

Jackson José de Souza

# **Identificação de alertas de segurança veiculados em redes sociais**

São Paulo - Brasil

9 de novembro 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

9 de novembro de 2013



# Agradecimentos



# Resumo

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





# Lista de figuras

Figura 1 – Exemplo de Alerta de segurança virtual . . . . .	17
Figura 2 – Exemplo de tuíte com os tipos de metadados mais comuns no Twitter .	18



# Lista de tabelas

Tabela 1	– Dado o dicionário de palavras acima e o seguinte documento: “O modelo sacola de palavras é bem simples, pois contém apenas palavras.” temos que a probabilidade da existência do documento acima dada a sacola de palavras é dada por Puni (sacola, palavras, simples, palavras) $= 0.3*0.2*0.15*0.2$ . . . . .	17
Tabela 2	– Matriz de confusão . . . . .	18



# Sumário

<b>I</b>	<b>Parte objetiva</b>	<b>13</b>
<b>1</b>	<b>Introdução</b>	<b>15</b>
1.1	Motivação do trabalho	15
1.2	Objetivo	15
1.3	Estrutura do trabalho	15
<b>2</b>	<b>Conceitos básicos</b>	<b>17</b>
2.1	Pré-processamento e transformação de dados	17
2.2	Estatística	17
2.3	Segurança	17
2.4	Segurança virtual	17
2.5	Redes sociais e Twitter	17
2.5.1	Twitter	18
2.6	Aprendizagem Computacional	18
<b>3</b>	<b>Desenvolvimento</b>	<b>19</b>
3.1	Classes do problema	19
3.2	Coleta dos dados	19
3.3	Pré-processamento dos dados	19
3.4	Seleção de conjunto de teste e treinamento	19
3.5	Seleção das características do problema	19
<b>4</b>	<b>Experimentos, discussões e resultados</b>	<b>21</b>
<b>5</b>	<b>Conclusão</b>	<b>23</b>
<b>II</b>	<b>Parte subjetiva</b>	<b>25</b>
	<b>Desafios e frustrações</b>	<b>27</b>
	<b>Relação entre o trabalho e as disciplinas cursadas no BCC</b>	<b>29</b>
	<b>Próximos passos</b>	<b>31</b>
	<b>Referências</b>	<b>33</b>



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 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. Redes sociais são sites que permite interação entre os usuários cadastrados na página. As interações permitidas por uma rede social entre usuários são bastante variadas e as mais comuns são a comunicação e o compartilhamento de informações sobre assuntos de interesse em comum. Usuários podem pessoas, organizações, organizações, instituições, etc. Elas se tornaram tão populares que entre as 10 páginas mais acessadas na internet 3 são redes sociais online<sup>1</sup>. As redes sociais online serão chamadas neste trabalho de redes sociais para a leitura deste termo não cansar o leitor.

As redes sociais 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 bilhões<sup>2</sup> 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

---

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

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

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 outros. Alguns deles, inclusive, utilizam heurísticas para detectar ameaças que não tenham sido identificadas e catalogadas. Apesar da existência de tais programas de tais softwares, a quantidade de ataques e invasões ainda causam grandes prejuízos. Estudos apontam que as perdas com crimes virtuais alcançam a casa das centenas de bilhões de dólares em prejuízos sofridos por usuários e empresas a cada ano (STRATEGIC; STUDIES, 2013).

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<sup>3</sup>. 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. Assim, 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 tuítes para serem classificados. Os participantes da pesquisa deveriam dizer se cada tuíte 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 tuítes. 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?”

