Uma Avaliação de Estratégias de Detecção de Conteúdo de Baixa Qualidade: Quais Atributos Ainda São Relevantes?*

Júlio Resende, Igor Moraes e Leonardo Rocha

[1]Universidade Federal de São João del Rei (UFSJ)

julio.cmdr@gmail.com, ti.igor.m@gmail.com, lcrocha@ufsj.edu.br

Abstract. *Millions of users have come to rely on the wide range of services provided by Social Networks. However, the ease use of social networks for communicating information also makes them particularly vulnerable to ill-intentioned users (spammers) whose main purpose is to proliferate of different types of malicious data and low-quality content (spams). Since Twitter is also rife with low-quality content, several researchers have devised various low-quality detection strategies that inspect tweets for the existence of spam contents. We carried out a literature survey of these low-quality detection strategies, evaluating which strategies are still applicable in the current scenario, taken into account that Twitter has undergone a lot of changes in the last few years.*

Resumo. *Milhões de usuários passaram a contar com a ampla gama de serviços fornecidos pelas Redes Sociais. Entretanto, a facilidade em utilizar essas redes para comunicação tornaram as mesmas vulneráveis a usuários mal intencionados (spammers), que têm objetivo de proliferar diferentes tipos de dados maliciosos ou difundir conteúdos de baixa qualidade (spams). Um dos principais exemplos dessas aplicações é o Twitter, para o qual diversas estratégias de detecção de spams vêm sendo propostas. No presente trabalho, realizamos uma pesquisa bibliográfica dessas estratégias. Por meio de uma avaliação experimental identificamos quais delas ainda são aplicáveis no cenário atual, considerando que o Twitter vem passando por mudanças constantemente.*

1. Introdução

A evolução da Web 2.0 permitiu aos usuários não apenas divulgar informações, como também interagir socialmente *online*. Utilizando-se dessa nova função da web, surgiram as redes sociais, que consolidaram a internet como o maior meio de comunicação a nível mundial. Redes sociais proporcionam um ambiente democrático e neutro, no qual usuários podem acessar e criar informações sobre diferentes assuntos e pontos de vista. Por essa razão, as redes sociais tornaram-se ferramentas fundamentais para auxiliar a tomada de decisão de diversos sistemas. Apesar disso, o potencial das redes sociais também é explorado de forma negativa por meio da produção de informações maliciosas, falsas ou simplesmente irrelevantes. Denominado por [Chen et al. 2017] como "conteúdo de baixa qualidade" (*low-quality content*), i.e., essencialmente, trata-se do *spam* em sentido amplo (além do *spam* do tipo *phishing*, considerado por diversos autores como o único tipo de *spam*). A publicação de conteúdo de baixa qualidade possui diversas motivações, que vão desde a divulgação de um produto até a propagação de *malwares*. Eles podem

*Esse trabalho foi parcialmente financiado por CNPq, CAPES, FINEP, Fapemig, e INWEB.

ser gerados tanto por seres humanos quanto por algoritmos de automatização de tarefas, os *bots*. Hoje, mais de 50% do tráfego da Internet não é de seres humanos [Łuksza 2018].

Nesse contexto, destaca-se a principal ameaça causada por conteúdos de baixa qualidade: se existem *bots* capazes de produzir tanto tráfego na Internet, e se as redes sociais possuem um papel tão importante na sociedade, é evidente que *bots* têm o potencial de influenciar a tomada de decisão dos usuários humanos. Uma das redes sociais mais afetadas por *spams* é o Twitter, que recebe mais de 700 milhões de publicações diariamente [Stats 2019]. Dentre tantos *tweets*[1], quase 10% são considerados *spams* [Ungerleider 2015]. A justificativa envolve muitos fatores: ao contrário dos serviços de e-mail, o Twitter não limita a frequência de publicações, e isso é um recurso desejável para os *spammers*. Além disso, usuários comuns também publicam *spams*, seja devido à vinculação da conta a aplicativos de terceiros, que geram conteúdo automático em nome dos usuários, ou porque a conta foi comprometida por algum *malware*.

