Detectando Avaliações Spam em uma Rede Social Baseada em Localização

Helen Costa[1]
Fabrício Benevenuto - Orientador[2]*, **Luiz H. C. Merschmann - Coorientador**[1]

[1]Departamento de Computação – Universidade Federal de Ouro Preto (UFOP)
Ouro Preto – MG – Brazil

[2]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte – MG – Brazil

helen@decsi.ufop.br, fabricio@dcc.ufmg.br, luizhenrique@iceb.ufop.br

Resumo. *Redes sociais baseadas em localização (Location-based Social Networks - LBSNs) como Foursquare e Yelp permitem que o usuário compartilhe a sua localização geográfica com sua rede social através de smartphones que possuem GPS, busquem por locais interessantes e também postem avaliações em locais existentes. Ao permitir que os usuários comentem sobre os locais, LBSNs cada vez mais têm que lidar com diferentes formas de ataques, que visam a propaganda de mensagens não solicitadas nas avaliações sobre os locais. Neste trabalho, investigamos a tarefa de identificar diferentes tipos de spam em avaliações de uma popular LBSN brasileira, chamada Apontador. Com base em uma coleção de avaliações pré-classificadas fornecida pelo Apontador e em informações coletadas sobre usuários e locais, identificamos três tipos de avaliações irregulares que denominamos como Comercial local, Boca-suja e Poluidora. Em seguida, utilizamos o nosso estudo de caracterização em uma abordagem de classificação que foi capaz de diferenciá-las com alta precisão.*

1. Introdução

Redes sociais baseadas em localização (*Location-based Social Networks* - LBSNs) são um novo tipo de sistema da Web 2.0 que vem atraindo cada vez mais novos usuários. Aproximadamente uma em cada cinco pessoas que possuem um *smartphone* acessa esse tipo de serviço através do seu dispositivo móvel [comScore 2013]. Redes como Foursquare e Yelp permitem que o usuário compartilhe a sua localização geográfica com sua rede social através de *smartphones* que possuem GPS. No Brasil, uma LBSN popular é o Apontador[1], um sistema que permite que usuários busquem por locais, cadastrem locais e façam *check-in* em locais usando um *smartphone*. Adicionalmente, o Apontador contém uma das funcionalidades mais interessantes de LBSN, a de permitir que os usuários postem avaliações em locais existentes. Através dessas avaliações (dicas) e de um *smartphone* com acesso a uma LBSN, um usuário pode não só encontrar locais próximos para visitar, como também ler sugestões sobre o que pedir, o que comprar e até mesmo o que evitar em um local específico.

Embora atraente como um mecanismo para enriquecer a experiência do usuário no sistema, avaliações abrem oportunidade para usuários disseminarem mensagens não solicitadas (spam). Essas mensagens podem ser avaliações que estão relacionadas com anúncios, pornografias e conteúdos irrelevantes que não são relacionados ao local. Esse tipo de comportamento também pode prejudicar a confiança dos usuários no sistema, comprometendo assim o seu sucesso em promover interações sociais baseadas em localização. Avaliações spam também podem comprometer a paciência e satisfação do usuário com o sistema, pois ele precisa separar spam do que vale a pena ler.

*Professor vinculado ao PPGCC da UFOP e da UFMG
[1]http://www.apontador.com.br/

Este trabalho tem como objetivo identificar diferentes tipos de avaliações spam em LBSNs, seguindo uma abordagem baseada em três etapas. A primeira etapa é categorizar avaliações spam em três diferentes classes considerando uma base de dados cedida pelo Apontador, que contém avaliações spam pré-classificadas manualmente. Em seguida, analisar vários atributos extraídos do conteúdo das avaliações e do comportamento dos usuários no sistema, com o intuito de entender o potencial desses atributos para distinguir entre diferentes classes de avaliações spam. E por fim, investigar a viabilidade da aplicação de um método de aprendizado de máquina supervisionado para identificar essas classes de avaliação spam.

