http://periodicos.letras.ufmg.br/index.php/textolivre
Ano: 2012 – Volume: 5 – Número: 1

ESTUDO DA OCORRÊNCIA DE CYBERBULLYING CONTRA PROFESSORES NA REDE SOCIAL TWITTER POR MEIO DE UM ALGORITMO DE CLASSIFICAÇÃO BAYESIANO

Rafael José de Alencar Almeida/Instituto Federal de Educação, Ciência e Tecnologia do Sudeste de Minas Gerais

RESUMO: O presente trabalho pretende contribuir com o mapeamento e a mensuração da ocorrência da violência virtual contra professores na rede social *Twitter*, por meio de técnicas computacionais baseadas em mineração de dados da Internet (*web mining*) e aprendizagem de máquina (*machine learning*). Para tal, foi realizada, durante uma semana, a coleta de postagens referentes a professores na rede social, as quais foram normalizadas e submetidas a um algoritmo de classificação Bayesiano capaz de realizar automaticamente a categorização do teor das mensagens como positivas, negativas ou neutras. Como resultado, obteve-se uma visualização gráfica hierárquica dos dados capaz de fornecer uma visão geral da gravidade e abrangência do fenômeno.

PALAVRAS CHAVE: *Cyberbullying*. Twitter. Mineração de Dados. Aprendizagem de Máquina. Filtro Bayesiano.

ABSTRACT: The present work aims to map and measure the occurrence of virtual violence against teachers at the social network Twitter, using computational techniques based on web mining and machine learning. During a week, we have collected a number of posts related to teachers in the social network, which were normalized and subjected to a Bayesian classifier capable of automatically categorize the content of the messages as positive, negative or neutral. As result, we obtained a hierarchical graphical visualization of the data able to provide an overview of the severity and the scope of the phenomenon.

KEYWORDS: Cyberbullying. Twitter. Data Mining. Machine Learning. Bayesian Filter.

INTRODUÇÃO

Na sociedade tecnológica contemporânea, o fenômeno do *cyberbullying* – *bullying* realizado através de meios digitais – vem se tornando um problema social de grandes proporções. Diferentemente do *bullying* tradicional, que normalmente se restringe ao espaço escolar, esta nova modalidade de agressão e intimidação revela-se como um tipo de violência mais cruel, uma vez que extrapola os ambientes escolares e as vítimas sofrem humilhações e ameaças constantes por meio de mensagens de celular, *e-mails*, postagens em redes sociais, dentre outras. Segundo pesquisa realizada pela fundação *Internet Safety Education* (i-SAFE Survey, 2004), 42% dos entrevistados afirmaram já ter sofrido *cyberbulying*, violência que tende a aumentar devido à crescente popularização do acesso à Internet.

Neste contexto, uma ferramenta de comunicação virtual cujo uso vem sendo subvertido para este tipo de agressão é a rede social Twitter <www.twitter.com>. Ela baseia-se no conceito de mensagens curtas de até 140 caracteres – chamadas *tweets* – que, ao serem postadas por um usuário, tornam-se acessíveis a todos aqueles que o acompanham na rede. Qualquer usuário pode propagar uma mensagem recebida com a opção *retweet*, que faz com que a mesma se torne acessível a todos usuários que acompanham aquele que a "*retweetou*". Com este mecanismo, um *tweet* pode ser veiculado pela rede a uma taxa exponencial, propagando-se de forma praticamente incontrolável e alcançando milhares de usuários. Quando usadas para fins de depreciação e humilhação, estas mensagens de amplo alcance e impacto podem infligir grande constrangimento às vítimas, que ficam praticamente impotentes frente a esta violência virtual de larga escala.

Assim como as demais redes sociais, o Twitter é um produto tecnológico relativamente recente, e como consequência, a ampla maioria de seus usuários é composta por jovens (WEBSTER, 2010). Aliado à mobilidade de acesso à Internet, principalmente em *smartphones*, estas redes de interação social vêm sendo utilizadas pelos alunos para intimidar e constranger - muitas vezes durante a própria aula – seus professores, os quais geralmente pertencem a uma geração que faz pouco uso da Internet, o que faz com que desconheçam o ocorrido ou não saibam como lidar com este tipo de provocação virtual.

De acordo com a pesquisa TIC Educação 2011, realizada pelo Centro de Estudos sobre as Tecnologias da Informação e da Comunicação em escolas públicas de todas as regiões do país, a maioria dos professores entrevistados considera que, pelo menos parcialmente, seus alunos sabem utilizar melhor o computador e a Internet do que eles próprios (CETIC, 2011). Desta forma, devido à sua grande exposição e pouca experiência com o uso da Internet e das redes sociais, alguns professores acabam sendo estigmatizados pela turma, tornando-se alvo de perseguições e chacotas realizadas por grupos de alunos nestas ferramentas virtuais.

