Detecção Automática de Spammers em Redes Sociais

Jhony M. Campanha, Johannes V. Lochter, Tiago A. Almeida
Departamento de Computação (DComp)
Universidade Federal de São Carlos (UFSCar)
18052-780, Sorocaba, São Paulo, Brasil
miller.jcamp@gmail.com, jlochter@acm.org, talmeida@ufscar.br

***Resumo*—As redes sociais são atualmente um dos serviços online mais utilizados diariamente. Relatórios recentes indicam que a maior parte do tempo dos usuários da Internet são dedicados a atualização de perfis e compartilhamento de informações em tais ferramentas. O massivo volume de usuários, aliado à alta frequência de acesso e a enorme quantidade de informações compartilhada, vêm atraindo a atenção de empresas e motivando-as a investir nesse canal de comunicação como meio de disseminação de propagandas e promoções, com o intuito de atingir rapidamente um grande público. Contudo, a facilidade encontrada na criação de perfis e inserção de conteúdo abre espaço para mensagens indesejadas, que prejudicam a experiência dos usuários, reduzem a qualidade das informações e, indiretamente, provocam prejuízos pessoais e econômicos. Diante desse cenário, esse trabalho apresenta uma análise de técnicas de aprendizado de máquina aplicadas na detecção automática de *spammers* em redes sociais. Experimentos realizados com uma base de dados real e pública, criada a partir de características das mensagens e do perfil do usuário, indicam que os métodos de florestas aleatórias, regressão logística e máquinas de vetores de suporte são promissores na tarefa de identificação de perfis falsos e podem ser empregados como *baselines* em comparações futuras.**

***Keywords*-social spam; redes sociais; classificação.**

I. Introdução

As redes sociais eletrônicas existem há décadas. Contudo, somente nos últimos anos elas se tornaram fenômeno de audiência na Internet, principalmente devido a dois importantes fatores: a inclusão digital, que permitiu alavancar o número de pessoas com acesso à Internet, e a evolução dos recursos empregados para melhorar a interação entre os usuários.

O crescimento no volume de usuários das redes sociais é tão expressivo que, atualmente, entre os dez sites mais visitados, a rede social Facebook ocupa o segundo lugar, seguida do Twitter e do LinkedIn[1]. No Brasil, estima-se que mais de um terço da população tenha conta no Facebook[2].

Com o aumento da popularidade das redes sociais, muitas empresas e comerciantes independentes passaram a vislumbrar um grande potencial para seus negócios dentro desta nova mídia, criando um ambiente de interação e disseminação de campanhas de marketing que visam aproximar o produtor ou vendedor do consumidor final. Neste cenário, pessoas mal intencionadas também visualizaram oportunidades para disseminar mensagens maliciosas e impertinentes entre o conteúdo legítimo publicado [1], [2].

Mensagens indesejadas e não-solicitadas, geralmente constituídas por informação de baixa qualidade, são denominadas *spam*. Elas normalmente incomodam os usuários, pois além de dificultar o acesso às informações que lhes interessam, costumam disseminar *links* que direcionam para sites maliciosos e perigosos. Usuários responsáveis pela disseminação de *spam* são denominados *spammers* [3].

Segundo a empresa de segurança computacional Nexgate, no primeiro semestre de 2013 houve um crescimento de 355% no volume de spam disseminado nas redes sociais, também conhecido como *social spam*[3]. Dentre as amostras analisadas, foi observado que apenas 15% do spam veiculado contém URL's que conduzem os sistemas de segurança a bloquear tais mensagens. Além disso, foi constatado que em cada 7 novas contas criadas por usuários, em média, 5 são perfis utilizados para disseminar spam. Ainda de acordo com o mesmo relatório, cada spammer utiliza, em média, 23 contas para divulgar mensagens indesejadas e a maior concentração de spammers está no Facebook e YouTube. A desproporção é tão grande que, para cada 1 spam encontrado em qualquer rede social, há outros 200 presentes no Facebook e YouTube. De acordo com dados oficiais do próprio Facebook, cerca de 200 milhões de ações mal-intencionadas são bloqueadas diariamente[4].

De acordo com o Akismet, empresa que presta serviço de filtragem de spam em blogs do Wordpress, mais de 45 bilhões de spam foram bloqueados em 2013, com uma média de 120 milhões por dia, o que representa um aumento de 75% em relação a 2012[5]. A Figura 1 ilustra o aumento no volume de spam disseminado nos blogs durante os últimos anos, conforme estatísticas reportadas pelo Akismet.

