São Paulo, Brazil
October 29-30, 2018

The Tenth International Conference on
FORENSIC COMPUTER SCIENCE and CYBER LAW

www.ICoFCS.org

DOI: 10.5769/C2018005 or http://dx.doi.org/10.5769/C2018005

# Detecção de Cibercrime em Redes Sociais: *Machine Learning*

Jackson Mallmann[1], Alex dos Santos Xavier[2], Altair Olivo Santin[3]

*(1) Instituto Federal Catarinense, jackson.mallmann@ifc.edu.br*
*(2) Pontifícia Universidade Católica do Paraná (PUC-PR), alex.xavier@ppgia.pucpr.br*
*(3) Pontifícia Universidade Católica do Paraná (PUC-PR), altair.santin@pucpr.br*

Resumo: O presente artigo expõe um trabalho de desenvolvimento que utiliza técnicas de Machine Learning para realizar a detecção de cibercrime em mensagens publicadas na rede social Twitter. As mensagens são pré-processadas, constrói-se um dicionário, e no uso do software WEKA são aplicas técnicas de agrupamento e de classificação (K-means, SVM, DT and NB) para detecção de cibercrimes. Em resultados experimentais afirma-se que o uso de Machine Learning foi essencial para o sucesso deste trabalho, sendo que o classificador SVM apresentou 98,77% de acertos na classificação de cibercrime.

Palavras-Chave: Machine Learning; Cibercrime; Redes Sociais, Agrupamento; Classificação.

Abstract: The proposed article presents an academic work which uses Machine Learning techniques to perform detection of cybercrime in messages posted in online social networks, based on tweets datasets. Through a pre-processed step to build a dictionary and using WEKA software, was applied the classify and cluster techniques (K-means, SVM, DT and NB) to detect cybercrimes. The experimental results showed that the use of Machine Learning was essential to the success of this work, where SVM classifier produced 98.77% of accuracy in detection.

Keywords: Machine Learning; Cybercrime; Social Networks; Cluster; Classify.

## I. Introdução

O uso de serviços virtuais também tem sido alvo de criminosos para o cometimento de delitos que com o advento da Internet, passaram a se chamar de cibercrimes [3,4,13,15]. Ódio, racismo e preconceito de gênero são alguns exemplos, sendo que o número de denúncias que envolvem cibercrimes está em gradativo aumento conforme a SaferNet[1].

Por sua vez, cibercrime pode ocorrer mediante conteúdo "impróprio" hospedado em URL (*Uniform Resource Locator*), como nos casos das redes sociais Facebook[2], Twitter[3], Instagram[4] e

1 new.safernet.org.br/denuncie#

2 facebook.com

3 twitter.com

4 instagram.com

YouTube[5]. E na ocorrência de cibercrime, é necessário que profissionais realizem uma investigação, respeitando-se questões de legislação, comumente chamado de constatação da autoria do cibercrime [5,7,11].

Desta forma, objetiva-se apresentar um trabalho de desenvolvimento que visa auxiliar na detecção do cibercrime e que automatize sua comprovação. Para tal, apresenta-se a realização dos experimentos através do uso de mensagens publicadas e extraídas da rede social Twitter [18].

Este trabalho se baseia no ambiente da Forense Computacional, a qual possui procedimentos e se utiliza de ferramentas computacionais com o objetivo do encontro de rastros digitais. Ainda, a justificativa para a realização deste trabalho baseia-se no trabalho de [1], que afirmam haver a necessidade da existência de métodos automáticos para detecção de cibercrimes. Ademais, contribui-se socialmente com a geração de um mecanismo que proporcione a formalização de investigações policiais.

O trabalho está organizado da seguinte forma: na Seção I se contextualiza o presente trabalho; na Seção II é descrito conceitos relativos a *Machine Learning* que são aplicados para a detecção de cibercrime; na Seção III detalha-se os trabalhos relacionados que realizam a detecção de cibercrime em diferentes contextos; na Seção IV apresenta-se a metodologia para a obtenção de resultados, já que na Seção V se expõe os resultados obtidos e de detalhes técnicos referentes aos experimentos; e por fim, na Seção VI é exposto as conclusões e direções futuras.