Apesar de não haver um consenso sobre a adequação de se definir como *bullying* um comportamento que não ocorre entre pares, neste trabalho será aceita uma definição mais ampla do fenômeno – assumindo-se que em um ambiente escolar deve prevalecer o respeito entre todas as pessoas –, considerando o *bullying* um comportamento violento intencional, realizado de forma repetitiva e praticado por por um indivíduo ou grupo, com o intuito de humilhar a vítima. Neste contexto, o *cyberbullying* aluno-professor apresenta-se como um fenômeno que necessita de maior atenção e estudo, em razão da grande maioria das pesquisas e debates existentes focarem-se apenas nas relações aluno-aluno – como se pode observar em propostas de caracterização do fenômeno como a realizada por Maidel (2009), que apesar de sua valiosa contribuição para conscientização sobre o *cyberbullying*, restringe-o ao uso das tecnologias digitais com o intuito de promover constrangimento moral ou psicológico, predominantemente entre pares de crianças e adolescentes.

1 METODOLOGIA

No presente trabalho, foram empregados algoritmos de aprendizado de máquina para investigar e mapear a ocorrência do *cyberbullying* contra professores no Twitter. Para esta tarefa, foi realizada a coleta de uma grande amostra de mensagens referentes a professores – e também professoras – publicadas na rede social durante uma semana. A mineração dos dados (*data mining*) foi realizada processando e classificando de forma automatizada os *tweets* coletados, dividindo-os nas categorias "positivo" (elogios e comentários positivos), "negativo" (xingamentos, ameaças e

outras formas de depreciação) e “neutro” (não categorizáveis ou indeterminados). Nesta etapa, um filtro de classificação Bayesiano foi utilizado, algoritmo que após “treinado” com palavras-chave e exemplos de postagens pré-classificadas, torna-se capaz de realizar a categorização automática dos *tweets* submetidos. Os resultados do estudo e sua análise crítica basearam-se na visualização gerada pelo processamento das postagens coletadas.

Para implementação dos algoritmos, foi utilizada a linguagem de programação Python, por ser uma linguagem *open-source*, multiplataforma e possuir um amplo suporte à extração e processamento de dados. Todos os algoritmos utilizados e desenvolvidos foram publicados como *software* livre, e estão disponíveis para o uso e modificação por outros pesquisadores, para qualquer finalidade, no repositório <https://github.com/rafjaa/analisador-cyberbullying-twitter>.

1.1 Extração dos dados

A primeira etapa do trabalho consistiu na extração de um grande volume de postagens referentes a professores no Twitter, durante uma semana – do dia 01/04/2012 ao dia 07/04/2012. Os *tweets* foram coletados por 6 horas diárias, durante o período de 09:00 às 15:00 horas. A estratégia para obtenção dos *tweets* relevantes foi a busca daqueles contendo os termos “meu professor” e “minha professora”, os quais possuem grande probabilidade de conter referências positivas ou negativas em relação aos professores mencionados. Para cada *tweet* foi considerado o número de *retweets* do mesmo, o qual serve como indicativo de que a mensagem está sendo enviada de forma repetitiva – forte indício da ocorrência de *cyberbullying* no caso de mensagens classificadas como negativas.

A coleta dos dados foi realizada através da API (*Application Programming Interface*) fornecida pelo Twitter, cuja interface de *software* permite a interação com a rede social e seus dados. Para efetuar a busca dos *tweets*, foram realizadas consultas via protocolo HTTP GET à URL <https://search.twitter.com/search.json?q=termo>, onde o parâmetro *q* refere-se ao termo a ser pesquisado – no caso, *q* recebeu os valores “meu professor” e “minha professora”.

Como resposta de cada busca, foi gerado um arquivo no formato JSON (*JavaScript Object Notation*), o qual fornece a listagem dos *tweets* correspondentes à pesquisa, com uma série de meta-informações associadas, como o horário das postagens, seu número de *retweets*, os usuários que as submeteram, dentre outras. Através da interoperabilidade do formato JSON, os dados coletados foram convertidos para estruturas de dados da linguagem Python, adequadas para serem processadas e persistidas em disco para análise posterior.