---

<sup>3</sup> <<https://www.twitter.com>>

“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 tuítes 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 redes sociais não ser simples, é inviável a um ser humano olhar cada possível alerta e classificá-lo manualmente, dado que 9.100 tuítes por segundo são postados no Twitter<sup>4</sup>. 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:

- **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

---

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



## 2 Conceitos básicos

Este capítulo apresenta a teoria que sustenta o desenvolvimento do trabalho desenvolvido. A [seção 2.1](#) apresenta conceitos envolvendo leitura de documentos, extração de termos deles e a posterior remoção, redução, contagem de termos e transformação dos dados para a extração das características dos tuítes. A [seção 2.2](#) apresenta os conceitos de estatística necessários para compreender os métodos utilizados para a análise de texto e classificação dos tuítes. A [seção 2.3](#) define os conceitos de segurança adotados neste trabalho e a [seção 2.4](#) define os conceitos envolvendo especificamente segurança virtual para que se entenda como foram escolhidas as classes dos tuítes. A [seção 2.5](#) apresenta uma definição de rede social, explica como elas funcionam e apresenta o Twitter, suas características principais e como os usuários podem compartilhar conteúdo nele em suas várias formas. A [seção 2.6](#) introduz conceitos relacionados à aprendizagem de máquina para entender o processo de classificação dos tuítes e os resultados do processo.

### 2.1 Pré-processamento e transformação de dados

Esta seção introduz definições de filtragem e tratamento de texto incluindo definições sobre específicos conjuntos de caracteres. No caso deste trabalho vamos nos restringir...

Uma classe é determinada por um conjunto de objetos que possui características em comum.

URL (*Uniform Resource Locator*), ou em português Localizador-Padrão de Recursos, é a designação usada para uma cadeia (conjunto) de caracteres que indicam a localização de um recurso em uma rede. Neste trabalho a URL pode ser entendida como o endereço de uma página na Internet. URL curta é um endereço reduzido de uma página que costuma ser utilizado para referenciar o endereço original de um site em textos cujo limite de caracteres para escrita é reduzido como o Twitter. Quando se deseja obter uma URL curta pode-se usar um serviço de encurtamento de URLs como o [<http://tinyurl.com/>](http://tinyurl.com/). Tokenização é o ato de decompor um documento em peças chamadas tokens. As peças são ocorrências específicas de cadeias de caracteres e são separadas por caracteres chamados delimitadores. Eis um exemplo:

Delimitadores (entre aspas simples): ‘,’ ‘;’ ‘ ’

Documento: Friends, Romans, Countrymen, lend me your ears;

Tokens:

Friends	Romans	Countrymen	lend	me	your	ears
---------	--------	------------	------	----	------	------

Note o caractere espaço ‘ ’ é também um delimitador de token.

Normalização de tokens é a tarefa de converter em uma forma canônica tokens que sejam diferentes mas que possuam um mesmo significado. Assim, temos o token canônico que dá nome à sua classe de equivalência e quaisquer tokens que sejam convertíveis ao canônico pertencem à mesma classe de equivalência. Qualquer token que pertence a uma dada classe de equivalência é representante dela e deve casar com qualquer outro token da classe, embora haja diferenças entre as cadeias de caracteres dos seus tokens.

Exemplos:

naive => Naive, Naïve, naïve, NAÏVE, NAIVE

usa => U.S.A., USA

Radicalizamento é o processo de reduzir palavras flexionadas ou derivadas à uma forma básica comum, o radical. O radical neste caso não precisa ser igual ao seu homônimo linguístico.

Exemplo: real, reais, realizar, realizável, realista => rea

Metadado é um dado sobre um dado. De outra forma, podemos dizer que um metadado é uma informação sobre um dado. Por exemplo, quando se busca tuítes podemos utilizar filtros que exibam como resposta apenas os tuítes em língua inglesa ou que tenham sido escritos no dia 17/04/2013. A língua e a data são exemplos de metadados que um tuíte possui, pois são informações sobre um tuíte que é o documento (dado) de interesse.

Aqui consideraremos como termo uma classe de todos os tokens que possuem exatamente a mesma cadeia de caracteres e que faz parte do dicionário de classificação. Em outras palavras, um token é uma cópia de um termo de forma que um documento pode conter várias cópias do seu representante, mas um termo é único.

Frequência de um termo é o número de vezes que um termo ocorre em um documento e é denotado por  $tft,d$  (transformar em fórmula).

O modelo sacola de palavras é uma representação simplificadora utilizada em problemas de classificação. No modelo o documento ou objeto é representado por uma coleção de termos (ou palavras) sem preocupação com a ordem dos termos. Porém, o número de ocorrência de cada termo é guardada.