Por essas razões supracitadas, técnicas de detecção e filtragem de *spams* apresentam-se como uma importante linha de pesquisa. Porém, o que é comumente apresentado na literatura é uma interpretação limitada do que é um *spam*. Normalmente, as interpretações mais comuns enfatizam a detecção de contas *spammers* ou a detecção de links maliciosos, o que corrobora para a geração de filtros incompletos. Correntes, informações falsas, conteúdo gerado automaticamente, *flood*, *clickbaits* e propagandas também são dados que não apresentam nenhum tipo de informação, e podem ser publicados também por contas legítimas, não sendo tratados por essas metodologias. Outro problema é que vários artigos se dispõem de diversos recursos de históricos para realizar as detecções, como *tweets* antigos de um usuário ou sua progressão de número de seguidores, sendo que alguns desses recursos não estão mais disponíveis devido a restrições impostas pela nova política do Twitter.

Nesse sentido, o presente trabalho apresenta os resultados de uma revisão da literatura relacionada ao problema de detecção de *spams*. Para realizar nosso estudo, utilizamos uma adaptação de uma metodologia de mapeamento sistemático [Petersen et al. 2008]. Assim, realizamos um pesquisa por estudos escritos em inglês e publicados entre 2010 e 2019 em cinco repositórios conhecidos: IEEE Explorer, ACM Digital Library, Springer, Scopus e Scielo. Além disso, utilizamos o mecanismo de pesquisa fornecido por cada um desses repositórios, bem como do mecanismo de pesquisa do Google Scholar. Durante a condução dessa revisão enfatizamos a verificação de quais metodologias e recursos (i.e., atributos) continuam sendo viáveis para abordar o problema de detecção online de *spams*. Adicionalmente, apresentamos uma metodologia que reúne todos os recursos que ainda se apresentam pertinentes para o problema, visando analisar, no cenário atual, quais são mais relevantes na construção de modelos de classificação. Nesse caso, utilizamos algumas técnicas de classificação automática clássicas para avaliar o impacto de tais recursos.

Enfatizamos que toda a concepção do trabalho, implementações e execuções de experimentos foram realizadas pelo aluno Júlio Resende, sob a orientação do professor Leonardo Rocha. O trabalho contou com a colaboração do aluno de mestrado do Programa de Pós-Graduação do DCOMP/UFSJ Igor Moraes nas avaliações dos resultados. Esse trabalho resultou na publicação de um artigo na ICCSA 2020.

[1]Tweet é o nome atribuído a uma publicação no Twitter.

2. Trabalhos Relacionados

O conteúdo de baixa qualidade ou spam [Chen et al. 2017], em redes sociais, é um conteúdo indesejado, produzido por fontes que expressam um comportamento diferente do pretendido pela plataforma, seja essa fonte um usuário legítimo, um *spammer* ou um *bot* [Wang et al. 2015]. O objetivo desse tipo de conteúdo inclui o compartilhamento de *links* maliciosos, informações fraudulentas, propagandas, correntes e outros, contando também com a indexação de *hashtags* não relacionadas para aumentar a visibilidade. Contudo, observamos que a maioria dos trabalhos publicados não compartilham dessa definição, utilizando apenas uma visão restrita, que não representa a totalidade do termo. A abordagem mais presente na literatura é considerar apenas o *spam* do tipo *phishing*, uma estratégia de disseminação de *malwares* por meio de URLs. A detecção desse tipo de conteúdo foi objeto de estudo de [Aggarwal et al. 2012, Martinez-Romo and Araujo 2013, Chen et al. 2015], que propuseram técnicas de verificação da ocorrência do link em lista negras cadastradas (i.e., catálogo de URLs suspeitas ou maliciosas) e técnicas que aprendem sobre as características dessas listas negras, a fim de aplicar o conhecimento gerado para classificar novas URLs. Um ponto negativo dessas estratégias é que *spammers* podem encurtar o link ou mesmo gerar um novo, com o intuito de driblar essa detecção. [Santos et al. 2014] aperfeiçoou essa metodologia, medindo a similaridade de palavras em *tweets*. Assim, caso um *tweet* possua uma URL desconhecida ou conteúdo textual similar a outros *tweets* com URLs suspeitas, ele é detectado como *phishing*. O Twitter utiliza dessa metodologia, verificando links com o *Google Safe Browsing*. Porém, links maliciosos encurtados continuam não sendo detectados como *spam*. Apenas são barrados os *tweets* que possuem links maliciosos por extenso, os idênticos aos já publicados em um curto período de tempo e os de menção dos quais o emissor e o receptor nunca interagiram entre si.