As principais contribuições deste trabalho são:

- Identificação de diferentes tipos de avaliações spam, caracterizando três tipos de avaliações irregulares no Apontador, classificadas como: *(i) Comercial Local*, avaliações contendo propagandas sobre o local alvo ou sobre serviços relacionados ao local; *(ii) Poluidora*, avaliações com conteúdo irrelevante ou não relacionado com o local ou mesmo perguntas, fazendo com que a área reservada para avaliações se torne um SAC (Serviço de Atendimento ao Consumidor); e *(iii) Boca-suja*, avaliações caracterizadas por conter comentários agressivos sobre o local, seu dono ou outros usuários, geralmente contendo palavras ofensivas e de baixo calão.
- Detecção automática de diferentes tipos de avaliação spam. Nossa abordagem não foi somente capaz de identificar corretamente uma parte significativa de avaliações spam e não-spam, mas também foi capaz de diferenciar avaliações spam em três diferentes classes.

Além de realizarmos um trabalho de aspecto prático e útil para uma empresa, nossa pesquisa também resultou em publicações importantes em veículos de grande impacto:

- COSTA, H.; BENEVENUTO, F; MERSCHMANN, L. H. C; BARTH, F.: *Pollution, Bad-mouthing, and Local Marketing: The Underground of Location-based Social Networks*. In: *Elsevier Information Sciences*, 279:123–137, 2014. [Qualis A1 - CC ; Fator de Impacto = 3.643] [2]
- COSTA, H.; BENEVENUTO, F; MERSCHMANN, L. H. C.: *Detecting tip spam in location-based social networks*. In: *Proceedings of the Annual ACM Symposium on Applied Computing (SAC)*, 2013, Coimbra, Portugal. [Qualis A1 - CC] [3]

2. Coleção de Dados

Neste trabalho foram utilizadas avaliações postadas no Apontador. Com 15 milhões de usuários distintos mensais, o Apontador é um *website* brasileiro de anúncio de locais e serviços que conta com uma base de dados georreferenciada contendo aproximadamente $7,5$ milhões de pontos de interesses no Brasil.

Nós construímos nossa base de dados a partir de uma coleção de avaliações de locais rotulada manualmente como "spam" ou "não-spam" por moderadores do próprio Apontador. Nossa base da dados é composta por dois conjuntos de dados, um contendo avaliações classificadas como poluidora, comercial local, boca-suja e não-spam e outro contendo dados que coletamos utilizando a própria API do Apontador [4], a fim de melhorar os atributos utilizados para diferenciar as classes de avaliações.

Com o intuito de identificar os diferentes tipos de spam em avaliações, pedimos a um grupo de voluntários do nosso grupo de pesquisa que fizessem uma verificação manual

[2] http://dx.doi.org/10.1016/j.ins.2014.03.108
[3] http://dx.doi.org/10.1145/2480362.2480501
[4] http://api.apontador.com.br/

das avaliações spam. Ao mesmo tempo, como a classificação manual do Apontador depende de julgamento humano para decidir quando uma avaliação é spam ou não, também investigamos se os voluntários concordavam com a classificação realizada pelos moderadores do Apontador. Sendo assim, pedimos aos voluntários que fizessem uma verificação manual de todas as avaliações spam, classificando-as em spam ou não-spam e, ao mesmo tempo, que fornecessem um nome que fosse capaz de descrever uma categoria da avaliação.

Ao analisar as categorias fornecidas, pudemos identificar três classes de avaliações spam: comercial local, poluidora e boca-suja. Avaliações comerciais locais são propagandas sobre o local alvo ou sobre serviços relacionados ao local. Poluidoras são avaliações que possuem conteúdo irrelevante ou não relacionado com o local ou mesmo perguntas, fazendo com que a área reservada para avaliações se torne um tipo de SAC (Serviço de Atendimento ao Consumidor). Finalmente, bocas-sujas são avaliações caracterizadas por conter comentários agressivos sobre o local, seu dono ou outro usuário, geralmente contendo palavras ofensivas e de baixo calão. A Tabela 1 resume como as avaliações rotuladas estão distribuídas entre as classes de spam.

