Crimes: reportes oficiais vs. postagens no Twitter

Gilvan O. dos Reis
Instituto de Computação
Universidade Federal do Amazonas
Manaus, Amazonas, Brasil
gor@icomp.ufam.edu.br

Eduardo F. Nakamura
Instituto de Computação
Universidade Federal do Amazonas
Manaus, Amazonas, Brasil
nakamura@icomp.ufam.edu.br

ABSTRACT

The Social Network Sites (SNS) has become a major source of communication. Through SNS, we can communicate with people who are in many part of the planet and by using the information contained in these sites, we can turn users into social sensors. Based on this, this project aims to compare crime reports collected through social sensors to official police reports. As a case study, we validate our model with real data from New York City. The main results obtained were the training of a classifier with precision of 86% and a classified database with 491.415 *tweets*. Even with this result, it was not possible to find a correlation between the official dataset and our dataset.

KEYWORDS

Social Sensors, Twitter, Crime Dataset

1 INTRODUÇÃO

Os sites de redes sociais (SRS) são serviços baseado na Internet que permite que indivíduos construam um perfil público ou semipúblico, que possam articular uma lista contendo outros usuários do site com os quais se compartilham conexões ou interesses e possam visualizar as listas de conexões feitas por outros usuários na rede [2].

Atualmente, os SRS (como por exemplo, *Foursquare*[1], *Twitter*[2] e *Swarm* [3]) se tornaram uma das maiores fontes de comunicação. Através deles, podemos nos comunicar com pessoas que estão em várias partes do planeta, sendo que cerca de 2 bilhões de pessoas utilizam pelo menos um SRS[4] e a quantidade média de horas que um usuário permanece conectado em um SRS é de 3 horas [3]. A popularização dos smartphones permitiu com que estes usuários estivessem conectados a qualquer momento. Com base neste cenário que um novo campo para estudos e pesquisas pode ser criado.

Sakaki et al. [7] por exemplo, utilizaram o *Twitter*, para desenvolver um sistema que detectasse terremotos no Japão através das análises dos *tweets* (mensagens curtas de até 140 caracteres) dos usuários deste SRS. Ele considera que cada usuário do *Twitter* é um sensor e cada *tweet* é uma informação sensorial. Estes sensores são nominados pelo autor como sensores sociais.

Neste trabalho, utilizamos os dados coletadas através dos sensores socais para treinar um classificador afim de identificar se uma mensagem contém um relato de crime e comparamos os resultados desta classificação com informações oficiais da polícia de Nova Iorque. Escolhemos a cidade de Nova Iorque como estudo de caso, porque esta cidade publica os dados oficiais da policia, possui uma grande concentração de pessoas e uma grande densidade de usuários do *Twitter*. A partir deste classificador criamos uma base de relatos de crimes apenas com informações providas pelos usuários do *Twitter*.

2 TRABALHOS RELACIONADOS

França et al. [4] demonstram uma abordagem para análise do volume de dados gerados nos SRS, incluindo também a coleta, tratamento e mineração destes dados e princípios de análise de interações sociais. Os autores explanaram as APIs (Application Programming Interface) de busca de cada um dos SRS mais populares (*Facebook*[5], *Twitter*, *Youtube*[6] e *Foursquare*) além de citar as suas limitações. Os autores também apresentam algumas dificuldades que podem ocorrer com a análise de grandes volumes de dados (por exemplo, o volume massivo de dados que se deve analisar em concorrência com a capacidade de processamento do hardware utilizado) e algumas soluções que permitem tornar este trabalho mais viável.

de Oliveira et al. [1] analisam os relatos de crimes a partir dos *tweets* feitos no estado de São Paulo. Os autores realizaram um estudo bem completo sobre este tema, explanando desde conceitos básicos de aprendizagem de maquina até os conhecimentos teóricos de cada classificador utilizado. Os autores compararam três das principais técnicas de aprendizado supervisionado: Naive Bayes, Árvores de decisão e SVM. Para comparar as técnicas, foram rotulados 26.503 *tweets* e os dividiu em duas partes: a primeira versão com 5.000 postagens e a segunda versão, com 10.000 postagens. Os autores compararam as técnicas usando o método de validação cruzada de tamanho 10. Os autores concluíram que o SVM apresentou o melhor resultado.