[1]Ranking disponível em: http://goo.gl/Hj4ybH. Acessado em 05/08/2014.
[2]Matéria disponível em: http://goo.gl/tV01pX. Acessado em 05/08/2014.
[3]Disponível em: http://goo.gl/sMMhyJ. Acessado em 05/08/2014.
[4]Reportagem *"Spam finds new target"*, disponível em: http://goo.gl/YoD567. Acessado em 05/08/2014.
[5]Disponível em: http://goo.gl/zNHuOk. Acessado em 05/08/2014.

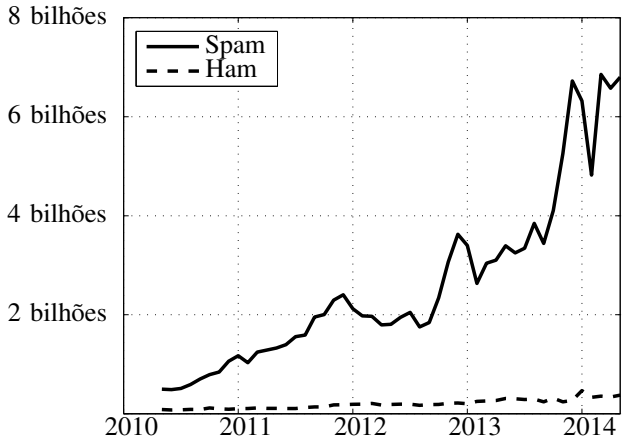

Figura 1. Quantidade de mensagens legítimas (*ham*) e indesejadas (*spam*) enviadas entre 2010 e 2014. (Dado obtido da página oficial do Akismet em Junho de 2014).

A prática normalmente usada pelos spammers para disseminar mensagens indesejadas nas redes sociais segue um certo padrão. Geralmente, um perfil falso começa a criar vínculos com usuários legítimos da rede para aumentar o alcance das suas mensagens. Uma vez que o vínculo entre o perfil falso e o perfil legítimo é estabelecido, o perfil falso começa a publicar mensagens. O usuário legítimo, por sua vez, ao clicar ingenuamente em algum link da mensagem indesejada pode disparar um mecanismo malicioso capaz de distribuir a mesma mensagem entre os seus contatos. Dessa forma, a mensagem se espalha rapidamente. Na grande maioria dos casos, o spam possui links que disparam ações diversas, desde a divulgação de algo vexatório sobre o usuário para os seus contatos até a instalação de *malware* que objetiva roubar dados pessoais da vítima.

O problema causado pelo spam nas redes sociais passou a ser seriamente discutido a partir de 2010 e, portanto, ainda há poucos trabalhos na literatura que o abordam. Destes, a maioria apresenta estudos sobre modelos e métodos de filtragem de spam aplicados no Twitter, já que tal rede social vem sendo fortemente afetada e impõe muitas dificuldades aos filtros existentes, pois as mensagens normalmente são muito curtas e repletas de gírias, símbolos e abreviações.

Nesse cenário, este artigo avalia o desempenho de diversos métodos de aprendizado de máquina que podem ser aplicados para detectar e bloquear automaticamente contas de usuários usadas para disseminar spam. O principal objetivo desse trabalho é encontrar métodos e configurações promissoras que possam ser efetivamente empregadas para tal propósito, além de oferecer um bom *baseline* para comparações futuras.

Este artigo está estruturado da seguinte forma: na Seção II são apresentados os trabalhos correlatos ao tema investigado. A metodologia experimental, incluindo base de dados, métodos e parâmetros, é apresentada na Seção III. Os principais resultados são apresentados na Seção IV. Finalmente, na Seção V são apresentadas as principais conclusões e perspectivas para trabalhos futuros.

II. Trabalhos correlatos

O problema causado pela prática de envio de spam nas redes sociais começou a ser seriamente avaliado a partir de 2010 e, em conseqüência, ainda há poucos trabalhos na literatura que o abordam. Em síntese, a maioria dos artigos disponíveis ataca o problema em duas frentes distintas: 1) pela filtragem automática de mensagens indesejadas ou 2) pela detecção e bloqueio das contas dos usuários que disseminam o spam [1]. Portanto, grande parte das pesquisas desenvolvidas adota estratégias semelhantes, primeiro buscando mapear o comportamento do spammer e, em seguida, desenvolvendo um processo reverso que venha a comprometer a proliferação das suas mensagens [4], [5].