Modelo de língua unigrama considera que a probabilidade da ocorrência de cada termo é independente da ocorrência prévia de quaisquer termos. Ou seja, a ordem da ocorrência dos termos não importa. Então, temos:  $Puni(t_1 \dots t_n) = \text{produtório da probabilidade de cada termo}$

OBS: Vale ressaltar que os vários conceitos de tratamento mencionados acima

Termo	Probabilidade
sacola	0.3
palavras	0.2
simples	0.15

Tabela 1 – Dado o dicionário de palavras acima e o seguinte documento: “O modelo sacola de palavras é bem simples, pois contém apenas palavras.” temos que a probabilidade da existência do documento acima dada a sacola de palavras é dada por Puni (sacola, palavras, simples, palavras) =  $0.3 \cdot 0.2 \cdot 0.15 \cdot 0.2$

podem ser aplicados de várias formas e isso depende do programa de computador utilizado. Ou seja, há mais de uma forma de se fazer radicalizamento, o que depende dos objetivos da necessidade do conteúdo do texto que está sendo analisado.

## 2.2 Estatística

Quando temos  $n$  eventos que são independentes entre si a probabilidade deles ocorrerem simultaneamente é dada por:

$$P(E_1 \cap E_2 \cap \dots E_n) = \prod_{i=1}^n P(E_i) = P(E_1) * P(E_2) * \dots P(E_n)$$

## 2.3 Segurança

Segurança consiste basicamente na proteção de um bem seja ele material ou imaterial e compreende também o controle de ameaças a tal bem. Um bem material pode ser uma pessoa, uma casa e um bem imaterial seria o conhecimento ou a cultura de um povo, por exemplo.

A preocupação das sociedades, instituições e países em proteger vários tipos de bens inspirou a divisão de tais bens em diversas categorias.

- **Segurança do interior**

Esta categoria, também chamada em inglês de *Homeland security*, se refere aos esforços nacionais para prevenir ataques terroristas, reduzir a vulnerabilidade de um país ao terrorismo e minimizar os danos consequentes de ataques e desastres naturais que ocorrerem. Esta categoria foi adaptada à preocupação atual dos EUA com o risco de ataques terroristas em seu país.

- **Segurança pública**

“A Segurança Pública é uma atividade pertinente aos órgãos estatais e à comunidade como um todo, realizada com o fito de proteger a cidadania, prevenindo e controlando manifestações da criminalidade e da violência, efetivas ou potenciais, garantindo o exercício pleno da cidadania nos limites da lei.” (MINISTÉRIO..., 2013).

- **Segurança nacional**

Trata-se do estado mensurável da capacidade de uma nação superar as múltiplas ameaças ao aparente bem estar da sua população e sua sobrevivência como um Estado-nação, a qualquer momento. Isto se faz pelo balanceamento da política de estado através da governança, que pode ser guiada pela computação, empiricamente ou de outra forma, e é extensível à segurança global por variáveis externas ao governo ([PALERI, 2008](#)).

- **Segurança física**

Segurança física compreende as medidas adotadas para negar acesso não autorizado a instalações, equipamentos e recursos, e proteger o pessoal e propriedade contra perdas e danos provocados por espionagem, roubo, ataques terroristas e desastres naturais. Traduzido e adaptado de ([HEADQUARTERS, 2001](#)).

Além das categorias citadas acima também existe o que se decidiu chamar de segurança virtual e é a esta categoria de segurança que é dedicada à [seção 2.4](#).

## 2.4 Segurança virtual

Esta seção aborda alguns conceitos e possui definições envolvendo segurança virtual que serão utilizados na classificação dos tuítes. Alguns dos conceitos e definições desta seção são baseados em ([FUTURE, 2011](#); [SECURITY, 2010](#); [WILSHUSEN, 2013](#); [CERT.PT; CSIRT, 2012](#); [SHIRLEY, 2007](#); [WILSHUSEN, 2011](#); [GLOSSARY... , 2009](#)).

