Caracterização dos perfis comerciais na rede social Instagram

Gabriela Enes Campos[1], **Helen Costa** [1]

[1]Departamento de Computação e Sistemas – Universidade Federal de Ouro Preto (UFOP)
João Monlevade – MG – Brasil

gabriela_enes@hotmail.com, helen@decsi.ufop.br

Abstract. *This paper presents a resulting study of users data analysis in Instagram, identifying the different existing profiles. Instagram is an application for sharring photos and videos that became popular in recent years and because of this, businesses are increasingly investing in dissemination of advertisement through it. This shows that brands are interested in disseminating their products in a relaxed atmosphere, so that their customers can have a closer relationship with the company. Thus, this research aims to characterize Instagram users, identifying and differentiating ordinary users from commercial ones. We leveraged our characterization study towards a classification approach able to differentiate these users with high accuracy.*

Resumo. *Neste artigo é apresentado um estudo resultante da análise de dados dos usuários na rede social Instagram, identificando os diferentes perfis existentes. O Instagram é um aplicativo para divulgação de fotos que se tornou popular nos últimos anos e devido a isto, empresas e comerciantes estão cada vez mais investindo na divulgação através dele. Isto mostra que as marcas estão interessadas em divulgar seus produtos de forma descontraída, fazendo com que seus clientes tenham um relacionamento mais próximo com a empresa. Sendo assim, esta pesquisa visa caracterizar as contas do Instagram, identificando e diferenciando usuários comuns de usuários comerciais. Em seguida, o estudo de caracterização é utilizado em uma abordagem de classificação que foi capaz de diferenciar estes usuários com alta precisão.*

1. Introdução

As redes sociais online (Online Social Network - OSN) têm sido um fenômeno na Internet, atraindo usuários cada vez mais. Isto é devido ao grande número de pessoas que utilizam computadores, smartphones e tablets com o acesso a Internet . Atualmente na Web existem diversos tipos de OSNs que propõem diversas funcionalidades, são redes de profissionais (ex., Linkedin), rede de amizades (ex., MySpace, Facebook) e redes para o compartilhamento de conteúdos específicos, tais como mensagens curtas (ex., Twitter), fotos (ex., Flickr, Instagram), vídeos (ex., YouTube), entre outras OSNs [Benevenuto, 2010].

O estudo das OSNs tem sido alvo de grandes pesquisadores, pois envolve fatores de interesse no mundo inteiro, como: comunicação, segurança, comércio e privacidade. Neste trabalho, será enfatizado o fator comercial, que está relacionado à divulgação de produtos e serviços, que é facilitado por OSNs, melhorando a interação com clientes e fornecedores de diversas localidades. Com isso, os usuários de redes sociais online

proporcionam uma melhor divulgação quando curtem, comentam, compartilham as informações sobre marcas, produtos e promoções, exercendo influência na sua rede social [Leskovec et al., 2007].

O Instagram é uma rede social utilizada para divulgação de fotos e vídeos curtos que se tornou muito popular nos últimos anos. Devido a isto, empresas e comerciantes estão cada vez mais investindo na divulgação através dele. Isto mostra que as marcas estão interessadas em divulgar seus produtos/serviços de forma descontraída, fazendo com que seus clientes tenham um relacionamento mais próximo da empresa. Além de empresas, usuários que são figuras públicas ou celebridades também fazem uso do Instagram para divulgação de sua própria imagem através de contas comerciais.

O Instagram ainda traz um grande desafio para estes tipos de usuários, pois não possui uma ferramenta de publicação direcionada para perfis comerciais, o que torna o processo de divulgação muito manual. Para contornar esta situação, o Instagram disponibiliza um manual para orientar empresas que querem divulgar suas marcas no aplicativo [1]. Este manual contém informações que vão desde a criação de uma conta com perfil mais comercial até o processo de divulgação dos produtos.