A técnica mais comum para detectar um spammer é verificar se as suas mensagens são spam, utilizando métodos simples que analisam o conteúdo da mensagem e outras características, tais como o intervalo de postagem [6] e a similaridade entre os conteúdos [1], [7]. Além destas características, há trabalhos que demonstram que os links contidos nas mensagens também podem servir de bons indicadores para descobrir se um usuário é um spammer ou não [3].

A análise do conteúdo das mensagens tem sido amplamente utilizada para obter atributos que auxiliem na detecção de spammers [8], [9]. Contudo, também existem outras fontes de informações que podem auxiliar na classificação das contas dos usuários. Por exemplo, de acordo com [10], dados coletados do perfil permitiram melhorar o desempenho da classificação, embora muitos destes métodos estejam intrinsecamente ligados às funcionalidades e particularidades de cada rede social. Nesse sentido, foi observado que no Twitter, se um usuário segue muitas contas e possui poucos seguidores, há alta probabilidade dele ser um spammer [11]. Além disso, para dar credibilidade às mensagens falsas, alguns estudos destacam que os spammers costumam postar suas mensagens através de citações de perfis famosos e em tópicos mais discutidos (*trend topics*).

A combinação de atributos extraídos a partir do conteúdo das mensagens e dos perfis vem apresentando bons resultados [12] e, consequentemente, a maioria dos trabalhos disponíveis na literatura emprega atributos obtidos a partir dessas fontes. Contudo, a própria rede social também pode prover métricas interessantes, como denúncias de spam reportadas pelos usuários [13].

Muitas vezes, os spammers camuflam suas mensagens com o intuito de evadir as regras de filtragem impostas pelas redes sociais. Nesse sentido, referências a fatos importantes ou personalidades famosas são inseridas nas mensagens com o intuito de ganhar a confiança dos destinatários, atrair

a atenção para o teor das mensagens e acessar os links maliciosos [14], [15].

Uma estratégia comumente empregada para rastrear contas de spammers é através do uso de *honeypots* – perfis falsos que fazem o papel de usuários legítimos criados para atrair a atenção dos spammers. A partir do momento que o spammer entra em contato com um *honeypot*, o seu perfil passa a ser monitorado com o objetivo de encontrar características marcantes que possam auxiliar nos processos de detecção e bloqueio [16], [17], [18].

Através dos trabalhos presentes na literatura é possível concluir que o desempenho das técnicas de filtragem está diretamente relacionado aos atributos empregados na tarefa de classificação. Em síntese, é possível observar que o emprego de atributos extraídos apenas do conteúdo das mensagens não são suficientes para obter resultados satisfatórios. De acordo com análises recentes, a combinação de atributos extraídos tanto das mensagens enviadas pelos spammers quanto dos perfis falsos vem apresentando os melhores registros de desempenho [12], [9].

III. Metodologia Experimental

Para tornar os resultados completamente reprodutíveis, são apresentadas nessa seção as configurações adotadas para cada classificador, além de informações gerais sobre a base de dados e a metodologia experimental empregada.

A base de dados utilizada foi criada por [7] e trata-se de uma coleção real, pública e não codificada, constituída por atributos extraídos de 1.065 perfis do Twitter[6]. As amostras estão distribuídas nas classes da seguinte forma: 33,3% (355) spammers e 66,7% (710) usuários legítimos.

Cada amostra representa a conta de um usuário do Twitter e é constituída por 62 atributos listados a seguir.

- Número de seguidores por conta sendo seguida;
- Fração de *tweets*: 1. respondidos, 2. com termos característicos de *spam* e 3. com URL;
- Existência de termos característicos de *spam* no nome do perfil;
- Número de: 1. contas sendo seguidas, 2. seguidores e 3. *tweets*;
- Número de contas sendo seguidas dos seguidores do usuário;
- Número de vezes que a conta foi mencionada por outro usuário em um *tweet*;
- Número de vezes que: 1. o usuário foi respondido e 2. respondeu a um *tweet*;
- Número de *tweets* de um usuário sendo seguido; e
- Idade da conta do usuário.