Segurança virtual consiste na prevenção de dano, proteção contra uso não autorizado, exploração e também envolve a restauração de dados, sistemas de comunicação e de informação para garantir confidencialidade, integridade e acessibilidade de dados e digitais programas de computador.

Um incidente de segurança virtual (ISV) pode ser considerado como um evento adverso, confirmado ou sob suspeita, que tem por consequência o acesso, extração, manipulação ou danificação da integridade, confidencialidade, segurança ou acessibilidade de dados ou programas de computador, sejam públicos ou privados, sem autorização legal. Um evento pode ser causado intencional ou não intencionalmente, ter um alvo específico ou vago, e pode fazer uso de variadas técnicas. Ele pode surgir a partir de diferentes fontes, incluindo um país fazendo espionagem ou guerra de informações contra outros países, criminosos, crackers, programadores de vírus, terroristas entre outros.

ISVs não intencionais podem ser causados por erro ou omissão humana e falhas de equipamentos. Por exemplo, a a operação de um sistema por funcionários displicentes ou sem treinamento, atualizações de programa de computador, realização de manutenções entre outros. Estes são exemplos de fontes de ISVs não intencionais que podem danificar



dados ou provocar interrupções ou mau funcionamento de sistemas. ISVs intencionais são provocados por um ente inteligente, como um cracker ou organização criminosa, e incluem ataques com alvo específico ou 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, a princípio, claro como é o caso de um vírus, *worm* ou *malware* liberado na internet sem alvo específico.

No contexto de segurança virtual um ataque consiste na tentativa de destruir, expor, alterar ou incapacitar algum software, sistema e ou dados contidos neles, ou qualquer outra falha de segurança em dispositivos eletrônicos (GLOSSARY..., 2009).

Há algumas formas de estruturar a classificação dos eventos e incidentes como em (WILSHUSEN, 2013; CERT.PT; CSIRT, 2012). Neste trabalho vamos adotar a classificação de ISV utilizada em (CERT.PT; CSIRT, 2012). Existem várias classes e tipos de incidentes que agrupam tipos de eventos. Para conhecer mais os tipos de eventos veja (CERT.PT; CSIRT, 2012).

Define-se um alerta de segurança virtual (ASV) como um aviso, geralmente de caráter urgente, sobre a ameaça, ocorrência, uma notícia de solução para, uso de ferramenta para, ou a explicação de como gerar um ISV.

Um exemplo de ASV pode ser visto no tuíte<sup>1</sup> a seguir:



Figura 1 – Exemplo de Alerta de segurança virtual

<sup>1</sup> Todos os tuítes deste trabalho seguem as regras de publicação de tuítes em trabalhos segundo o Twitter. Ver <<https://twitter.com/logo>> seção: Offline (static uses and publications)

## 2.5 Redes sociais e Twitter

Esta seção faz uma introdução ao tema das redes sociais e coloca em destaque o Twitter, que é a fonte dos dados utilizados no trabalho. Assim, esta seção explica o que é um tuíte, qual o seu conteúdo, e como eles se propagam dentro do Twitter. As fontes utilizadas na escrita desta seção foram (BOYD; ELLISON, 2007; TWITTER. . . , 2013).

Redes sociais são serviços hospedados na Internet que permitem a indivíduos construir um perfil, se expressar a outros indivíduos com os quais ele possui alguma conexão na rede social, visualizar sua lista de conexões e participar de outras listas que tenham sido criadas por outros. As conexões podem ser os amigos em uma rede social e as listas podem ser conjuntos de integrantes de um grupo, de evento etc. Vale mencionar que a classificação de um relacionamento em uma rede social como sendo o de amizade não significa que os usuários realmente sejam amigos no sentido denotativo da palavra.

A visibilidade do conteúdo gerado ou compartilhado por usuários depende das restrições impostas pela rede social e pelos usuários. Por exemplo, o perfil do usuário pode ser total ou parcialmente público e a sua visibilidade na Internet depende de como o provedor do serviço controla as informações da rede social. Em outras palavras, o perfil do usuário pode não ser visível a todos na Internet, ou seja, aos indivíduos que não são usuários da rede social na qual o perfil foi criado. O mesmo acontece com outros tipos de conteúdo publicados na rede social. O conteúdo pode ser público para os integrantes da rede social, mas invisível a não usuários dela, também como pode ser restrito a determinadas conexões ou a todos os usuários que não possuem conexão com o divulgador do conteúdo.