1.2 Pré-processamento dos dados

Nesta etapa, os dados coletados passaram por um processo de tratamento e organização, extraindo-se apenas seu conteúdo relevante para análise: o texto da mensagem, sua data de publicação e o número de *retweets* de cada postagem. Uma vez que a classificação automatizada dos *tweets* baseia-se na ocorrência estatística das palavras relevantes, também foi realizado um pré-processamento do texto da mensagem, no qual foram removidos termos e caracteres desnecessários, como conjunções, palavras com menos de 3 caracteres e o próprio termo da busca – uma vez que este é uma constante em todos os resultados. Por questões de privacidade, também foram removidos dos *tweets*: *links*, endereços de *e-mail* e referências ao usuário que postou a mensagem e a outros

usuários.

Ao final do pré-processamento, obteve-se um conjunto de postagens em letras minúsculas e sem acentos, com seu conteúdo normalizado para análise e classificação, conforme ilustrado na Tabela 1, por meio de uma das postagens coletadas:

TEXTO ORIGINAL	Tenho mó vontade de dar 1 chute na cabeça achatada do meu professor, sei lá, às vezes ele pede
TEXTO PRÉ-PROCESSADO	tenho vontade dar chute cabeca achatada sei vezes pede

Tabela 1: exemplo de *tweet* pré-processado para classificação

1.3 Classificação automatizada

Para classificação automatizada dos *tweets* coletados, foi utilizado um filtro de classificação Bayesiano. Este tipo de algoritmo, cujo emprego é muito comum em detectores de *spam*, permite classificar um texto em categorias por meio da análise estatística da ocorrência de suas palavras em outros textos pré-classificados. Portanto, é um método de aprendizagem de máquina supervisionado, que requer uma etapa de treinamento, com a inserção de um conjunto de *tweets* classificados manualmente, em categorias preestabelecidas.

Devido à abordagem de classificação "ingênua" do filtro Bayesiano (*naive Bayes classifier*), não levando em conta a ordem e o relacionamento entre as palavras do texto avaliado, seu uso requer um baixo custo computacional em relação a outras técnicas de classificação, como redes neurais e máquinas de vetor de suporte (SEGARAN, 2008), o que o torna altamente viável para a classificação de uma grande amostra de *tweets*. Para o presente estudo foi utilizada uma implementação em Python do classificador Bayesiano <http://pypi.python.org/pypi/Reverend>, para o qual foram definidas as categorias "positivo", "negativo" e "neutro".

O treinamento do algoritmo foi realizado com a inserção de palavras positivas – como elogios - associadas à categoria "positivo" e palavras negativas – xingamentos, palavrões e outros termos depreciativos - associadas à categoria "negativo". Também foram coletados 300 *tweets*, os quais foram pré-processados e classificados manualmente dentro de uma das três categorias possíveis (conforme exemplo na Tabela 2), dos quais 200 foram empregados no treinamento do classificador Bayesiano.

TWEET COLETADO	CLASSIFICAÇÃO MANUAL
Meu professor de geo saca muito da matéria, além de ser ótima pessoa.	Positivo
Pqp esse meu professor de química é um gordo idiota, só fala merda!	Negativo
Meu professor tem Facebook gente!!!!	Neutro

Tabela 2: amostra de *tweets* coletados e sua classificação manual

Os 100 *tweets* restantes foram empregados na validação do classificador – etapa que visa determinar a acurácia do algoritmo após o treino –, sendo submetidos para classificação

automática, e tendo a resposta comparada com a classificação manual. Conforme a Tabela 3, o índice geral de acerto do filtro foi de aproximadamente 87%, e suas taxas individuais de acerto por categoria mantiveram-se acima de 80%, podendo ser considerada uma elevada precisão.

CATEGORIA	TAXA DE ACERTO
Positivo	88.8 %
Negativo	83.7 %
Neutro	88.8 %
Geral	87.1 %

Tabela 3: acurácia do algoritmo de classificação na etapa de validação, após seu treinamento

2 ANÁLISE DOS RESULTADOS

No período de uma semana, foram coletados mais de 6900 *tweets* contendo referências a professoras e professores. Este grande fluxo de dados foi submetido à etapa de pré-processamento e normalização, após a qual foi realizada sua classificação automática através do filtro Bayesiano, já treinado e validado. A Figura 1 representa a visualização hierárquica do resultado gerado pelo algoritmo, com a distribuição dos *tweets* por semana e gênero (professor/professora), e sua ocorrência por categoria.