Entretanto, considerar como definição de *spam* apenas mensagens contendo *phishing* é uma abordagem muito limitada. Buscando ampliar o espectro de detecção, [Bosma et al. 2012] baseou sua definição nos relatórios de denúncia feitos por outros usuários, o que seria uma abordagem melhor, se não fosse pelo fato de que os relatórios não são disponíveis publicamente, impedindo a replicação ou aperfeiçoamento do método. Outra abordagem, explorada por [Thomas et al. 2011, Sridharan et al. 2012, Hu et al. 2014], foi a criação de uma base de dados para detecção de *tweets*, rotulando-os como *spam* caso a conta emissora tenha sido suspensa pelo Twitter em uma solicitação de validação posterior. Porém, basear-se na política de suspensão do Twitter não é o melhor caminho, já que a publicação de *spam* não é o único motivo para que uma conta seja suspensa. Além disso, *spammers* podem publicar *tweets* normais para evitarem serem detectados, enquanto usuários legítimos podem publicar conteúdo *spam*, mesmo que inconscientemente. Outros trabalhos que também focam na detecção de contas *spammers*, tais como [McCord and Chuah 2011, Almaatouq et al. 2014, Miller et al. 2014], utilizam atributos de conta extraídos por meio de API's do próprio Twitter, para gerar modelos classificadores com aprendizado de máquina. Porém, o foco nesses atributos pode enviesar a pesquisa, por serem facilmente fabricados por ferramentas externas, como obtenção de seguidores ou publicações agendadas.

Uma alternativa desenvolvida com o objetivo de detectar variações inorgânicas de valores desses atributos é a geração de modelos de comportamento temporal, explorada por [Benevenuto et al. 2010, Gao et al. 2011, Jin et al. 2011, Lee et al. 2011,

Yang et al. 2011, Tan et al. 2012, Yang et al. 2013, Zheng et al. 2015]. Considerando então dados do histórico do usuário, os modelos detectam anomalias e então sinalizam a conta como *spammer*. A abordagem é muito eficiente na detecção de *spammers*, mas não considera que usuários legítimos também publicam *spams*. Outro empecilho é que a utilização de recursos temporais infere que os *spammers* só serão detectados depois de enviarem várias mensagens de spam, pois este tipo de detecção requer um tempo mínimo para que as contas analisadas gerem "provas" que evidenciem atividades ilegais. Existem outros trabalhos cuja a metodologia de detecção é baseada em grafos, que buscam traçar todo o caminho comunicativo percorrido pelos usuários. O método é utilizado por [Wang 2010, Song et al. 2011, Fakhraei et al. 2015], nos quais os atributos extraídos determinam a quantidade de interações feitas por usuário, assim como a distância entre o emissor e o receptor em *tweets* direcionados. Com isso, a acurácia é melhorada significativamente, ao custo do tempo de resposta. Nessa mesma linha de comportamento temporal, podemos citar o trabalho apresentado por [Gao et al. 2011]. A metodologia proposta é a clusterização de *tweets* que utilizam uma mesma URL (apresentando uma sobreposição com o grupo de trabalhos descritos no primeiro parágrafo dessa seção), tratando-se, portanto, de uma detecção de *phishing*. Posteriormente, utilizando-se de listas negras, o grupo que apresenta uma URL maliciosa é considerado uma campanha de *spam*, ou seja, uma divulgação em massa de um mesmo conteúdo malicioso. Dessa forma, todas essas estratégias levam algum tempo até que o comportamento inadequado possa ser identificado. Devido ao prejuízo causado por esses conteúdos, é de suma importância a consideração de metodologias que realizem a detecção em tempo real.

Assumindo como metodologia ideal: (1) a detecção dos variados tipos de *tweets* de baixa qualidade; (2) sendo essa detecção em tempo real; e (3) utilizando atributos disponíveis publicamente – foram encontrados apenas dois artigos que cumprem todos esses requisitos [Wang et al. 2015, Chen et al. 2017]. [Wang et al. 2015] propõem uma estratégia para detectar spams em seu sentido amplo. Porém, o *dataset* utilizado para avaliação foi composto apenas por *tweets* que apresentam no mínimo uma URL. Não é possível, portanto, mensurar sua capacidade de detectar *tweets* que não apresentam esse atributo. Já em [Chen et al. 2017] foi rotulado uma base de dados de acordo com a perspectiva dos usuários sobre o que é um conteúdo de baixa qualidade. Este é um ponto de partida interessante, dada a definição abstrata do que é um conteúdo "indesejado ou irrelevante". A comparação desse trabalho com outros que adotam uma definição restrita sobre o que é conteúdo de baixa qualidade não é estatisticamente válida, devido os trabalhos terem focos diferentes. No entanto, foi apresentado nesse trabalho uma comparação de resultados com os de [Wang et al. 2015], sendo que os resultados de [Chen et al. 2017] se demonstraram mais eficientes. Um ponto crítico de [Chen et al. 2017], ao que se refere aos atributos utilizados para se construir o modelo de classificação, foi a utilização de alguns atributos que não devem ser considerados na detecção em tempo real, como o número de *retweets* e favoritos recebidos pelo *tweet* a ser analisado. Como a detecção deve ser aplicada ao publicar um *tweet*, esse valor sempre será 0. O mesmo artigo também propõe uma metodologia que faz o uso de atributos denominados indiretos, que dependem de uma análise de todas as postagens realizadas pelo usuário. Apesar desta metodologia ser capaz de aumentar significativamente o desempenho dos classificadores, o uso excessivo de informações do usuário não tem sido visto com bons olhos nos últimos anos, fato que tem pressionado o Twitter a reduzir os limites para obtenção de informações.