As conexões podem ser estabelecidas de forma unidirecional e bidirecional. Ou seja, uma conexão bidirecional depende da aprovação de ambos os usuários envolvidos nela a respeito do status do relacionamento enquanto que a unidirecional depende apenas da vontade de um usuário. Em algumas redes sociais para duas pessoas serem amigas é necessário o consentimento de ambas, mas um usuário pode liberar o acesso do conteúdo que ele publica a outros usuários sem que estes façam o mesmo.

As comunicações entre conexões podem ser feitas de várias formas. Entre elas, existe a troca de mensagens visível a todos os usuários da rede, aos usuários conectados a pelo menos um dos participantes da troca de mensagens e apenas entre os participantes da troca de mensagens. O tipo de conteúdo publicado varia desde texto em língua natural até foto, áudio, vídeo entre outros tipos de conteúdo. Finalmente, como as redes sociais possuem o intuito de serem o mais acessíveis possível, elas possibilitam o seu uso por meio de computadores de mesa, notebooks, *smartphones* e até aparelhos celulares comuns.

O que torna as redes sociais únicas é o fato de que elas não apenas permitem que indivíduos conheçam estranhos, mas também tornam possível aos usuários se comunica-

rem e tornar visíveis seus grupos de conexões. Ao se expressar dentro da rede social o conteúdo gerado pelo usuário pode se dispersar entre todos os seus grupos de conexões e isto possibilita a criação de conexões entre indivíduos, que não seriam possíveis de outra forma. Apesar de, em geral, não existir o objetivo de se criar tais conexões a publicação de conteúdo na rede social faz com que as conversas travadas com os outros usuários da rede social tenham como consequência a criação de novas conexões na rede social devido à existência de interesses em comum entre indivíduos que não se conhecem pessoalmente.

Porém, vale ressaltar que os usuários, em várias das redes sociais, não estão necessariamente buscando fazer troca de conhecimento com pessoas que possuem interesses em comum nem buscando criar novas conexões. Na verdade, elas estão basicamente se comunicando com as suas redes de conexões, ou contatos.

### 2.5.1 Twitter

O Twitter é uma rede social que funciona também como microblog permitindo aos usuários lerem e enviarem tuítes. Tuítes são mensagens de texto com até 140 caracteres. Os tuítes podem ser enviados por meio de aplicativos para *smartphones*, página na Internet ou por SMS, em alguns países. Os tuítes dos usuários, por padrão, são visíveis a qualquer um que tenha acesso a Internet, mas seu acesso também pode ser limitado apenas aos usuários conectados ao usuário que envia tais mensagens, os seguidores. Um seguidor no Twitter é uma conexão que possui permissão para ler as mensagens de um dado usuário e permite a esse usuário enviar ‘mensagens diretas’ ao seguidor. Se um usuário quiser ele pode deixar de seguir alguém ou bloquear algum seguidor. ‘Mensagens diretas’ são tuítes que apenas o remetente e o destinatário podem ver. Usuários podem ser pessoas, empresas, instituições etc.

Os tuítes podem ser agrupados por assunto utilizando *hashtags* - palavras ou frases precedidas de uma cerquilha ‘#’. Da mesma forma, é possível mencionar um usuário em um tuíte usando um ‘@’ sucedido do nome de um usuário (sem espaços). Isto significa que o usuário mencionado poderá ver tal tuíte. Uma resposta é um caso particular de menção em que o tuíte começa com o ‘@’ seguido do nome do usuário ao qual se está respondendo. Além disso, também é possível replicar um tuíte contanto que o seu dono não tenha limitado seu acesso apenas a seus seguidores. Os tuítes replicados possuem o acrônimo RT (*retweet*) seguido pela menção do usuário que originalmente escreveu o tuíte. Outra característica comum nos tuítes é o uso de URLs curtas devido ao limite do número de caracteres de um tuíte. Caso o usuário esteja escrevendo um tuíte com uma URL não encurtada (mais de 20 caracteres) o Twitter utiliza seu próprio encurtador. Assim, o tamanho do tuíte é contado de acordo com o número de caracteres da URL encurtada em substituição à original.