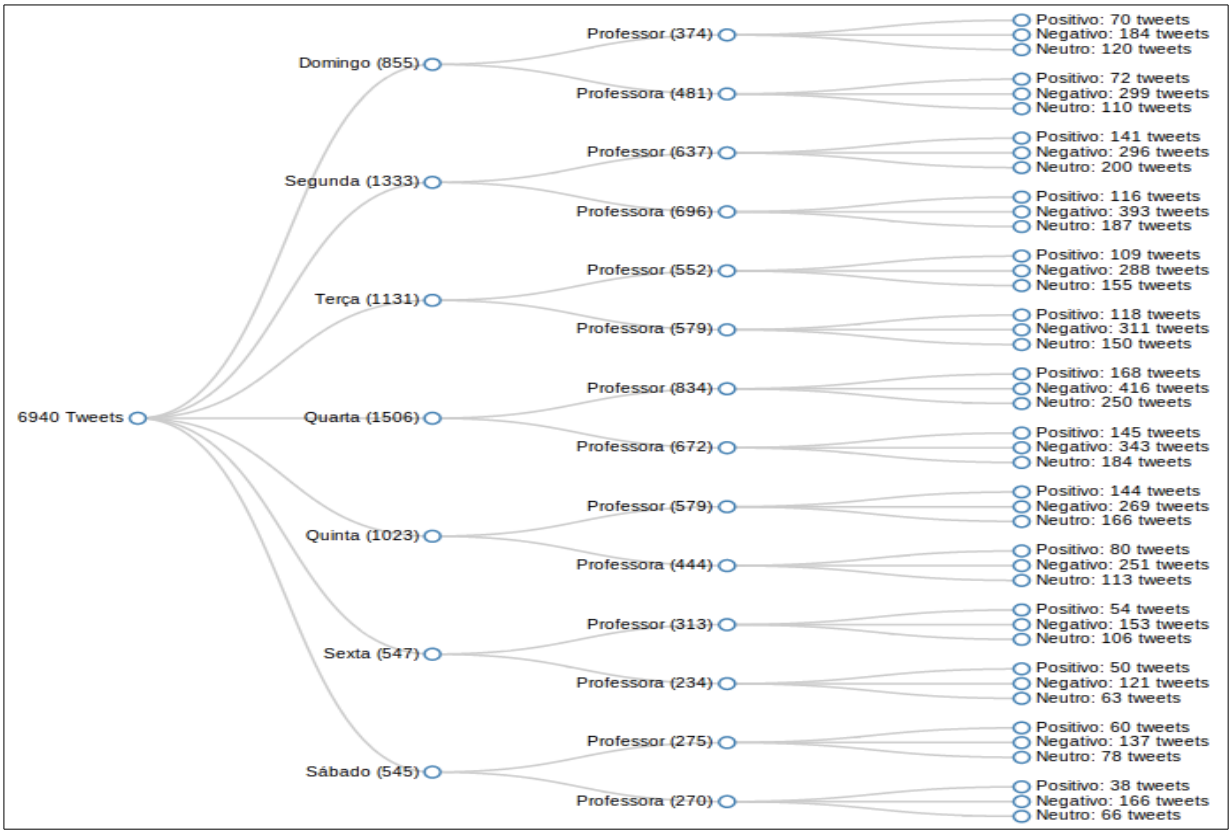

Figura 1: visualização hierárquica da classificação automatizada dos dados coletados

A primeira observação em relação aos dados é de que a violência virtual contra professores é um fenômeno *real*, de ocorrência diária e de enorme proporção, tendo sido registrados 6940 *tweets* referentes a professores durante a semana de análise, dos quais 3627 (52%) foram classificados como negativos – triste percepção de que mais da metade das mensagens direcionadas aos professores no Twitter constituem um ato de agressão virtual. Em contrapartida, esta volumosa amostra de dados – que com base nas 6 horas de coleta diária de dados representa uma taxa média de 165 postagens por hora – deixa claro o enorme potencial da rede como ferramenta para o estudo do comportamento social na Internet.

Em relação à divisão das postagens por semana, observa-se que os alunos vêm fazendo uso da rede social para manifestar suas opiniões acerca de seus professores predominantemente durante os dias de aula, registrando-se uma maior ocorrência de *tweets* de segunda-feira a quinta-feira, reduzida em quase 50% na sexta-feira de feriado (06/04/2012 – Sexta-Feira da Paixão) e no sábado, e mantendo-se proporcionalmente baixa no domingo.

Apesar da redução da quantidade de postagens durante o feriado e o final de semana, as mensagens de caráter negativo predominaram em todos os dias da semana, com índice sempre superior a 50%, seguidas das mensagens neutras, as quais apresentaram uma taxa de ocorrência um pouco acima das mensagens positivas. De acordo com a classificação fornecida pelo algoritmo, domingo foi o dia em que os ânimos mais se exaltaram – possivelmente pela iminência das aulas após o período de descanso no fim de semana –, sendo registrado o pior cenário, com o maior índice de *tweets* negativos (56,5%) e o menor índice de *tweets* positivos (16,6%). A maior quantidade de *tweets* positivos ocorreu na quinta-feira, véspera de feriado (21,9%), e o melhor cenário pôde ser observado no feriado de sexta-feira, com a ocorrência da menor taxa de postagens negativas (50,1%) e da maior taxa de *tweets* neutros (30,9%).