[1] www.foursquare.com
[2] www.twitter.com
[3] www.swarmapp.com
[4] www.statista.com/topics/1164/social-networks/

In: XIV Workshop de Trabalhos de Iniciação Científica (WTIC 2017), Gramado, Brasil. Anais do XXIII Simpósio Brasileiro de Sistemas Multimídia e Web: Workshops e Pôsteres. Porto Alegre: Sociedade Brasileira de Computação, 2017.

ISBN 978-85-7669-380-2.

[5] w ww.facebook.com
[6] www.youtube.com

Sakaki et al. [7] investigaram a interação em tempo real entre os usuários do *Twitter*, para detectar eventos específicos em determinadas localizações geográficas. Eles decidiram focar em grandes eventos que possam influenciar na vida de muitas pessoas, como tornados e terremotos, pois este tipo de evento é frequente no Japão. Os autores fazem a análise morfológica dos *tweets* utilizando o software Mecab[7] e a partir deste, usam o *SVM* [6] para classificar se um determinado *tweet* está relatando determinado evento. Os autores informaram que usaram uma base de teste contendo 597 exemplos positivos. Eles também assumem que cada *tweet* pode estar associado com uma localização em um determinado tempo. Esta característica é muito importante, pois é a partir desta informação que eles podem executar um modelo probabilístico para estimar a posição do evento.

3 METODOLOGIA

Nesta seção é descrita a metodologia utilizada ao longo deste trabalho. Este processo é composto pelas seguintes fases: (i) Coleta de dados; (ii) Rotulação e (iii) Classificação.

3.1 Coleta dos dados

Os dados oficiais da polícia de Nova Iorque foram coletados através de um site governamental[8] que os disponibiliza para consulta do público. Nele é possível exportar a base oficial em vários formatos além de ser possível visualizar os dados em um mapa. Com isso, coletamos 316.977 relatos de crimes que ocorreram entre os meses de maio a dezembro de 2016.

Para realizar a coleta dos *tweets* foi utilizada a API[9] provida pelo *Twitter*. Neste caso é informado para a API uma área geográfica e uma consulta, e o resultado são *tweets* que: foram postados em até nove dias, estão dentro da área geográfica delimitada e satisfaçam esta determinada consulta.

Os autores recomendam coletar os dados no máximo a cada sete dias ao invés de nove, para ter no mínimo dois dias extras para coletar novamente os dados caso haja algum imprevisto, como problema de conexão ou problema de energia. Também, em cenários parecidos, não recomendamos salvar apenas dados geolocalizados, visto que apenas uma parte muito pouco representativa dos dados são geolocalizados (Dos 2.919.652 *tweets* coletados no período de março a dezembro de 2016 apenas 2.233, 0,076%, deles continham geolocalização), tornando assim inviável a coleta de apenas dados geolocalizados.

Neste trabalho foi utilizado a área geográfica que representa a cidade de Nova Iorque e foram utilizadas sete diferentes consultas criadas com base no *Uniform Crime Reporting Handbook*[10], sendo elas: (1) crime OR crimes, (2) robbery, (3) homicide, (4) rape, (5) assault, (6) burglary e (7) theft.

Também foi decidido analisar apenas os *tweets* coletados a partir do mês de maio até o mês de dezembro de 2016. Os *tweets* do mês de março foram excluídos por apresentarem pouca quantidade (pois neste período apenas dados geolocalizados eram coletados) e os *tweets* do mês de abril foram excluídos devido a erros no *script* de coleta que poderiam interferir na análise dos dados. Os tweets dos meses de janeiro a julho de 2017 não foram considerados neste experimento devido a uma alteração na formatação dos dados oficiais, devendo ser melhor analisada futuramente para evitar adicionar qualquer ruído no experimento.

Após realizada a coleta dos *tweets*, os dados foram pré-processados para então serem rotulados de acordo com a informação presente no *tweet*. Neste pré-processamento foi retirado todos os *retweets* (uma espécie de encaminhamento de *tweets*) visando fazer com que apenas os dados únicos sejam analisados. Para realizar esta remoção foram excluídos todos os *tweets* que possuíam o campo *retweeted_status*, que segundo a documentação da API do *Twitter* [11] afirma que este campo só é presente nos *retweets*. Vale ressaltar que esta informação não será descartada e que futuramente planeja-se utilizar deste valor para verificar a relação entre a quantidade de *retweets* e a veracidade de um determinado relato de crime.

3.2 Rotulação

Foram utilizados nove rótulos, sendo eles: **informação**, **crime**, **robbery**, **homicide**, **rape**, **assault**, **burglary**, **larceny-theft** e **motor vehicle theft**. A seguir estão as explicações dos valores possíveis em cada um dos rótulos.