Tabela 1. Classe de spam

Classe	Número de Avaliações	Porcentagem
Comercial Local	1.063	30, 1%
Poluidora	1.716	48, 5%
Boca-suja	759	21, 4%
Total	3.538	100%

3. Identificando Comportamentos em LBSN

Ao contrário de usuários comuns de LBSNs, pessoas que postam spam (*spammers*) têm como objetivo o conteúdo comercial, a auto-promoção e a depreciação de ideias e reputação [Heymann et al. 2007]. *Spammers* e usuários não-spam têm objetivos diferentes no sistema e, portanto, esperamos que eles também se comportem de maneira diferente. Sendo assim, nós analisamos vários atributos que refletem o comportamento do usuário no sistema com o objetivo de investigar o poder discriminativo desses atributos para distinguir entre as classe de avaliação. Consideramos quatro grupos de atributos: atributos de conteúdo, que são propriedades do texto postado pelo usuário numa avaliação; atributos de usuário, que são propriedades específicas do comportamento do usuário no sistema; atributos de local, que são informações do local onde a avaliação foi postada; e atributos sociais, que capturam as relações estabelecidas entre os usuários através da rede social.

Alguns atributos considerados são informações fornecidas pelo próprio Apontador na base de dados cedida, porém, a maioria dos atributos foram gerados durante a nossa pesquisa. Detalhes sobre a descrição de cada atributo e como foram calculados podem ser vistos no Capítulo 4 da dissertação. Nós ainda avaliamos o poder relativo dos 60 atributos calculados para discriminar cada classe das outras através de um ranqueamento dos atributos feito pelo **cálculo de importância de atributo** do classificador *Random Forest* [Breiman 2001]. O ranqueamento que utilizamos foi o Diminuição Média da Precisão (*Mean Decrease Accuracy* - MDA) de um atributo, apresentado no Capítulo 4, Seção 5 da dissertação. A Tabela 2 sumariza os resultados. Podemos notar que os 15 atributos mais discriminativos são distribuídos entre as quatro categorias, o que demonstra a importância de termos investigado cada uma delas.

Tabela 2. *Ranking* dos atributos

Categoria	Ranking MDA	Descrição
Conteúdo	**3**	Número de endereços de e-mail no texto
32 atributos	**4**	Número de contato no texto
	5	Número de URLs no texto
	6	Número de telefone no texto
	7	Número de caracteres numéricos
	8	Polaridade de sentimento SentiStrength
	9	Polaridade de sentimento Combined-method
	10	Número de palavras
	12	Número de 1-grama distintos
	15	Número de letras maiúsculas
	16	Polaridade de sentimento SentiWordNet
	18	Polaridade de sentimento SenticNet
	25,23,19,45,27	Valor do coeficiente Jaccard (média, mediana, max., min., desvio)
	21	Polaridade de sentimento Happiness Index
	22	Polaridade de sentimento SASA
	26	Fração de 1-grama
	31	Número de palavras ou regras spam
	32	Valor de "Tem palavra ofensiva"
	34	Número de palavras com todos os caracteres em maiúsculo
	40	Número de palavras ofensivas
	43	*Clicks* no link "Esta avaliação me ajudou"
	50	Polaridade de sentimento PANAS-t
	52	*Clicks* no link "Reportar abuso"
	56	Polaridade de sentimento Emoticons
	57	Número de 3-grama distintos
	58	Fração de 3-grama
	59	Número de 2-grama distintos
	60	Fração de 2-grama
Usuário	**13**	Número de avaliações postadas pelo usuário
11 atributos	17	Número de fotos postadas pelo usuário
	20	Número de locais cadastrados pelo usuário
	38,36,42,46,24	Distância entre os locais avaliados (média, mediana, max., min., desvio)
	47	Entropia das avaliações do usuário
	51	Número de áreas diferentes onde o usuário postou uma avaliação
	55	Foco das avaliações do usuário
Local	**1**	Número de avaliações do local
5 atributos	**2**	Nota do local
	28	*Clicks* no botão "Recomendo" do local
	29	*Clicks* no botão "Não recomendo" do local
	35	*Clicks* na página do local
Social	**11**	Fração de seguidores por seguidos
12 atributos	**14**	Grau de entrada (número de seguidores)
	44,37,41,30	Assortatividade (entrada/entrada, entrada/saída, saída/entrada e saída/saída)
	33	Grau do nodo
	39	Coeficiente de clusterização
	48	Grau de saída (número de seguidos)
	49	Pagerank
	53	Reciprocidade
	54	Betweenness