3. Metodologia de Avaliação

No intuito de analisar a relevância de cada atributo e a viabilidade de uma detecção de conteúdos de baixa qualidade, propomos uma metodologia empregada nas estratégias dos principais trabalhos existentes. Para realizar nossa avaliação, selecionamos uma base manualmente classificada quanto a qualidade de mensagens postadas no Twitter e utilizada em [Chen et al. 2017]. Recentemente, o Twitter proibiu que bases coletadas por pesquisadores pudessem ser compartilhadas diretamente. No entanto, em [Chen et al. 2017], os autores disponibilizaram os IDs das mensagens coletadas e, por meio desses IDs, uma nova coleta pôde ser realizada. No trabalho original, 100 mil *tweets* foram analisados manualmente, entretanto, em nossa coleta pelos IDs, percebemos que muitas das mensagens foram apagadas do Twitter. Dessa forma foi possível coletar um total de 43.857 *tweets*, no qual 3.967 estão classificados como de baixa qualidade e o restante como normal. Embora nossa base apresente valores distintos da coleção original, a proporção entre mensagens ruins e normais foi mantida.

No processo de obtenção da base de dados, além do texto dos *tweets*, foram coletados também os atributos existentes junto ao texto do *tweet* na API do Twitter. Como apresentado na Seção 2, atributos não diretos, que são provenientes de associações com *tweets* anteriores do usuário ou de seguidores do mesmo, podem melhorar a eficiência dos classificadores. No entanto, as novas políticas do Twitter impossibilitam a coleta de grandes volumes de dados e, portanto, tais atributos não foram abordados nas análises. Também não foram abordados os atributos que indicam a quantidade de *retweets* e favoritos que o *tweet* recebeu, uma vez que o uso destes em classificações em tempo real não é algo relevante. Os atributos abordados são descritos na Tabela 1 e foram propostos neste contexto pelos trabalhos relacionados que adotam um conceito amplo sobre conteúdo de baixa qualidade. Os atributos de ID 3, 4, 6 ao 9, 15 ao 18 e 35 foram propostos por [Chen et al. 2017]. Os atributos de ID 12, 13, 19 a 28, 30, 32 e 34 foram propostos por [Wang et al. 2015]. Já os atributos de ID 1, 2, 5, 10, 11, 29, 31 e 33 foram propostos por ambos. É importante ressaltar que outros trabalhos, mesmo que tenham uma abordagem diferente sobre spam, também consideraram alguns destes atributos em suas metodologias.

Id	Descrição	Id	Descrição
1	Quantidade de seguidores do usuário	19	Quantidade total de palavras no tweet
2	Quantidade de amigos do usuário	20	Quantidade de caracteres no tweet
3	Quantidade de listas do usuário	21	Quantidade de dígitos no tweet
4	Quantidade de tweets favoritados pelo usuário	22	Quantidade de espaços no tweet
5	Quantidade de tweets postados pelo usuário	23	Quantidade de palavras maiúsculas no tweet
6	Indicador de URL no perfil do usuário	24	Atributo 23 / Atributo 19
7	Indicador de descrição no perfil do usuário	25	Comprimento da maior palavra do tweet
8	Indicador de perfil padrão do usuário	26	Comprimento médio das palavras do tweet
9	Indicador de avatar padrão do usuário	27	Quantidade de exclamações (!) no tweet
10	Meses desde a criação da conta	28	Quantidade de interrogações (?) no tweet
11	Atributo 1 / Atributo 2	29	Quantidade de URLs no tweet
12	Atributo 1 / (Atributo 1 + Atributo 2)	30	Atributo 29 / Atributo 19
13	Média de amigos por mês desde a criação da conta	31	Quantidade de hashtags no tweet
14	Média de tweets por semana desde o início da conta	32	Atributo 31 / Atributo 19
15	Indicador de localização da postagem do tweet	33	Quantidade de menções a usuários no tweet
16	Ferramenta utilizada para postar o tweet	34	Atributo 33 / Atributo 19
17	Tipo do tweet: regular, menção, resposta ou retweet	35	Quantidade de símbolos no tweet
18	Indicador de conteúdo sensível no tweet		