Apesar do Instagram possuir um manual para criação de conta diferenciada, nem todos tem o conhecimento sobre a sua disponibilidade. Com isso, nota-se que a própria empresa reconhece a necessidade de diferenciação entre perfis de usuários, que podem ter objetivos diferentes na rede. Além disso, a criação de um mecanismo automático de detecção de contas comerciais permitiria que o próprio Instagram pudesse disponibilizar aos seus usuários uma ferramenta de identificação desses tipos de usuários num mecanismo de busca dentro do próprio sistema, por exemplo. Assim, erros como o caso de um designer Andrés Iniesta (usuário comum), que teve sua conta excluída por ter o mesmo nome de um jogador de futebol famoso, não precisariam acontecer [Estado de Minas].

Neste sentido, este trabalho visa identificar quais características são eficientes para diferenciar usuários comerciais de usuários comuns no Instagram. Adicionalmente, é proposto um método de detecção de usuários comerciais utilizando a tarefa de classificação da mineração de dados.

Desta forma, o objetivo geral deste trabalho é a descoberta de padrões dos usuários do Instagram, através da criação de uma base de dados coletada utilizando a API do Instagram. Ela contém informações relativas aos dados públicos disponibilizadas nos perfis dos usuários. A partir de tais informações é possível encontrar as características que proporcionem a diferenciação dos tipos de usuários. Nesse sentido, foi proposto um método de detecção de usuários comerciais.

O restante do artigo está organizado como a seguir. A Seção 2 apresenta trabalhos relacionados com este assunto. A Seção 3 descreve como foi feita a coleta de dados e a estratégia usada para rotular usuários comerciais e comuns. A Seção 4 investiga o conjunto de atributos gerados e sua capacidade de discriminar usuários comerciais de usuários comuns. A Seção 5 descreve e avalia a estratégia usada para detectar os diferentes tipos de usuários. Finalmente, a Seção 6 oferece conclusões e direções para trabalhos futuros.

[1] Instagram: `https://help.instagram.com/454502981253053/`

2. Trabalhos Relacionados

Khosla et al. [2014] analisaram os principais fatores que tornam uma imagem popular. Os dados foram coletados do Flickr, que é uma rede social que permite a seus usuários criarem álbuns para armazenarem suas fotografias. Para prever a popularidade da imagem através do conteúdo de imagem, foram utilizados duas metologias: recursos de interpretação humana de imagem, como cor e variações de intensidade e; recursos de visão computacional de baixo nível. Para analisar cor e recurso simples de imagem, foi feita uma relação de pixels de diferentes canais de cores, mostrando que este tipo de recurso tem uma correlação muito pequena com popularidade. Nesta abordagem também foi utilizado um histograma de cores, que consistiu de um padrão de 50 cores distintas, que diagnosticou a importância das cores na previsão de popularidade. Utilizando os recursos de visão computacional, foram investigados cinco atributos de imagem:

- *Gist*: analisa a essência da imagem;
- *Texture*: analisa as texturas das imagens;
- *Color Pacthes*: analisa as imagens de acordo com uma paleta de cores;
- *Gradient*: analisa o gradiente das imagens, utilizando algoritmo de visão computacional HOG - Histograma Orientado a Gradiente;
- *Deep Learning*: analisou as imagens à partir de um vetor treinado, contendo um conjunto de palavras, para identificar a relação de popularidade quando esses objetos estava presentes nas imagens.

Outro contexto analisado foi sobre a previsão da popularidade da imagem através do conteúdo social, onde foram analisados: média de visualizações das imagens publicadas, número de fotos públicas, número de grupos, número de membros dos grupos, há quanto tempo é membro, *IsPro* (conta especial). Dessa maneira, os resultados obtidos utilizando os recursos de interpretação humana, verificou-se que imagens com cores fortes (vermelho/laranja) tem mais importância levando a um maior número de visualizações. Com as análises de visão computacional, observou que objetos como pessoas contribuem positivamente para a popularidade da foto. Os cinco atributos de imagem analisadas mostraram que o conteúdo da imagem faz muita diferença na previsão da popularidade. Este trabalho se diferencia do descrito, pois é focado somente nos usuários. Porém, são analisados os atributos da imagem do perfil, utilizando a ferramenta desenvolvida por esses autores.