No rótulo **informação** podemos ter 4 valores:

- 0: caso a informação do *tweet* refere-se a um crime.
- 1: caso a informação do *tweet* cite outra cidade ou localização fora da cidade de Nova Iorque.
- 2: caso a informação do *tweet* seja duvidosa, ou seja, existe a necessidade de olhar uma outra fonte de informação (como um link ou uma foto) para entender o significado do *tweet*. Por exemplo, "*omg that is such a crime*".
- 3: caso a informação do *tweet* não se refere a um crime. Por exemplo, "*Why not read a Fictional Forensic Accountant Crime Book. My author's page: https://t.co/yJdOcczJeR https://t.co/Izd0kuW5h2 #Ad*".

O rótulo **crime** só será rotulado caso o rótulo **informação** tenha valor igual a 0. Neste rótulo podemos ter 6 valores:

- 0: caso o rótulo **informação** tenha valor diferente de 0.
- 1: caso a informação contenha um relato de um crime.
- 2: caso a informação contenha um processo jurídico como sentenças ou prisões.
- 3: caso a informação contenha uma citação de crime de forma metáfora ou uma pegadinha. Por exemplo, "*@darrenrovell highway robbery*".
- 4: caso a informação contenha opiniões do usuário sobre algum crime.
- 5: caso a informação contenha notícias policiais que não reportam crimes.

[7] http://mecab.sourceforge.net/
[8] https://maps.nyc.gov/crime/
[9] https://dev.twitter.com/rest/public
[10] Disponível em https://www2.fbi.gov/ucr/handbook/ucrhandbook04.pdf
[11] https://dev.twitter.com/overview/api/tweets

Os demais rótulos são binários, ou seja, são rótulos que possuem apenas dois valores: 1 caso o *tweet* atenda a condição daquele rótulo e 0 caso contrário. Estes rótulos só serão analisados caso o valor do rótulo **crime** tenha valor igual a 1. As condições de cada rótulo são:

- **robbery**: caso o crime seja um roubo.
- **homicide**: caso o crime seja um homicídio.
- **rape**: caso o crime seja um assédio sexual.
- **assault**: caso o crime seja uma agressão.
- **burglary**: caso o crime seja uma invasão.
- **larcerny-theft**: caso o crime seja um furto.
- **motor vehicle theft**: caso o crime seja um roubo de veículo.

Para criar a base rotulada deste experimento foram escolhidos, de forma arbitraria, os *tweets* dos meses de julho e novembro de 2016. Neste período foram coletados 802.691 *tweets* dos quais apenas 120.199 não eram *retweets*. Destes foram rotulados um total de 7.286 *tweets*. Para prover uma análise mais linear dos dados, foram rotulados os vinte primeiros *tweets* após as 12:00h de cada dia e os vinte últimos *tweets* antes das 24:00h de cada dia.

3.3 Classificação

Antes de realizar a classificação, a base passou por um processo de filtragem, onde seus textos foram analisados e foram removidos quaisquer *hashtags*, menções a outros usuários, links e *emojis*. Depois foram removidas as *stopwords* e foi realizado um processo de *stemming* para recuperar a raiz da palavra ao retirar os prefixos e sufixos das palavras, visando melhorar as correspondências entre os textos. Além disso, foram retirados da base todos os *tweets* que continham citações de outras cidades ou países, pois esta base visa focar apenas nos relatos de crime da cidade de Nova Iorque. Após esta remoção o tamanho final da base foi de 491.414 *tweets*.

Para a classificação dos dados foi utilizado o método *Support Vector Machine* (*SVM*). Devido à baixa quantidade de *tweets* em cada rótulo, foi decidido treinar apenas o classificador do rótulo com maior número de entradas (**crime**). O vetor de características para o classificador analisar foi criado por meio do *Term Frequency - Inverse Document Frequency* (*tf-idf*) de cada palavra.

O *SVM* foi utilizado em duas variantes: classe única e duas classes. Para o *SVM* de classe única foi considerado que caso um dado não seja similar aos dados pertencentes a classe negativa, ele será considerado como sendo da classe positiva. A base de treino também teve duas variantes em relação ao seu tamanho. A primeira variante foi utilizar toda a base rotulada para treinar o classificador e a segunda foi utilizar uma base de treino balanceada, ou seja, os dados da base rotulada pertencentes à classe negativa foram escolhidos de forma randômica até que estes atingissem a mesma quantidade dos dados da classe positiva. Além disso, a base de treino também teve duas variantes em relação ao tratamento dos seus dados, onde uma foi processada pelo mesmo processo de filtragem da base de dados e outra não foi.