Tabela 1. Atributos analisados nesta pesquisa

Muitos atributos não são fornecidos diretamente pela API do Twitter, mas podem ser derivados facilmente. Como exemplo, é possível destacar o atributo de ID 17, que

informa se o tweet é uma postagem regular, um retweet, uma resposta ou uma menção a outras contas. Este atributo não está disponível diretamente através da API do Twitter, mas pode ser inferido por meio de atributos booleanos, que indicam se o tweet pertence a cada uma das quatro categorias. Outros atributos, como os atributos de ID 19, 20, 21 e 22 podem ser inferidos por meio da aplicação de expressões regulares no texto do tweet.

Atenção especial foi dada ao atributo de ID 16, que informa a ferramenta utilizada para a postagem do tweet. Havia na base mais de uma centena de valores para o mesmo, alguns com ocorrência única em toda base. Uma vez que não foi informado pelos trabalhos que utilizaram este atributo qual a estratégia utilizada para a escolha dos valores, foi adotado neste trabalho uma metodologia para esta finalidade. Sendo assim, foi calculado a frequência de valor do atributo de ID 16 no conjunto de instâncias da base classificadas como de baixa qualidade e também nas instâncias que não são de baixa qualidade. Após esta etapa, foi calculado o módulo da diferença da frequência de cada valor entre os dois conjuntos da base, sendo os valores ordenados de forma decrescente, de acordo com a diferença de frequência de cada um. Para a escolha dos valores, foi aplicada uma função de interpolação na diferença das frequências, e definido um ponto de corte na função. O ponto de corte foi definido levando em consideração o Δy da função de ponto a ponto, sendo realizado o corte no momento em que o valor para de decrescer. Dessa forma, foram selecionados os valores *Twitter for iPhone, Twittascope* e *Twitter for Android*. Os demais valores foram sumarizados como "Outros". Na Figura 1, exibimos a frequência de cada termo nos conjuntos de tweets de baixa qualidade e tweets normais, sendo que os valores cortados no passo anterior já foram sumarizados como "Outros".

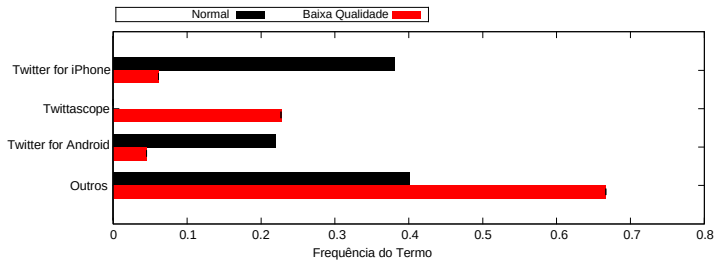

Figura 1. Frequência dos termos para o atributo 16

Para descobrir quais os atributos mais relevantes, realizamos uma análise com base em algoritmos de seleção de características. Para tal, foram aplicadas duas métricas de seleção de características do tipo filtro, que independem do classificador utilizado: Chi2 [Liu and Setiono 1995] e *Information Gain*. Tais métricas consistem, basicamente, em aplicar uma pontuação para cada atributo da base. Essas estratégias apenas apresentam uma pontuação para cada atributo. Sendo assim, é necessário o uso de uma metodologia extra capaz de determinar a quantidade de atributos que devem ser considerados. Neste trabalho, fizemos testes com todas as quantidades de atributos para os dois métodos de ranqueamento. O classificador utilizado nos testes foi o *Random Forest* (RF), sendo este o classificador que apresentou melhores resultados na literatura. A avaliação do modelo de classificação gerado para cada conjunto foi realizada por meio da métrica *F-measure* da classe de baixa qualidade, utilizando validação cruzada com 10 *folds*.