Além dos 14 atributos mencionados, outros 48 foram computados através de medidas estatísticas simples (média, mediana, mínimo e máximo), conforme segue.

- Número de *hashtags* pelo número de palavras em cada *tweet*;
- Número de URL por número de palavras em cada *tweet*;
- Número de caracteres por *tweet*;
- Número de *hashtags* por *tweet*;
- Número de citações por *tweet*;
- Quantidade de caracteres numéricos por *tweet*;
- Número de URLs em cada *tweet*;
- Número de palavras por *tweet*;
- Número de vezes que o *tweet* foi reenviado, contado pela presença de "RT @username" no texto;
- Intervalo de tempo entre *posts*; e
- Número de *tweets* postados por dia e por semana.

Neste trabalho foram avaliados os seguintes métodos de classificação: naïve Bayes, máquinas de vetores de suporte, árvores de decisão, florestas aleatórias, k-vizinhos mais próximos, regressão logística e classificador baseado em regras (PART). A escolha de tais métodos reside no fato de terem sido avaliados e listados como as melhores técnicas de mineração de dados e classificação atualmente disponíveis [19].

A técnica de floresta aleatória também foi escolhida, mesmo não estando na lista originalmente proposta por [19], pois ela faz uma combinação de árvores de decisão. Logo, como o método C4.5 está na lista dos melhores métodos, acredita-se que a combinação de árvores de decisão também possa conduzir a bons resultados.

A Tabela I apresenta os métodos de classificação avaliados.

Tabela I
Métodos de classificação avaliados neste trabalho.

Classificadores
Naïve Bayes [20]
Máquinas de Vetores de Suporte (SVM) [21]
Árvores de Decisão (C4.5) [22]
Florestas Aleatórias [23]
k-vizinhos mais próximos (k-NN) [20]
Boosted C4.5 [24]
Boosted naïve Bayes [24]
Regressão Logística [25]
PART [26]

Como o desempenho dos métodos florestas aleatórias, regressão logística e SVMs são notoriamente sensíveis a ajustes de parâmetros, foi empregada a técnica de busca em *grid* para encontrar o melhor *setup* para cada uma dessas técnicas. Nesse caso, a busca foi realizada com um conjunto de treino (80% dos dados) e teste (20% dos dados),

[6] *Twitter Spammers Collection*. Disponível em http://goo.gl/0SXJOC. Acessado em 05/08/2014.

escolhidos aleatoriamente para cada conjunto de atributos avaliado, conforme descrito a seguir.

- O "número de árvores" do método florestas aleatórias foi testado entre 10 e 200, com passo de 10;
- O estimador de pico (*ridge*) da regressão logística multinomial foi testado de 10^{-10} até 10^{10}, com passo incremental 1 na potência;
- O método SVM com *kernel* linear foi testado com o atributo *C* de 10^{-3} até 10^{3}, com passo incremental 1 na potência; e
- O método SVM com *kernel* radial e polinomial foi testado com o atributo *C* e γ de 10^{-3} até 10^{3}, com passo incremental 1 na potência.

Após a execução da busca em *grid*, os seguintes parâmetros foram escolhidos e empregados nos experimentos.

- Florestas aleatórias – 110 árvores;
- Regressão logística – *ridge* = 10;
- SVM (Linear) – $C = 1000$;
- SVM (Radial) – $C = 100$ e $\gamma = 0.1$; e
- SVM (Polinomial) – $C = 0.1$ e $\gamma = 1$.

O SVM foi implementado utilizando a biblioteca LIBSVM [27] e os demais métodos foram implementados usando a ferramenta WEKA [28] com parâmetros padrões definidos na ferramenta.

Todos os experimentos foram repetidos 10 vezes utilizando validação cruzada com 5 partições. Para avaliar os classificadores, foram computadas as médias e desvios padrões de medidas de desempenho bastante populares na literatura de aprendizado de máquina e detecção de spammers, tais como:

- Acurácia (Acc) – porcentagem de amostras corretamente classificadas;
- *Spammers Caught* (SC) – porcentagem de spammers corretamente bloqueados;
- *Blocked Legitimate Users* (BLU) – porcentagem de usuários legítimos incorretamente bloqueados;
- F-Medida – média harmônica entre precisão e sensitividade;
- Micro-F1 – média ponderada das F-medidas de ambas as classes; e
- Macro-F1 – média aritmética das F-medidas de ambas as classes.