Eis um exemplo de um tuíte que reúne todos os tipos de metadados supracitados:



Figura 2 – Exemplo de tuíte com os tipos de metadados mais comuns no Twitter

Além do tuíte clássico de 140 caracteres também é possível publicar tuítes expandidos. Estes tuítes podem conter fotos, vídeos e cartões. Estes conteúdos multimídia podem ser adicionados usando a própria plataforma do Twitter, para fazer o *upload* de fotos, ou aplicativos como o <https://vine.co/> para inserir vídeos no tuíte. O conteúdo multimídia é disponibilizado via um link, respeitando o limite de caracteres de um tuíte. Um cartão de tuíte ou *tweet card* é um conteúdo multimídia expandido utilizado para exibir fotos, vídeos, propaganda, resumo de notícias etc. O cartão é gerado a partir da inserção de alguns metadados no tuíte que permitem visualizar o conteúdo. Note que este é um meio adicional utilizado para publicar fotos e vídeos no Twitter, mas não é o único.

## 2.6 Aprendizagem Computacional

Aprendizagem de Máquina é uma subárea de Inteligência Artificial cujo objetivo é construir sistemas que são treinados a executar uma dada tarefa aperfeiçoando o seu desempenho conforme ganham experiência em realizar tal tarefa. 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.

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. Conforme o sistema possui mais mensagens spam ou não-spam (dados) ele consegue separar de forma cada vez mais próxima do ideal as mensagens spam das 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 recebidas por um usuário que irão para a caixa de entrada e as que irão para a pasta de spam.

Também é possível definir a aprendizagem de máquina de forma mais geral: DANIEL: quero deixar como está abaixo, mas vou tirar anotação de tabela =p

*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)

Em aprendizagem há algumas formas de aprendizado. Para efeitos de simplificação divide-se grosseiramente a aprendizagem em supervisionada e não-supervisionada. Na aprendizagem supervisionada o sistema recebe um conjunto de dados de treinamento e outro de testes, ambos separados por categorias, as classes. O conjunto de treinamento é utilizado para aprender a identificar as categorias dos dados e o conjunto de testes é utilizado pra testar se as categorias dos dados são identificadas corretamente. A detecção de spam é um caso de aprendizado supervisionado.

O aprendizado não-supervisionado recebe um conjunto de dados não categorizados e tenta separar os dados em grupos cujos dados compartilham características próprias do grupo. Como as categorias dos dados não são conhecidas é difícil avaliar o desempenho do sistema na separação dos dados.

Classificação: Dado um conjunto de classes busca-se determinar a qual classe um objeto pertence. Um problema de classificação possui 2 ou mais classes e embora na maioria dos problemas um objeto pertença a apenas um classe também é possível atribuir mais de uma classe a cada objeto. Um modelo de aprendizagem de máquina que faz a classificação de objetos é chamado de classificador.

Categorização é o ato de atribuir uma classe a cada objeto utilizado na construção do classificador, o que envolve o conjunto de treinamento, teste e validação do classificador.

Daniel, a definição dos itens abaixo está incompleta.

Um problema de classificação possui 6 fases:

- A aquisição dos dados consiste na coleta e categorização dos dados.
- Pré-processamento é a fase em que são removidos os
- Extração das características
- Seleção do modelo de aprendizagem

- Treinamento do classificador é fase em o modelo adotado irá utilizar o conjunto de treinamento para aprender os critérios de decisão irá utilizar para realizar a classificação.
- Ajuste das características e validação do classificador

(construir diagrama e colocar abaixo dos itens acima)

Tabela 2 – Matriz de confusão

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

Precisão:

$$\frac{pv}{pv + fp}$$

Por exemplo de documentos classificados como pertencentes à classe x e que pertencem a x / número de documentos que foram classificados como pertencentes a uma classe x.