Ao final, foi realizada uma comparação do modelo mais eficiente resultante dos testes que levaram em conta os métodos de ranqueamento de atributos com outros três conjunto de dados. O primeiro conjunto de dados foi formado por todos os 35 atributos

exibidos na Tabela 1; o segundo conjunto de dados foi formado por todos os atributos do usuário (Atributos 1 ao 14); e o terceiro conjunto foi formado por todos os atributos referentes ao conteúdo em si (Atributos 15 ao 35). Nessa análise, além do classificador RF, também foi utilizado o *Support Vector Machine* (SVM). A análise foi realizada também através da técnica de validação cruzada com 10 *folds*. Os resultados dessas análises são apresentados na próxima seção.

4. Resultados e Discussões

Após a aplicação dos algoritmos *Chi2* e *Information Gain*, foram obtidas duas listas com os atributos ordenados por uma pontuação. Para efeito de visualização, os pesos de cada atributo foram normalizados de acordo com o atributo de maior peso, que teve a sua pontuação alterada para 100. Os resultados dos algoritmos de ranqueamento de atributos podem ser visualizados na Figura 2, onde cada atributo está representado pelo ID exibido na Tabela 1, sendo todos ordenados de forma decrescente, de acordo com a média dos ganhos calculados pelas duas métricas.

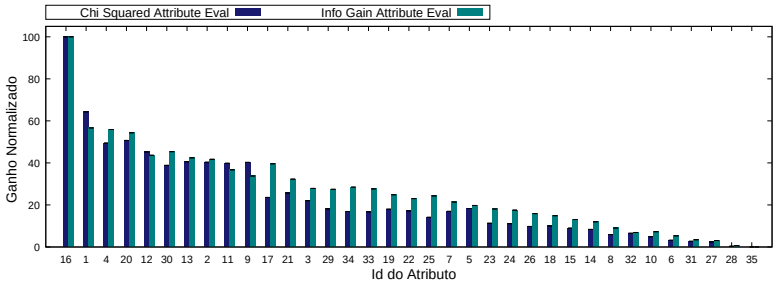

Figura 2. Pontuação de cada atributo gerada pelos algoritmos *Chi2* e *Info. Gain*

Como pode ser visualizado na Figura 2, o atributo de ID 16, que se refere à ferramenta utilizada para postar o tweet, foi o que recebeu a maior pontuação. Esse fato pode ser evidenciado de forma mais clara por meio do gráfico ilustrado na Figura 1, que exibe que todas instâncias que assumem o valor *Twittascope* no atributo de ID 16 foram consideradas como conteúdo de baixa qualidade pelos usuários que rotularam a base de [Chen et al. 2017]. Já no gráfico da Figura 3 é possível visualizar a influência da quantidade de atributos na classificação do conteúdo de baixa qualidade no Twitter. Para ambos os métodos, o maior valor para a métrica *F-Measure* foi obtido com 33 atributos, descartando os atributos de ID 28 e 35, que foram os que apresentaram as menores pontuações para os dois métodos de ranqueamento abordados.

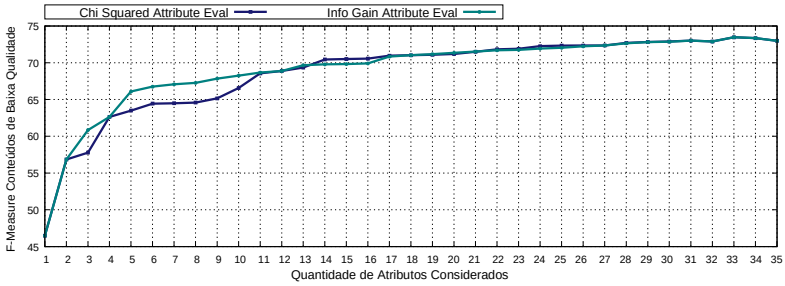

Figura 3. F-Measure da classe conteúdos de baixa qualidade para variadas quantidades de atributos segundo os algoritmos *Chi2* e *Information Gain*

Sendo assim, a Tabela 2 apresenta uma comparação mais abrangente entre diferentes conjuntos de dados, buscando analisar também a eficácia de atributos do usuário e atributos do tweet, quando aplicadas de forma separada. Os modelos gerados pelos

classificadores estão avaliados em função da taxa de verdadeiros positivos (TPR) e da métrica *F-measure* (F1), ambos em relação a classe referente a *tweets* de baixa qualidade. Também consta os valores referentes a acurácia e a Macro F1, que corresponde à média da F1 das duas classes.