IV. Resultados

Nessa seção, são apresentados os resultados da detecção automática de spammers obtidos pelos métodos de aprendizado de máquina. A Tabela II apresenta a média e o desvio padrão dos resultados obtidos por cada método de classificação. Os valores estão ordenados por acurácia e os campos em negrito destacam os melhores desempenhos para cada medida de avaliação.

A maioria dos métodos avaliados foi capaz de detectar uma quantidade expressiva de spammers. Observa-se que, em todos os testes, a maior parte das técnicas avaliadas foi capaz de bloquear corretamente mais de 70% dos perfis usados indevidamente para disseminar spam. Por outro lado, é importante também observar que muitos métodos bloquearam erroneamente uma grande quantidade de usuários legítimos, resultando em taxas inaceitáveis de usuários legítimos bloqueados (BLU). É relevante mencionar que a taxa de usuários legítimos bloqueados é crítica para a tarefa de detecção de spammers e, portanto, métodos que apresentam altas taxas de BLU (superiores a 5%) são inaceitáveis e não recomendados para aplicação em cenários reais [8]. Nesse quesito, pode-se notar que a menor taxa de bloqueio de usuários legítimos foi obtida pelo SVM com kernel polinomial e pelo método de florestas aleatórias.

Os resultados dos experimentos indicam que os métodos de classificação que obtiveram desempenhos mais equilibrados entre detecção correta de contas de spammers e bloqueio incorreto de usuários legítimos foram florestas aleatórias, regressão logística e máquinas de vetores de suporte com kernel radial e linear. Contudo, o melhor desempenho geral em termos de acurácia, micro e macro-F1, foi obtido pela técnica de florestas aleatórias. Por outro lado, os métodos PART e k-NN obtiveram resultados insatisfatórios, principalmente evidenciados pelas altas taxas de bloqueio de usuários legítimos.

Para comparar os resultados obtidos, foi realizada uma etapa de validação estatística usando o teste-T pareado [29] com variância igual a 95%. Nesse caso, foi devidamente comprovado que o método de florestas aleatórias é estatisticamente equivalente à regressão logística multinomial e às máquinas de vetores de suporte com kernel radial e linear. Entretanto, ficou evidente que, para a variância indicada, a técnica de florestas aleatórias é estatisticamente superior aos demais métodos analisados.

Para facilitar a avaliação dos métodos, também é apresentada na Tabela III uma comparação entre os melhores resultados deste trabalho e o disponível na literatura de detecção automática de spammers que emprega a mesma coleção de dados e metodologia de avaliação.

É importante notar que os resultados obtidos estão em linha com o apresentado na literatura e que os métodos que obtiveram os melhores desempenhos podem ser empregados com sucesso para auxiliar o processo de detecção automática de spammers e usados como *baselines* para comparações futuras.

Tabela II
RESULTADOS OBTIDOS PELOS MÉTODOS DE CLASSIFICAÇÃO AVALIADOS NA TAREFA DE DETECÇÃO AUTOMÁTICA DE CONTAS DE SPAMMERS.

	Acurácia (%)	**SC (%)**	**BLU (%)**	**Macro-F1**	**Micro-F1**	**Tempo (s)**
Florestas aleatórias	**88,63 ± 1,74**	72,28 ± 5,45	3,20 ± 1,58	**0,86 ± 0,02**	**0,88 ± 0,02**	2,53
Regressão logística	87,83 ± 2,10	72,17 ± 6,17	4,34 ± 1,59	0,85 ± 0,03	0,87 ± 0,03	0,82
SVM - Radial	87,45 ± 1,92	71,27 ± 5,72	4,45 ± 1,52	0,85 ± 0,02	0,87 ± 0,02	0,92
SVM - Linear	87,33 ± 2,12	70,51 ± 5,74	4,25 ± 1,84	0,85 ± 0,03	0,87 ± 0,03	2,44
Boosted C4.5	86,17 ± 2,44	**73,58 ± 6,17**	7,54 ± 2,44	0,84 ± 0,03	0,86 ± 0,02	3,08
Boosted Naïve Bayes	86,10 ± 1,98	71,24 ± 5,27	6,46 ± 1,96	0,84 ± 0,02	0,86 ± 0,02	0,82
Naïve Bayes	86,10 ± 1,98	71,24 ± 5,27	6,46 ± 1,96	0,84 ± 0,02	0,86 ± 0,02	0,32
5-NN	86,07 ± 1,87	69,66 ± 5,11	5,73 ± 2,17	0,83 ± 0,02	0,86 ± 0,02	**0,00**
3-NN	85,52 ± 1,72	70,48 ± 4,77	6,96 ± 2,17	0,83 ± 0,02	0,85 ± 0,02	0,01
SVM - Polinomial	85,30 ± 1,82	61,41 ± 5,35	**2,76 ± 1,33**	0,82 ± 0,03	0,84 ± 0,02	1,08
C4.5	84,07 ± 2,87	71,75 ± 5,53	9,77 ± 3,06	0,82 ± 0,03	0.84 ± 0,03	0,98
1-NN	82,51 ± 2,15	72,06 ± 4,88	12,27 ± 2,42	0,80 ± 0,02	0,82 ± 0,02	**0,00**
PART	82,20 ± 2,41	73,41 ± 5,18	13,41 ± 2,54	0,80 ± 0,03	0,82 ± 0,02	1,49