No quesito gênero, não parece haver distinção por parte dos alunos no foco de seus comentários na rede social. As postagens dirigem-se de forma razoavelmente equilibrada entre professores (51,35%) e professoras (48,65%), com uma pequena predominância das mensagens relativas a professoras nos três primeiros dias da semana, e com o contrário ocorrendo nos demais.

Entretanto, em relação ao conteúdo das postagens no geral, observa-se uma maior agressividade contra as professoras, com um total de 56% de mensagens negativas e 18% de mensagens positivas, contra 49% negativas e 21% positivas no caso dos professores. A neutralidade média nas mensagens relacionadas às professoras também cai, de 30% em relação aos professores, para 26% em relação a elas. Um dos fatores que pode colaborar para este cenário é uma possível visão machista por parte dos alunos, de que as professoras, por serem do sexo feminino, são mais vulneráveis e complacentes, sentindo-se mais intimidadas com este tipo de violência, havendo menor probabilidade de reagirem a esta forma de *bullying*.

CONCLUSÃO

O presente trabalho buscou mensurar a ocorrência do *cyberbullying* contra professores nas redes sociais, especificamente no Twitter, visando obter uma maior compreensão do fenômeno. Por meio da coleta e classificação automatizada de uma grande amostra de dados, foi possível perceber a dimensão e gravidade desta modalidade de *bullying*, com a impressionante constatação de que mais da metade das mensagens relativas a professores que trafegam nesta rede social são de

teor negativo.

Os resultados obtidos também proporcionaram uma visão mais ampla da ocorrência diária e por gênero desta violência virtual, permitindo observar que ela ocorre com maior frequência nos dias de aula – como foi possível observar na queda da taxa de postagens durante o feriado e finais de semana –, e de forma mais intensa contra profissionais do sexo feminino.

Apesar do uso comum de abreviaturas (devido ao limite de 140 caracteres por mensagem), gírias e erros gramaticais nas postagens dos alunos no Twitter, o algoritmo de classificação automatizada obteve, após treinado, uma elevada taxa de acerto em seu processo de validação, onde comparou-se a classificação gerada pelo algoritmo com a resposta dada por um humano. Esta constatação – associada também à consistência no relacionamento dos dados após sua categorização e visualização – demonstrou a viabilidade e eficácia do emprego de classificadores Bayesianos para categorização de postagens em língua portuguesa no Twitter.

O desenvolvimento deste trabalho também contribuiu com a produção de um projeto *open-source* em linguagem Python, com algoritmos para coleta, classificação e visualização de dados referentes ao comportamento de usuários na rede social. Com licença de uso livre, o código-fonte produzido poderá ser adaptado para novas pesquisas relacionadas, sem restrições de uso.

REFERÊNCIAS

I-SAFE AMERICA. *National i-SAFE Survey Finds Over Half of Students Are Being Harassed Online*. 2004. Disponível em: <http://www.dbprescott.com/internetbullying6.04.pdf>. Acesso em 30 mar. 2012.

MAIDEL, Simone. Cyberbullying: um novo risco advindo das tecnologias digitais. *Revista Electrónica de Investigación y Docencia* - ISSN: 1989-2446. Número 2, junho de 2009.

PYTHON PROGRAMMING LANGUAGE. *Python v2.7.3 documentation index*. Disponível em: <http://docs.python.org/>. Acesso em 31 mar. 2012.

SEGARAN, Toby. *Programando a inteligência coletiva*. Rio de Janeiro: Alta Books, 2008.

TIC EDUCAÇÃO 2011 – Pesquisa sobre o Uso das Tecnologias da Informação e da Comunicação no Brasil. Disponível em: <http://www.cetic.br/educacao/2011/p-barreiras01.htm>. Acesso em 5 jul. 2012.

TWITTER API. *Twitter platform documentation*. Disponível em: <https://dev.twitter.com/docs>. Acesso em 31 mar. 2012.

WEBSTER, Tom. *Twitter Usage In America: 2010*. 2010. Disponível em <http://images.publicaster.com/ImageLibrary/account2782/documents/Twitter_Usage_In_America_2010.pdf>. Acesso em 30 mar. 2012.