Como pode ser observado na Tabela 2, o algoritmo RF obteve os melhores resultados em todos os conjuntos quando comparado com o SVM. Já em uma análise mais genérica, todos os conjuntos de dados obtiveram uma acurácia acima de 90% para todos os classificadores. No entanto, devido ao desbalanceamento da base, levar em consideração a métrica referente ao TPR e a F1 da classe de menor representação propicia uma avaliação mais fidedigna em relação ao modelo gerado por cada classificador, no qual é possível perceber que o conjunto de atributos formado pelos 33 atributos mais relevantes obtém os melhores resultados. Não muito distante, o conjunto referente a apenas os atributos do tweet também apresenta resultados satisfatórios, diferente dos atributos de usuário, que sozinhos não são capazes de apresentar bons resultados. Essa é uma importante constatação desse trabalho, ou seja, utilizar apenas informações referentes às mensagens é suficiente para se alcançar um desempenho tão bom quanto à estratégia que considera todos eles, incluindo aqueles referentes a informações dos usuários. Conforme mencionado anteriormente, o uso de informações de usuários tem sido cada vez mais desencorajado.

Conjunto de dados	Classificador	Acurácia	TPR	F1	Macro F1
33 atributos mais relevantes	SVM	93.66	43.31	55.28	75.93
	Random Forest	**96.15**	60.43	**73.81**	**85.87**
Todos Atributos	SVM	93.6	43.31	55.28	75.93
	Random Forest	96.0	59.57	73.09	85.47
Apenas Atributos do usuário (1 - 14)	SVM	91.48	14.84	22.68	59.01
	Random Forest	93.50	36.45	50.36	73.44
Apenas Atributos do tweet (15 - 35)	SVM	94.95	45.27	61.87	79.58
	Random Forest	95.77	**60.73**	72.19	84.95

Tabela 2. Resultados de classificação para 4 conjuntos de atributos distintos

5. Conclusão e Trabalhos Futuros

Neste trabalho apresentamos uma revisão da literatura focada na detecção de spams ou conteúdos de baixa qualidade em redes sociais, em especial no Twitter. O foco desta revisão foi avaliar quais trabalhos ainda seriam interessantes para a detecção de spam no cenários atual do Twitter, onde diversas informações não são mais disponibilizadas, não existem mais ou dependem de histórico de uso, portanto não são aplicadas para detecção online.

Dos trabalhos que ainda se apresentavam como relevantes, foi avaliada a relevância dos atributos considerados nos seus modelos de classificação. Separamos esses atributos em dois grupos: *(i)* aqueles focados em características dos usuários e *(ii)* aqueles focados no conteúdo das mensagens em si. Para essa avaliação, consideramos mensagens ainda disponíveis no Twitter utilizadas em análises anteriores [Chen et al. 2017] e geramos modelos de classificação automática considerando os atributos selecionados utilizando dois algoritmos de classificação: RF e SVM. De acordo com os nossos resultados, utilizar apenas informações sobre mensagens é tão bom quanto utilizar todas as informações. Tal evidência inicial fornecida pelos nossos resultados é importante pois o uso de informações de usuários tem sido cada vez mais desencorajado.

Como trabalho futuro, nosso objetivo é propor novos atributos que possam não apenas melhorar a tarefa de detecção de conteúdos de baixa qualidade, como aplicar

essas estratégias em outras redes sociais além do Twitter. Também pretendemos utilizar métricas de seleção de atributos das abordagens *wrapper*, que podem resultar em conjuntos de atributos promissores para determinados classificadores, como por exemplo as redes neurais convolucionais (CNN), que possuem alta capacidade de aprendizagem.

Referências

Aggarwal, A., Rajadesingan, A., and Kumaraguru, P. (2012). PhishAri: Automatic real-time phishing detection on twitter. In *2012 eCrime Researchers Summit*. IEEE.

Almaatouq, A., Alabdulkareem, A., Nouh, M., Shmueli, E., Alsaleh, M., Singh, V. K., Alarifi, A., Alfaris, A., and Pentland, A. S. (2014). Twitter: who gets caught? observed trends in social micro-blogging spam. In *Proceedings of the 2014 ACM conference on Web science - WebSci*. ACM Press.

Benevenuto, F., Magno, G., Rodrigues, T., and Almeida, V. (2010). Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, volume 6, page 12.

Bosma, M., Meij, E., and Weerkamp, W. (2012). A framework for unsupervised spam detection in social networking sites. In *European Conference on Information Retrieval*, pages 364–375. Springer.

Chen, C., Zhang, J., Chen, X., Xiang, Y., and Zhou, W. (2015). 6 million spam tweets: A large ground truth for timely twitter spam detection. In *2015 IEEE International Conference on Communications (ICC)*. IEEE.