Tabela III
COMPARAÇÃO ENTRE OS MELHORES RESULTADOS OBTIDOS NESTE TRABALHO E OS RESULTADOS DISPONÍVEIS NA LITERATURA.

Classificadores	**Acurácia**	**Macro-F1**	**Micro-F1**	**Tempo (s)**
Resultados de Benevenuto *et al.* [7]				
SVM - linear	87,61	0,85	**0,88**	**–**
Melhores resultados obtidos neste trabalho				
Florestas aleatórias	**88,63 ± 1,74**	**0,86 ± 0,02**	**0,88 ± 0,02**	2,53
Regressão logística	87,83 ± 2,10	0,85 ± 0,03	0,87 ± 0,03	**0,82**
SVM - radial	87,45 ± 1,92	0,85 ± 0,02	0,87 ± 0,02	0,92
SVM - linear	87,33 ± 2,12	0,85 ± 0,03	0,87 ± 0,03	2,44

V. CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho apresentou uma análise comparativa do desempenho obtido pelos métodos de classificação mais tradicionais da literatura na tarefa de detecção automática de perfis de spammers em redes sociais. Os atributos empregados, disponíveis em uma base de dados real e pública, foram extraídos e computados através de dados encontrados nas mensagens enviadas pelos usuários e informações disponíveis em seus perfis do Twitter.

Os resultados dos experimentos indicaram que a maioria das técnicas avaliadas foi capaz de detectar uma taxa expressiva de spammers. Porém, muitos desses métodos bloquearam indevidamente uma grande quantidade de usuários legítimos evidenciando a inviabilidade da aplicação prática de tais técnicas em cenários reais.

Dentre os métodos avaliados, a técnica de florestas aleatórias e regressão logística, em média, obtiveram os melhores desempenhos, inclusive apresentando taxas de acerto superiores às presentes na literatura e, portanto, demonstraram ser adequadas para auxiliar na detecção automática de spammers e usadas como *baselines* para comparações futuras. Contudo, a análise estatística dos resultados indicou que, para uma variância de 95%, esses métodos apresentaram resultados estatisticamente equivalentes aos encontrados pelas máquinas de vetores de suporte com *kernel* radial e linear e foram superiores aos demais métodos de classificação.

Trabalhos futuros compreendem o estudo de formas de adaptar os métodos mais promissores para otimizar seu desempenho, além da aplicação na detecção de spammers em outros cenários, tais como Facebook e Youtube.

AGRADECIMENTOS

Os autores são gratos a Fapesp e CNPq pelo apoio financeiro ao desenvolvimento desse projeto.

REFERÊNCIAS

[1] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao, "Detecting and characterizing social spam campaigns," in *The 10th ACM SIGCOMM Conference on Internet Measurement (ICM'10)*, Melbourne, Australia, 2010, pp. 35–47.

[2] M. Bouguessa, "An unsupervised approach for identifying spammers in social networks," in *Proceedings of the 23rd IEEE International Conference on Tools with Artificial Intelligence (ICTAI'11)*, Gatineau, Quebec, Canada, 2011, pp. 832–840.

[3] D. Wang, D. Irani, and C. Pu, "A social-spam detection framework," in *The 8th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference (CEAS'11)*, Perth, Australia, 2011, pp. 46–54.