# 2. Fundamentação

*Machine Learning* é uma subárea em que usuários, no uso de algoritmos realizam o treinamento de sistema informático, e que posteriormente realizará classificações. No treinamento é aplicado análise de dados. Os dados são representados por atributos, instâncias e classes. Os atributos são as características dos dados. Por sua vez, as instâncias são conhecidas como registros, e compostos por atributos, sendo que um dos atributos da instância visa caracterizar a instância, ou seja, a classe a que pertence a instância [20].

Após o treinamento pode-se realizar classificações, comumente chamado de predições. A predição pode utilizar diferentes métricas, como por exemplo a acurácia (ACC).

Esta métrica é baseada na somatória da taxa de categorização correta e dividida pela somatória de todas as classificações (corretas, incorretas e aquelas não classificadas). Esses valores podem ser retirados da Matriz de Confusão (*Confusion Matrix*). Assim é possível realizar a comparação entre diferentes técnicas na solução de um mesmo problema [20].

Várias técnicas de *Machine Learning* podem ser aplicadas na detecção de cibercrime, como por exemplo, a classificação e o agrupamento, os quais serão descritos nas próxima Subseções.

### *A. Classificação*

A classificação objetiva predizer o valor de um atributo mediante uma aprendizagem supervisionada [16], já que a classificação necessita de treinamento por parte do usuário.

Existem vários algoritmos de classificação. Os mais usuais: SVM (*Support Vector Machine*), NB (*Naïve Bayes*) e DT (*Decision Tree*) [16].

### *B. Agrupamento*

O agrupamento tem o objetivo de classificar dados em grupos que tenham padrões comuns. Os grupos não são previamente padronizados (treinados), e por isso, a técnica de agrupamento é considerada uma técnica não supervisionada.

O funcionamento desta técnica é baseado na execução de um algoritmo. Por exemplo, o agrupamento é utilizado onde é necessário a determinação de padrões que simbolizam um

5 youtube.com

conjunto de textos. Citam-se como exemplo os algoritmos K-means, DBSCAN e Cobweb [6].

Sendo assim, na próxima Seção apresenta-se o estudo de trabalhos relacionados com a detecção de cibercrimes.

# 3. Trabalhos Relacionados

No trabalho de [2] é apresentado o resultado da aplicação de e-mails em 3 diferentes tipos de técnicas de agrupamento: K-means, fuzzy c-means e *neural networks backpropagation*. Foi utilizado uma base de e-mails disponibilizada pela Enron's e-mail *dataset*, e desta forma, analisados 1412 e-mails divididos em: 452 com conteúdo de análise sentimental negativa, 537 com conteúdo de análise sentimental positiva, e 423 eram com conteúdo de análise sentimental neutra. Para análise dos e-mails foi realizado um pré-processamento dos e-mails onde foram posteriormente extraídas as características para a criação de um vetor de características. Após o pré-processamento os e-mails foram submetidos as técnicas de agrupamento, em que o neural *networks backpropagation* conseguiu o melhor resultado: 97,91% de reconhecimento. Além do mais, os autores afirmam que e-mails com texto que possuam sentimento negativo em seu texto podem ser usados como gerador de evidência para casos de cibercrimes.

Embora esse trabalho seja interessante, ele não exibe como os autores fizeram os cálculos. Além de não apresentarem com clareza como foi realizada a extração das características. E, não informam de que forma os algoritmos foram analisados (mediante a implementação de um framework, ou outro).