Após todas essas considerações foram criados seis classificadores, sendo eles os seguintes:

(1) Uma classe com base de treino completa e não processada;
(2) Uma classe com base de treino completa e processada;
(3) Duas classes com base de treino balanceada e não processada;
(4) Duas classes com base de treino balanceada e processada;
(5) Duas classes com base de treino completa e não processada;
(6) Duas classes com base de treino completa e processada.

4 RESULTADOS

Embora a consulta "rape" tenha apresentado a terceira maior quantidade de *tweets* (612.144), ela conteve apenas a segunda menor quantidade de relatos oficiais (775). Isto poderia indicar que as pessoas que sofrem um estupro preferem falar deste assunto nas redes sociais ao invés de ir fazer uma denúncia na polícia, devido a possibilidade do anonimato e de evitar o julgamento das pessoas que ela convive. Por outro lado, foram rotulados apenas 25 *tweets* como relatos de estupro (rótulo **rape**). Este número é muito pequeno se comparado aos outros rótulos da base e isto informa que a grande quantidade de dados coletados pela consulta "rape" são geralmente pessoas comentando sobre esse assunto, mas dificilmente reportando o acontecimento deste crime.

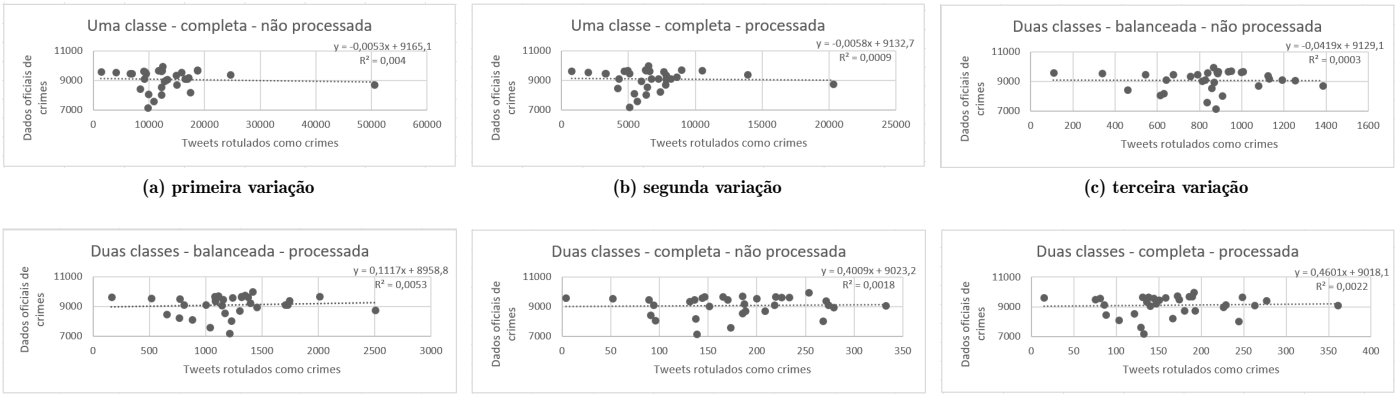
Também podemos ver que a quantidade de homicídios é muito baixa tanto nos relatos oficiais (225) quanto nos *tweets* (63.272). Isto aconteceu devido a um esforço da polícia de Nova Iorque para diminuir este tipo de crime[12] e este impacto pode ser observado nas redes sociais.

Em relação a quantidade de dados pertences a cada um dos rótulos na base rotulada podemos ver que dos 3.435 *tweets* pertences ao rótulo **informação 0** apenas 570 (7,82%) são relatos de crimes (**crime 1**). A maioria destes dados são comentários sobre crimes (**crime 4**, 1.127) e notícias (**crime 5**, 1.206).

Na tabela 1 pode-se ver as estatísticas de cada variação do classificador do rótulo **crime**. Neste experimento foi utilizado validação cruzada fator 10 para gerar estas estatísticas. A última coluna da tabela é composta pela razão entre a quantidade de dados classificados pertencentes a classe positiva e a quantidade de dados classificados pertencentes a classe negativa. Para esta base de treino, a razão entre as classificações pertencentes ao rótulo **crime** é de 0,0782322.