Chen, W., Yeo, C. K., Lau, C. T., and Lee, B. S. (2017). A study on real-time low-quality content detection on twitter from the users' perspective. *PLOS ONE*, 12(8):1–22.

Fakhraei, S., Foulds, J., Shashanka, M., and Getoor, L. (2015). Collective spammer detection in evolving multi-relational social networks. In *Proceedings of the 21th SIGKDD*. ACM Press.

Gao, H., Chen, Y., Lee, K., Palsetia, D., and Choudhary, A. (2011). Poster. In *Proceedings of the 18th ACM conference on Computer and communications security*. ACM Press.

Hu, X., Tang, J., Gao, H., and Liu, H. (2014). Social spammer detection with sentiment information. In *2014 IEEE International Conference on Data Mining*. IEEE.

Jin, X., Lin, C. X., Luo, J., and Han, J. (2011). Socialspamguard: A data mining-based spam detection system for social media networks. In *Proceedings of the international conference on very large data bases*.

Lee, K., Eoff, B. D., and Caverlee, J. (2011). Seven months with the devils: A long-term study of content polluters on twitter. In *Fifth International AAAI Conference on Weblogs and Social Media*.

Liu, H. and Setiono, R. (1995). Chi2: feature selection and discretization of numeric attributes. In *Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence*. IEEE Comput. Soc. Press.

Martinez-Romo, J. and Araujo, L. (2013). Detecting malicious tweets in trending topics using a statistical analysis of language. *Expert Systems with Applications*, 40(8):2992–3000.

McCord, M. and Chuah, M. (2011). Spam detection on twitter using traditional classifiers. In *Lecture Notes in Computer Science*, pages 175–186. Springer Berlin Heidelberg.

Miller, Z., Dickinson, B., Deitrick, W., Hu, W., and Wang, A. H. (2014). Twitter spammer detection using data stream clustering. *Information Sciences*, 260:64–73.

Petersen, K., Feldt, R., Mujtaba, S., and Mattsson, M. (2008). Systematic mapping studies in software engineering. In *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering*, EASE'08, page 68–77, Swindon, GBR. BCS Learning & Development Ltd.

Santos, I., Miñambres-Marcos, I., Laorden, C., Galán-García, P., Santamaría-Ibirika, A., and Bringas, P. G. (2014). Twitter content-based spam filtering. In *Advances in Intelligent Systems and Computing*, pages 449–458. Springer International Publishing.

Song, J., Lee, S., and Kim, J. (2011). Spam filtering in twitter using sender-receiver relationship. In *Lecture Notes in Computer Science*, pages 301–317. Springer Berlin Heidelberg.

Sridharan, V., Shankar, V., and Gupta, M. (2012). Twitter games. In *Proceedings of the 28th ACSAC*. ACM Press.

Stats, I. L. (2019). Internet Live Stats - 1 second. `https://www.internetlivestats.com/one-second/`. Accessed: 2019-07-03.

Tan, E., Guo, L., Chen, S., Zhang, X., and Zhao, Y. (2012). Spammer behavior analysis and detection in user generated content on social networks. In *2012 IEEE 32nd International Conference on Distributed Computing Systems*. IEEE.

Thomas, K., Grier, C., Song, D., and Paxson, V. (2011). Suspended accounts in retrospect. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. ACM Press.

Ungerleider, N. (2015). Almost 10% of twitter is spam. `https://www.fastcompany.com/3044485/almost-10-of-twitter-is-spam`. Accessed: 2019-07-02.

Wang, A. H. (2010). Don't follow me: Spam detection in twitter. In *2010 International Conference on Security and Cryptography (SECRYPT)*, pages 1–10.

Wang, B., Zubiaga, A., Liakata, M., and Procter, R. (2015). Making the most of tweet-inherent features for social spam detection on twitter. *arXiv preprint arXiv:1503.07405*.

Yang, C., Harkreader, R., and Gu, G. (2013). Empirical evaluation and new design for fighting evolving twitter spammers. *IEEE Transactions on Information Forensics and Security*, 8(8):1280–1293.

Yang, C., Harkreader, R. C., and Gu, G. (2011). Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers. In *International Workshop on Recent Advances in Intrusion Detection*, pages 318–337. Springer.

Zheng, X., Zhang, X., Yu, Y., Kechadi, T., and Rong, C. (2015). ELM-based spammer detection in social networks. *The Journal of Supercomputing*, 72(8):2991–3005.

Łuksza, K. (2018). Bot traffic is bigger than human. make sure it doesn't affect you!