Em [1,12], os autores apresentam trabalhos que utilizam técnicas de *Machine Learning* (SVM, NB e AdaBoost) mediante utilização do *software* WEKA [19] para classificação de tweets. A classificação foi realizada em dois tipos de classes. Para realização dos experimentos, tweets foram coletados e selecionados em categorias. Após isso, todos os tweets selecionados foram pré-processados (remoção de anotações, remoção de re-tweets, e remoção de URL´s), para que os atributos pudessem ser extraídos e representados em vetores. Optaram pelo uso de atributos estilométricos, atributos baseados no tempo, e atributos baseados em sentimento. Em [12] apresentam como resultado palavras e uma lista de verbos utilizadas tendo como base os tweets analisados naquele trabalho. E em [1], resultados da aplicação das técnicas de *Machine Learning* via respectivos algoritmos: SVM com 99,1%, NB com 99,9% e AdaBoost com 100% de eficácia.

Os trabalhos [1,12], embora interessantes, divulgaram apenas um dicionário com as principais palavras utilizadas. Entretanto os tweets analisados não foram divulgados, nem tão pouco os dados e/ou aplicativo(s) utilizado(s) para o pré-processamento dos tweets, método para extração das palavras e verbos utilizados, o que dificulta a replicabilidade deste trabalho.

Dando sequência no assunto, no trabalho de [8], objetivou-se determinar se mensagens publicadas em rede social Facebook é proveniente do usuário proprietário do perfil ou de um intruso. Para os experimentos, os autores fizeram a coleta de mensagens de 30 usuários do Facebook, sendo que cada mensagem tinha uma média de 20,6 palavras (103 caracteres). Após fizeram o pré-processamento, representação em vetores de atributos e aplicação de técnica de *Machine Learning* (SVM *light*), em que obtiveram 79,6% de acertos.

Os trabalhos relacionados e estudados podem ser divididos em três grupos: no primeiro é apresentado em [2] a análise de e-mails, sendo que os autores concluíram que e-mails com teor negativo podem ser usados como gerador de evidência de cibercrime; no segundo, [1,12], faz-se a classificação de tweets, e no terceiro, trabalho de [8], apresenta-se uma maneira de evitar a ocorrência de cibercrime.

Sendo assim, finaliza-se esta Seção, onde foram apresentados trabalhos que obtiveram resultados aceitáveis via aplicação de técnicas de *Machine Learning* e que são utilizados na detecção de cibercrime, objeto de estudo do atual trabalho.

# 3. Metodologia

Objetiva-se apresentar resultados da detecção de cibercrime em redes sociais. Para isso, utilizou-se a metodologia de desenvolvimento [9]. Classificam-se tweets em cibercrime, e não cibercrime.

A proposta envolve a utilização de duas bases de dados: a primeira são de tweets públicos e disponíveis, a qual foi denominada de cibercrime. Por sua vez, a segunda, tweets de personalidades que não tem contato com o suposto cibercrime, entretanto não são criminosos, e, por questões éticas não serão divulgados.

Na Figura 1 é apresentado um fluxo da proposta deste trabalho. Após a coleta dos tweets, os mesmos são importados e pré-processados: transforma-se todo o conteúdo do texto em sua forma maiúscula, retiram-se todos os espaços, acentos, símbolos HTML, números, re-tweets, e links.

Na sequência as duas bases são representadas em dois vetores (Extração de Atributos e Representação em Vetores). No primeiro, foi verificado se as palavras de maior incidência existentes nas duas bases, e no segundo, tendo-se um dicionário de palavras como referência.

Este dicionário foi formalizado mediante contabilização das palavras com maior ocorrência na base cibercrime, totalizando 775 palavras (Figura 2 – Attributes).

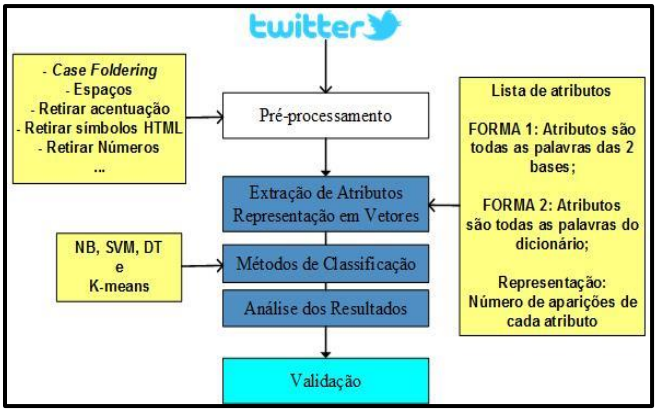


Figura 1. Fluxo.