Segundo os dados mostrados, as variações que apresentarem os melhores resultados, tanto na medida F1 quanto na aproximação da proporção da base de treino, foram as variações 3 e 4. Isto mostra que utilizar uma grande de dados pertencentes a classe negativa causa uma perda da eficiência do classificador neste cenário. Também podemos ver que

[12] http://www.newsday.com/news/new-york/nyc-shootings-homicides-down-in-2016-records-show-1.12830158



(a) primeira variação **(b) segunda variação** **(c) terceira variação**

(d) quarta variação **(e) quinta variação** **(f) sexta variação**

Figura 1: Nuvens de dispersão de cada variação do classificador

após o pré-processamento da base de treino, a precisão do classificador tende a aumentar.

Tabela 1: Estatísticas de cada variação do classificador do rótulo crime.

Variação	Precisão	Revocação	F1	Razão entre as classificações
1	0,888918	0,498064	0,638420	0,017501
2	0,894140	0,490071	0,633129	0,509843
3	0.869589	0.761404	0.805989	0,088705
4	0.866543	0.767296	0.809709	0,061586
5	0,809492	0,371930	0,488561	0,012816
6	0,814563	0,413941	0,537045	0,011939

Após plotagem dos dados de cada variação em uma nuvem de dispersão com 32 pontos, sendo cada ponto representando uma semana (01-07, 08-14, 15-21 e 21-28) de cada mês dos dados da base, podemos calcular a função que representa a regressão linear e o coeficiente de determinação (R^2)[5] de cada variação. A figura 1 mostra estes dados.

Podemos ver que os valores do coeficiente de determinação foram muito baixos, sendo que o maior valor foi da variação 4 que apresentou 0,53% de dependência entre as duas bases.

5 CONCLUSÕES

Este projeto teve como objetivo criar uma base de relatos de crimes coletados do *Twitter* e comparar com relatos oficiais da polícia de Nova Iorque. Para isto foi necessário coletar os dados de um site de rede social, realizar certos processamentos a fim de tentar melhorar o resultado e treinar um classificador para realizar a identificação dos crimes. Após experimentação foi identificado que os classificadores *SVM* de duas classes com uma base de treino balanceada (variações 3 e 4) apresentaram os melhores resultados em comparação com as outras variações. Entretanto nenhum dos classificadores apresentaram correlação linear, pois todos eles apresentaram coeficientes de determinação muito baixos, tendo o melhor resultado com 0,53% da base da polícia relacionada. Estimasse que este resultado pode ser melhorado ao adicionar mais pontos na nuvem de dispersão e aumentar a base de treino com mais tweets pertencentes a classe positiva.

A base de dados criada e todos os arquivos intermediários gerados podem ser acessados através do arquivo compactado disponibilizado pelos autores[13] e poderão ser utilizados em vários outros trabalhos pela comunidade. Um documento explicando cada um dos arquivos pode ser encontrado dentro do arquivo compactado. Em trabalhos futuros, a metodologia utilizada neste projeto poderá ser aplicada para criar bases de dados de crimes em cidades que não possuem uma base pública de crimes bem organizada, como é a situação de várias cidades brasileiras.

REFERÊNCIAS

[1] Derick M de Oliveira, Roberto CSNP Souza, Denise EF de Brito, Wagner Meira Jr, Gisele L Pappa, and Belo Horizonte-MG-Brasil. 2015. Uma Estratégia não Supervisionada para Previsão de Eventos usando Redes Sociais.. In *SBBD*. 137–148.

[2] Nicole B Ellison et al. 2007. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication* 13, 1 (2007), 210–230.

[3] Nicole B Ellison, Charles Steinfield, and Cliff Lampe. 2007. The benefits of Facebook "friends:" Social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication* 12, 4 (2007), 1143–1168.

[4] Tiago Cruz França, Fabrício Firmino de Faria, Fabio Medeiros Rangel, Claudio Miceli de Farias, and Jonice Oliveira. 2014. Big Social Data: Princípios sobre Coleta, Tratamento e Análise de Dados Sociais. *XXIX Simpósio Brasileiro de Banco de Dados–SBBD* 14 (2014).

[5] Joseph F Hair, William C Black, Barry J Babin, Rolph E Anderson, and Ronald L Tatham. 2009. *Análise multivariada de dados*. Bookman Editora.

[6] Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. *Machine learning: ECML-98* (1998), 137–142.

[7] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*. ACM, 851–860.

[13] https://drive.google.com/open?id=0B2bt4bROkPsLUnBIdFRmUGFwSEU