4. Detectando Tipos de Spam em Avaliações

Nós investigamos a viabilidade da aplicação de um algoritmo de aprendizado supervisionado, a fim de detectar avaliações boca-suja, poluidora e comercial local no Apontador. O problema aqui abordado pode ser visto como um problema de classificação hierárquica, dado que as classes obedecem uma hierarquia. Nessa hierarquia, enquanto o primeiro nível é composto pelas classes spam (S) e não-spam (NS), o segundo nível é formado pelas classes descendentes de spam, a saber, boca-suja (BS), poluidor (PL) e comercial local (CL). Para resolver esse problema, duas abordagens de classificação foram consideradas: a plana e a hierárquica. A abordagem de classificação plana ignora a hierarquia das classes, realizando predições diretamente nos nós folha da hierarquia. Sendo assim, um classificador único é treinado para distinguir entre as avaliações não-spam, boca-suja, poluidora e comercial local. Diferentemente da plana, a abordagem hierárquica leva em conta a hierarquia das classes. Nessa abordagem, para cada nó pai da hierarquia, um classificador é construído para distinguir entre seus nós filhos.

Neste trabalho, os experimentos foram realizados utilizando os classificadores SVM (*Support Vector Machine*) [Tsochantaridis et al. 2005] e RF (*Random Forest*), que são o estado da arte em técnicas de classificação.

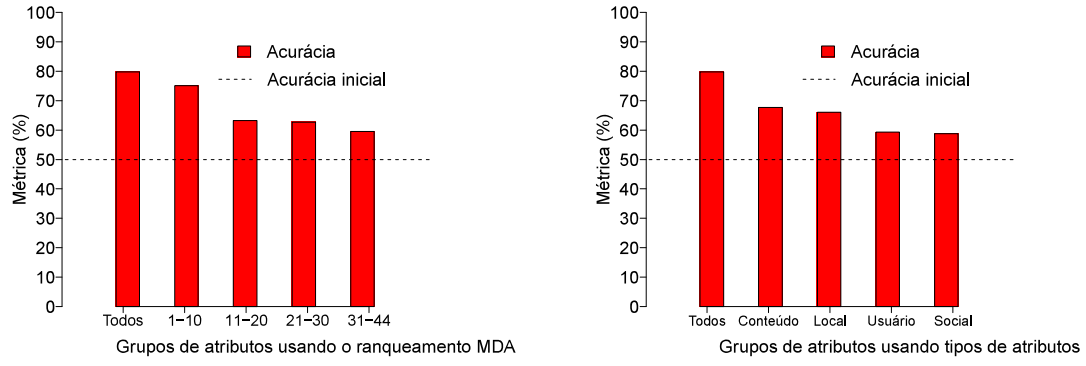
Resumindo os melhores resultados obtidos nos experimentos, a Figura 1 apresenta a matriz de confusão resultante da abordagem hierárquica utilizando o classificador RF. Observamos que 93.1% das avaliações não-spam, 76.4% das comerciais locais, 68.4% das poluidoras e 56.6% das bocas-sujas foram corretamente classificados pelo RF. Apesar dos bons resultados obtidos para as classes comercial local e não-spam (*recall* $> 75\%$), uma fração significativa das avaliações poluidoras e bocas-sujas foram classificadas erroneamente. No entanto, a abordagem hierárquica foi a que proporcionou melhores resultados para essas classes e, além disso, para a classe boca-suja, a maior porcentagem de erro da classificação (32.4%) pertence à classe poluidora, que também é uma subclasse de spam.