A extração de atributos, representação em vetores, aplicação dos métodos de agrupamento/classificação, análise de resultados e validação foram formalizados no uso do *software* WEKA versão 3.8.1 [19].

Posteriormente foram gerados dois arquivos com extensão ARFF. Ambos foram submetidos a técnica de agrupamento (K-means), e as técnicas de classificação (NB, SVM e DT).

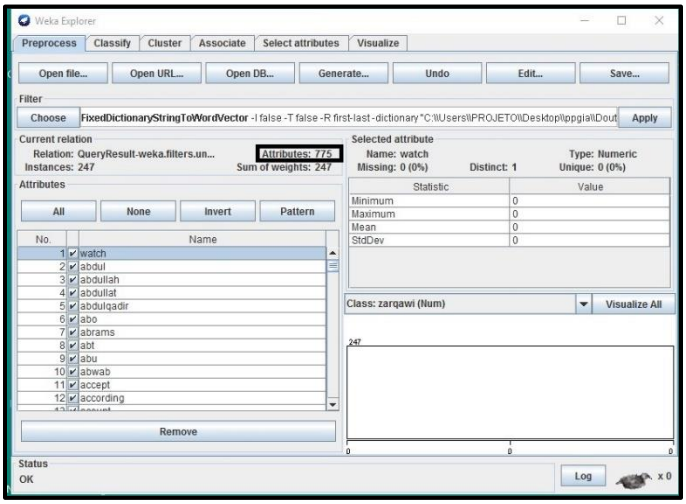


Figura 2. Aplicação de Filtros – *Software* WEKA.

Na sequência, os resultados provenientes dos Métodos de Classificação são analisados e validados, sendo que na próxima Seção, é apresentado os resultados dos experimentos de agrupamento/classificação, assim como da etapa de Análise de Resultados e Validação.

# 4. Resultados e Discussão

A primeira base, denominada de base cibercrime, está disponível no site www.kaggle.com e é composta de 27.182 tweets, e a segunda base (2.354 tweets) que foi coletada mediante o uso de um aplicativo implementado em linguagem Python [14] com o uso da biblioteca Tweepy [17], o qual pode-se visualizar parcialmente na Figura 3. Ambas as bases estavam traduzidas em língua inglesa.

Após a coleta, todos os tweets foram armazenadas no banco de dados MySQL, e então pré-processados. Como resultado, restaram apenas palavras nos tweets: 168 tweets da primeira base, e 76 da segunda, ou seja, um total de 244 instâncias.

Durante o pré-processamento das bases, verificou-se que muitas linhas (geralmente re-tweets) estavam em branco ou que continham apenas um valor que não tinha representatividade. Para tal, optou-se em excluir estas linhas.

```
#!/usr/bin/env python
# encoding: utf-8

import tweepy #https://github.com/tweepy/tweepy
import csv

#Twitter API credentials
consumer_key = "--------"
consumer_secret = "--------"
access_key = "--------"
access_secret = "--------"


def get_all_tweets(screen_name):

    #authorize twitter, initialize tweepy
    auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
    auth.set_access_token(access_key, access_secret)
    api = tweepy.API(auth)

    #initialize a list to hold all the tweepy Tweets
    alltweets = []

    #make initial request for most recent tweets (200 is the maximum allowed count)
    new_tweets = api.user_timeline(screen_name = screen_name,count=200)
```

Figura 3. Código Parcial Python.

Por sua vez, aplicou-se na sequência a extração de atributos e geração de vetores, ou seja, a construção de arquivos com extensão ARFF. Para esta formalização, aplicaram-se 3 classes disponíveis no software WEKA (NominalToString, FixedDictionaryStringToWordVector e Add).