[4] A. Zinman and J. Donath, "Is britney spears spam?" in *The 4th Conference on Email and Anti-Spam (CEAS'07)*, Mountain View, CA, USA, 2007.

[5] Y. Chen, H. Gao, K. Lee, D. Palsetia, and A. Choudhary, "Towards online spam filtering in social networks," in *The 19th Annual Network and Distributed System Decurity Symposium (NDSS'12)*, San Diego, CA, United States, 2012, pp. 1–6.

[6] S. A. Golder, D. M. Wilkinson, and B. A. Huberman, "Rhythms of social interaction: messaging within a massive online network," in *The 3rd International Conference on Communities and Technologies (CT'07)*, East Lansing, MI, USA, 2007, pp. 142–151.

[7] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on twitter," in *The 7th Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference (CEAS'10)*, Redmond, WA, USA, 2010.

[8] S. Ghosh, B. Viswanath, F. Kooti, N. K. Sharma, G. Korlam, F. Benevenuto, N. Ganguly, and K. P. Gummadi, "Understanding and combating link farming in the twitter social network," in *Proceedings of the 21st international conference on World Wide Web*, New York, NY, USA, 2012, pp. 61–70.

[9] Y. Suhara, H. Toda, S. Nishioka, and S. Susaki, "Automatically generated spam detection based on sentence-level topic information," in *The 22nd International World Wide Web Conferece (WWW'13)*, Rio de Janeiro, Brazil, 2013, pp. 1157–1160.

[10] A. H. Wang, "Don't followme: Spam detection in twitter," in *The 5th International Conference on Security and Cryptography (SECRYPT'10)*, Athens, Greece, 2010, pp. 142–151.

[11] D.-H. Park, E.-A. Cho, and B.-W. On, "Social spam discovery using bayesian network classifiers based on feature extractions," *2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications*, vol. 0, pp. 1808–1811, 2013.

[12] M. McCord and M. Chuah, "Spam detection on twitter using traditional classifiers," in *The 8th International Conference on Autonomic and Trusted Computing (ATC'11)*, Banff, Canada, 2011, pp. 175–186.

[13] M. Bosma, E. Meij, and W. Weerkamp, "A framework for unsupervised spam detection in social networking sites," in *The 34th European conference on Advances in Information Retrieval (ECIR'12)*, Barcelona, Spain, 2012, pp. 364–375.

[14] K. Thomas, C. Grier, V. Paxson, and D. Song, "Suspended accounts in retrospect: An analysis of twitter spam," in *The 11th ACM SIGCOMM Conference on Internet Measurement (IMC'11)*, Berlin, Germany, 2011, pp. 243–258.

[15] Z. Chu, I. Widjaja, and H. Wang, "Detecting social spam campaigns on twitter," in *The 10th International Conference on Applied Cryptography and Network Security (ACNS'12)*, Singapore, 2012, pp. 455–472.

[16] K. Lee, J. Caverlee, and S. Webb, "Uncovering social spammers: Social honeypots + machine learning," in *The 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10)*, Geneva, Switzerland, 2010, pp. 435–442.

[17] C. Yang, R. C. Harkreader, and G. Gu, "Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers," in *The 14th International Conference on Recent Advances in Intrusion Detection (RAID'11)*, Menlo Park, CA, USA, 2011, pp. 318–337.

[18] K. Lee, B. D. Eoff, and J. Caverlee, "Seven months with the devils: A long-term study of content polluters on twitter," in *The 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11)*, Barcelona, Spain, 2011.

[19] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, 2008.

[20] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *Eleventh Conference on Uncertainty in Artificial Intelligence*. San Mateo: Morgan Kaufmann, 1995, pp. 338–345.

[21] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[22] R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers, 1993.

[23] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[24] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Thirteenth International Conference on Machine Learning*. San Francisco: Morgan Kaufmann, 1996, pp. 148–156.

[25] S. le Cessie and J. van Houwelingen, "Ridge estimators in logistic regression," *Applied Statistics*, vol. 41, no. 1, pp. 191–201, 1992.

[26] E. Frank and I. H. Witten, "Generating accurate rule sets without global optimization," in *Fifteenth International Conference on Machine Learning*, J. Shavlik, Ed. Morgan Kaufmann, 1998, pp. 144–151.

[27] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[28] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009.

[29] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.