		Classe predita			
		NS	CL	PL	BS
Classe verdadeira	NS	**93.1%**	0.6%	4.2%	2.1%
	CL	5.7%	**76.4%**	17.0%	0.9%
	PL	19.2%	3.8%	**68.8%**	8.2%
	BS	9.9%	1.1%	32.4%	**56.6%**

Figura 1. Resultados finais da classificação hierárquica

4.1. O Impacto de Reduzir o Conjunto de Atributos

A fim de avaliar o desempenho do classificador considerando diferentes subconjuntos de atributos, realizamos experimentos utilizando subconjuntos de 10 atributos que ocupam posições contíguas no ranqueamento MDA (ou seja, os primeiros 10 melhores atributos, os próximos 10 atributos e assim por diante). A Figura 2(a) mostra o valor da métrica acurácia quando utilizamos todos os atributos, quando utilizamos diferentes subconjuntos de atributos e quando utilizamos um classificador inicial, que considera todas as avaliações como não-spam (isto é, quando não utilizamos um classificador treinado). Também realizamos experimentos usando subconjuntos de acordo com cada tipo de atributo (conteúdo, usuário, local e social) da base de dados. A Figura 2(b) mostra os resultados obtidos nesses experimentos. Nossa classificação proporciona ganhos em relação à acurácia inicial em todos os subconjuntos de atributos avaliados, isto é, mesmo atributos mal ranqueados têm algum poder discriminativo. Além disso, melhorias significativas em relação à acurácia inicial podem ser alcançadas mesmo



(a) Ranqueamento MDA (b) Ranquenamento por tipo de atributo

Figura 2. Desempenho do classificador com subconjuntos de atributos

quando apenas um tipo de atributo (por exemplo, atributos sociais) considerado em nossos experimentos pode ser obtido.

5. Conclusão

Neste trabalho, abordamos o problema de detectar diferentes tipos de spam em avaliações de uma popular LBSN brasileira. Esperamos que a identificação, caracterização e diferenciação de classes de spam em LBSNs apresentadas aqui possam ter implicações para outros sistemas baseados em avaliação e também possam ser combinadas com outras estratégias de defesa. Como exemplo, notamos que avaliações bocas-sujas são postadas em locais com notas baixas. Assim, após a detecção de avaliações bocas-sujas, pode-se tentar diferenciar se elas são avaliações verdadeiras de usuários que não gostaram mesmo do local ou se elas estão relacionadas a um ataque maliciosamente combinado contra à classificação (nota) do local. Isto poderia ser feito por meio de um mecanismo de defesa de classificação, como o Iolaus, proposto por Molavi *et al.* [Kakhki et al. 2013].

Outra possibilidade importante que a nossa abordagem revela está relacionada com as avaliações comerciais locais. Notamos que os comerciantes locais são usuários ativos que cadastram locais, e portanto, contribuem positivamente para o sistema em alguns aspectos. Ao identificá-los, o sistema poderia oferecer-lhes um contrato de publicidade para seus serviços em certas áreas do site como "avaliações patrocinadas" ao invés de simplesmente remover suas avaliações do local ou até mesmo de expulsá-los do sistema.

Como contribuição final, deixamos nossa base de dados de avaliações spam disponível para a comunidade acadêmica.

Referências

[Breiman 2001] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

[comScore 2013] comScore (Acessado em janeiro de 2013). Nearly 1 in 5 smartphone owners access check-in services via their mobile device, `http://bit.ly/mgaCIG`.

[Heymann et al. 2007] Heymann, P., Koutrika, G., and Garcia-Molina, H. (2007). Fighting spam on social web sites: A survey of approaches and future challenges. *IEEE Internet Computing*, 11:36–45.

[Kakhki et al. 2013] Kakhki, A. M., Kliman-Silver, C., and Mislove, A. (2013). Iolaus: Securing Online Content Rating Systems. In *Int'l World Wide Web Conference (WWW'13)*, pages 919–930.

[Tsochantaridis et al. 2005] Tsochantaridis, I., Joachims, T., Hofmann, T., and Altun, Y. (2005). Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research (JMLR)*, 6:1453–1484.