Com o primeiro é realizado a conversão de atributos nominais para *strings*. Posteriormente, no segundo, converte-se os atributos *string* em um conjunto de atributos que representam informações de ocorrência de palavras. Por fim, no terceiro, adiciona-se um novo atributo ao conjunto de dados, como por exemplo, a classe.

Foi realizado experimento no uso das técnicas de classificação e de agrupamento. São elas: K-means, SVM, NB e DT. Para todas elas, utilizou-se a opção de treinamento (*Use training set*). Na Tabela I apresenta-se os resultados da aplicação das técnicas em cada arquivo ARFF gerado. Analisando a Tabela I, verifica-se que no uso do arquivo ARFF contendo todas as palavras do dicionário proporcionou melhores resultados do que no uso das palavras de todas as palavras.

| | K-means |
|---|---|
| 244 tweets (atributos: todas as palavras) | 71,72% |
| 244 tweets (atributos: palavras do dicionário) | **97,16%** |

Tabela I. Taxa de Acertos – Agrupamento.

Na Tabela II é apresentado os resultados da aplicação das técnicas de classificação. Foram utilizados os algoritmos: SVM, NB e DT para os dois arquivos.

| | **SVM** | **NB** | **DT** |
|---|---|---|---|
| 244 tweets (atributos: todas as palavras) | 97,12% | 92,31% | 93,12% |
| 244 tweets (atributos: palavras do dicionário) | **98,77%** | 85,65% | 86,47% |

Tabela II. Taxa de Acertos – Classificação.

Na Figura 4 expõem-se detalhes do melhor resultado: SVM (98,77%): aplicação do algoritmo SVM no arquivo ARFF palavras do dicionário.

```
=== Summary ===

Correctly Classified Instances         241               98.7705 %
Incorrectly Classified Instances         3                1.2295 %
Kappa statistic                          0.971
Mean absolute error                      0.0123
Root mean squared error                  0.1109
Relative absolute error                  2.8627 %
Root relative squared error             23.9437 %
Total Number of Instances              244

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 1,000    0,039     0,982      1,000    0,991     0,971    0,980     0,982     Cibercrime
                 0,961    0,000     1,000      0,961    0,980     0,971    0,980     0,973     Naocibercrime
Weighted Avg.    0,988    0,027     0,988      0,988    0,988     0,971    0,980     0,979

=== Confusion Matrix ===

   a   b   <-- classified as
 168   0 |   a = Cibercrime
   3  73 |   b = Naocibercrime
```

Figura 4. WEKA - SVM (98,77%).

Comparando com os resultados proporcionados por [1], onde houve um alcance de 100% de eficácia, o que é questionável. Entretanto neste trabalho foi obtido 98,77% no uso do classificador SVM. Os resultados demonstram ser possível o uso de *Machine Learning* na detecção de cibercrimes, assim como parcialmente realizado em [1].

Também se verifica que os tweets que não continham característica de cibercrime (base 2, ou tweets não cibercrime), somente 3 foram classificados incorretamente, além do que, não é surpresa a classificação dos 168 tweets contendo

cibercrime de forma correta, já que as palavras do dicionário foram obtidas a partir da base cibercrime.

## 4. Conclusões e Direções Futuras

Com os experimentos desenvolvidos objetivou-se apresentar e comprovar o auxílio na detecção de cibercrime mediante o uso de técnicas de *Machine Learning*. Para tal, foram aplicados os algoritmos: SVM, NB, DT e K-means. Baseado nos resultados proporcionados nos experimentos apresentados neste trabalho, verificou-se que o algoritmo SVM propositou os melhores resultados, alcançando uma taxa de acerto de 98,77%, quando foi utilizado as palavras do dicionário como atributo.

Os resultados dos experimentos podem ser melhorados. Para tal, como trabalhos futuros, apresenta-se a necessidade de melhorar a quantidade da amostra utilizada, melhorar os atributos, refazer os experimentos no uso de outro aplicativo, como por exemplo, o *software* MOA, ou mesmo através de implementação própria dos algoritmos de classificação/agrupamento.

