pessoas ou um único indivíduo realiza um ataque contra um sistema de infraestrutura crítica. Um ataque de alvo vago ocorre quando o alvo definido para a realização do ataque não é claro, como é o caso de um vírus, *worm* ou *malware* liberado na internet sem alvo específico.

3.1.2 Redes sociais e Twitter

([BOYD; ELLISON, 2007](#))

3.2 Aprendizagem Computacional

Aprendizagem de Máquina é uma subárea de Inteligência Artificial cujo objetivo é construir sistemas que consigam aprender uma tarefa e melhorar seu desempenho ao executá-la conforme aumenta o seu conhecimento em como lidar com ela. Aprendizagem de Máquina é uma técnica em que se busca utilizar o raciocínio indutivo extraindo regras e padrões a partir de um conjunto de dados para aprender sobre eles. O aprendizado é bem sucedido se ele se aperfeiçoa conforme aumenta a exposição aos dados relacionados à tarefa de interesse. Uma outra forma mais sucinta de como funciona a aprendizagem de máquina é a seguinte:

Um programa de computador aprende a partir da experiência E com respeito a uma classe de tarefas T e medida de desempenho P , se a sua performance em T , segundo a medição P , melhora com a experiência E . Tradução livre de ([MITCHELL, 1997](#))

Por exemplo, um sistema que utiliza Aprendizagem de Máquina pode ser treinado para distinguir mensagens enviadas por e-mail e separá-las em mensagens spam e não spam. Após os resultados do treinamento nesta tarefa serem considerados satisfatórios o sistema será utilizado para separar, ou classificar, as mensagens de e-mail que irão para a caixa de entrada e as que irão para a pasta de spam.

3.3 Estatística

3.4 Análise de dados

4 Desenvolvimento

4.1 Classes do problema

4.2 Coleta dos dados

4.3 Pré-processamento dos dados

Para pré-processar os dados é necessário conhecê-los. Ou seja, como são representados, o significado dos metadados presentes neles e os tipos de ruído que eles possuem. É importante destacar que o filtro da api do twitter utilizado para coletar os tweets possui uma detecção que não funciona muito bem para identificar o idioma em que o tweet foi escrito. Por isso foi utilizado um classificador que identifica o idioma do tweet.

Levando em conta as ferramentas disponíveis para efetuar a filtragem dos dados o pré-processamento dos dados foi feito em duas fases:

1. Um script ruby olha o texto de cada tweet, pega o texto presente na headline de cada página cujo link se encontra no texto do tweet e adiciona ao tweet. Depois são decodificados os caracteres html presentes no texto do tweet e após isso são removidos os tweets que não possuem informações suficientes para serem classificados segundo os critérios definidos abaixo.
2. Um script perl extrai o conteúdo do tweet que contém idioma escrito, desconsiderando os metadados, e identifica o idioma dele, usando um processo de classificação. Os tweets que não estão escritos em idioma inglês são removidos.

Critérios que um tweet deve satisfazer para ser considerado no problema:

4.4 Seleção de conjunto de teste e treinamento

4.5 Seleção das características do problema

5 Experimentos, discussões e resultados

6 Conclusão

Parte II

Parte subjetiva

Desafios e frustrações

Relação entre o trabalho e as disciplinas cursadas no BCC

Próximos passos

Referências

BOYD, D. M.; ELLISON, N. B. Social network sites: Definition, history, and scholarship. *J. Computer-Mediated Communication*, v. 13, n. 1, p. 210–230, 2007. Citado na página [24](#).

FUTURE, D. of H. S. *Blueprint for a Secure Cyber Future: The Cybersecurity Strategy for the Homeland Security Enterprise*. [S.l.], 2011. Citado na página [23](#).

MITCHELL, T. M. *Machine Learning*. 1. ed. New York, NY, USA: McGraw-Hill, Inc., 1997. ISBN 0070428077, 9780070428072. Citado na página [24](#).

SANTOS, L. A. F. et al. Análise de mensagens de segurança postadas no twitter. *Anais do simpósio brasileiro de sistemas colaborativos (SBSC)*, n. 3, p. 20–28, 2012. Citado na página [16](#).

SECURITY, U. D. of H. *Privacy Impact Assessment for the Initiative Three Exercise*. [S.l.], 2010. Citado na página [23](#).