Recuperação (*Recall*):

$$\frac{pv}{pv + fn}$$

Por exemplo de documentos classificados como pertencentes à classe x e que pertencem a x / de documentos pertencentes à classe x

Acurácia:

$$\frac{pv + nv}{pv + fp + fn + nv}$$

Por exemplo: de documentos classificados como pertencentes à classe x e que pertencem a x + de documentos classificados como não pertencentes à classe x e que não pertencem a x / todos os documentos Em outras palavras é o número de documentos classificados corretamente

## 3 Desenvolvimento

### 3.1 Classes do problema

### 3.2 Coleta dos dados

### 3.3 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 tuítes possui uma detecção que não funciona muito bem para identificar o idioma em que o tuíte foi escrito. Por isso foi utilizado um classificador que identifica o idioma do tuíte.

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

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

### 3.4 Seleção de conjunto de teste e treinamento

Como a ordem dos termos na sacola de palavras não importa, então qualquer ordenação de um conjunto de termos possui a mesma probabilidade de ocorrência. Assim, temos que a probabilidade de ocorrência de tais termos segue a distribuição multinomial.

### 3.5 Seleção das características do problema





## 4 Experimentos, discussões e resultados



## 5 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 17.
- CERT.PT; CSIRT, R. N. *Taxonomia Comum para a Rede Nacional de CSIRTs*. 2012. <[www.cert.pt/images/docs/Taxonomiav2.5.pdf](http://www.cert.pt/images/docs/Taxonomiav2.5.pdf)>. Citado na página 17.
- 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 17.
- GLOSSARY of IT Security Terminology Terms and definitions. [S.l.]: TeleTrusT Germany, 2009. <[http://www.teletrust.de/uploads/media/ISOIEC\\_JTC1\\_SC27\\_IT\\_Security\\_Glossary\\_TeleTrusT\\_Documentation.pdf](http://www.teletrust.de/uploads/media/ISOIEC_JTC1_SC27_IT_Security_Glossary_TeleTrusT_Documentation.pdf)>. Citado na página 17.
- HALL, M. et al. The WEKA data mining software: an update. *SIGKDD Explorations*, v. 11, 2009. Issue 1. Nenhuma citação no texto.
- HEADQUARTERS, U. S. D. of A. *Field Manual 3-19.30: Physical Security*. 2001. <<http://www.globalsecurity.org/military/library/policy/army/fm/3-19-30/ch1.htm>>. Capítulo 1. Citado na página 17.
- MINISTÉRIO da Justiça - Órgão de Segurança. 2013. <<http://portal.mj.gov.br/>> em Órgãos de Segurança, Conceitos Básicos. Acessado: 26-10-13. Citado na página 17.
- MITCHELL, T. M. *Machine Learning*. 1. ed. New York, NY, USA: McGraw-Hill, Inc., 1997. ISBN 0070428077, 9780070428072. Citado na página 18.
- PALERI, P. National security: Imperatives and challenges. In: \_\_\_\_\_. [S.l.]: Tata McGraw-Hill, 2008. p. 57. ISBN 9780070656864. Citado na página 17.
- 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 15.
- SECURITY, U. D. of H. *Privacy Impact Assessment for the Initiative Three Exercise*. [S.l.], 2010. Citado na página 17.
- SHIRLEY, R. *Internet Security Glossary*. 2007. <<http://www.ipa.go.jp/security/rfc/RFC4949-00EN.html>>. Citado na página 17.
- STRATEGIC, C. for; STUDIES, I. *The Economic Impact of Cybercrime and Cyber Espionage*. [S.l.], 2013. Citado na página 15.
- TWITTER Help Center - Get started: FAQs and the basics. 2013. <<https://support.twitter.com/groups/50-welcome-to-twitter>>. Acessado: 28-09-13. Citado na página 17.

WILSHUSEN, G. C. *Cybersecurity: Continued Attention Needed to Protect Our Nation's Critical Infrastructure*. [S.l.], 2011. Citado na página 17.

WILSHUSEN, G. C. *Cyber Threats Facilitate Ability to Commit Economic Espionage*. [S.l.], 2013. Citado na página 17.