A principal limitação para fazer esta análise é que o Twitter possui poucas palavras em cada "postagem". É evidente a necessidade da diminuição do pré-processamento realizado nos experimentos, entretanto salienta-se de que, para detecção e comprovação de cibercrime, em especifico da rede social Twitter é necessário investigar primeiramente a mensagem principal ("postagem"), pois em várias situações, o cibercrime é constatado pela mensagem principal, e não por re-tweets.

Leitores que desejarem o código fonte, scripts com os códigos para pré-processamento dos tweets, aquisição dos arquivos ARFF, entre outros, devem solicitar pelo e-mail dos autores.

Ademais, realizada análise dos resultados corrobora-se com a detecção de cibercrime. Todas as mensagens de cibercrime foram classificadas corretamente, uma vez que o dicionário foi formalizado pelas mensagens pertencente a essa classe.

## References

[1] Ashcroft, M.; Fisher A.; Kaati L.; Omer E.; Prucha N. (2015) "Detecting Jihadist Messages on Twitter". European Intelligence and Security Informatics Conference – IEEE Computer Society.

[2] Bogawar, P. S. Mrs.; Bhoyar, K. K. (2016) "Soft Computing Approaches to Classification of Emails for Sentiment Analysis", International Conference on Informatics and Analytics (ICIA-16), Agosto.

[3] Broadhurst, R. (2006) "Developments in the global law enforcement of cyber-crime", Policing: An Int. Journal of Police Strategies & Management. V. 29, N. 3, p.408-433.

[4] Cia, S. Ó. (2004) "An Extended Model of Cybercrime Investigations", Int. Journal of Digital Evidence, V. 3, 1ª. Edição.

[5] Delmanto, C. (2000) "Código penal comentado", Editora Renovar. 5ª. Edição atual. e ampl., Rio de Janeiro.

[6] Jain, A. K.; Murty, M. N.; Flynn, P. J. (1999) "Data Clustering: A Review", ACM Computing Surveys, V. 31, N. 3.

[7] Jesus, D. E. (2002) "Direito Penal", Editora Saraiva. São Paulo.

[8] Li, J. S.; Monaco, John V.; Chen, L.; Tappert, C. C. (2014) "Authorship Authentication Using Short Messages from Social Networking Sites", International Conference on e-Business Engineering, IEEE 11th.

[9] Maren, V. D. – J. M. (1999) "Méthodes de recherche pour l´Éducation", Montréal: De Boeck, 1999.

[10] MySQL. (2018) "Banco de dados de código aberto", https://www.mysql.com/.

[11] Noronha, E. M. (1998) "Curso de Direito Processual Penal", Editora Saraiva. 26ª. Edição, São Paulo.

[12] Omer, E. (2015) "Using machine learning to identify jihadist messages on Twitter",http://uu.diva-portal.org/smash/get/diva2:846343/FULLTEXT01.pdf.

[13] Pinheiro, P. P. (2009) "Direito Digital", Editora Saraiva. 3ª. Edição, São Paulo, 2009.

[14] Python. (2018) "Linguagem de programação Python", https://www.python.org/.

[15] Surendran, A. C.; Platt, J. C.; Renshaw, E. (2005) "Automatic Discovery of Personal Topics to Organize Email", Proc. 2nd Conference on Email and Anti-Spam, CEAS.

[16] Tam, P-N.; Steinbach, M.; Kumar, V. (2005) “Introduction to Data Mining”, AddisonWesley.

[17] Tweepy (2018) “Twitter for Python!”, https://github.com/tweepy/tweepy.

[18] Twitter. (2018) “Site de rede social Tweeter”, https://twitter.com.

[19] Weka. (2018) “Data Mining Software in Java”, http://www.cs.waikato.ac.nz/ml/weka/.

[20] Witten, I.H.; Frank, E.; Hall, M.A.; Pal, C.J. (2016) “Practical Machine Learning Tools and Techniques”, Editora Morgan Kaufmann. 4th Edition. ISBN: 9780128042915.