Silva et al. [2013] propõem um estudo do comportamento dos usuários do Instagram na identificação de pontos de interesse de uma cidade, através do acesso das informações geográficas disponibilizadas pela RSP (Rede de Sensoriamento Participativo). O Instagram possibilita que fotos publicadas na sua rede sejam integradas ao Twitter. Desta forma, os autores coletaram sua base de dados através do Twitter, sendo 2.272.556 *tweets* contendo fotos georreferenciadas, postadas por 482.629 usuários distintos. Com isso, as análises foram feitas em regiões proporcionalmente distribuídas, para que a probabilidade de uma área aleatória ser sensoreada em um horário aleatório seja bem baixo, uma vez que os usuários compartilham fotografias em diferentes escalas de tempo. Os resultados mostraram que a aplicação criada conseguiu identificar os PDIs (Pontos de Interesses) em um contexto espaço-temporal, o que é fundamental, uma vez que os PDIs são dinâmicos e mudam ao longo do tempo. Diferentemente deste trabalho, não será feita uma caracterização de usuários utilizando informações de geolocalização, mas serão utilizadas outras características relacionadas ao comportamento do usuário.

Com o objetivo de adquirir uma compreensão inicial sobre o tipo de fotos compartilhadas por indivíduos no Instagram, Hu et al. [2014] coletaram informações sobre perfil, as 20 primeiras fotos postadas, legendas e tags associadas com fotos, seguidos e seguidores de 50 usuários escolhidos aleatoriamente. As análises consistiram de técnicas de visão computacional e humanas, onde as imagens foram separadas em 8 categorias. Assim, eles identificaram que 46,6 % representam imagens das categorias *selfie* (24,2%) e amigos (22,4%). As categorias animais e moda foram as menos populares, compreendendo apenas 5%. Outras categorias como, alimentação, tecnologia e fotos com texto representaram um pouco mais de 10% do número total de imagens. Eles também identificaram os tipos de usuários do Instagram e categorizaram como: usuários que gostam de comida; usuários que postam fotos com texto; usuários que praticam atividades físicas; usuários que amam tirar *selfies* e; usuários que postam selfies, mas também postam fotos com amigos. Tanto esse trabalho quanto o presente estudam o comportamento do usuários na identificação de padrões que possam ser categorizados, sendo ambos de extrema importância para o entendimento dos diversos interesses dos usuários na rede.

3. Coleta de dados

Para analisar o comportamento de usuários no Instagram, foi necessário primeiro criar uma base de dados através de informações de usuários que estivessem disponíveis publicamente e pudessem ser coletadas através da API[2]. Essa API utiliza o protocolo Oauth 2.0 [3], que fornece uma forma padronizada de acessar os dados protegidos. Ele proporciona autorização específica para várias aplicações. O aplicativo desenvolvido seguiu a política e os termos de uso de dados da API do Instagram, garantindo a privacidade dos usuários e respeitando a conduta desta OSN.

Para criar a base de dados proposta, primeiramente foi criada uma lista com 50 palavras que estão relacionadas as marcas empresariais, com base em uma lista disponível na página da revista Exame[4]. Utilizou-se esta lista para dar início a coleta. Em seguida, foi desenvolvido um *crawler* que, para cada termo da lista, recuperou usuários no Instagram relacionados com o termo, utilizando o método de busca da OSN. Para cada termo pesquisado, foram escolhidos os cinco primeiros usuários listados. E para cada usuário escolhido, foram coletados também os cinco primeiros seguidores e os cinco primeiros seguidos deste usuário.

A coleta foi feita no período de 23 a 29 de março de 2015 e o *crawler* desenvolvido coletou **1389** usuários distintos. Adicionalmente, também foram coletadas informações públicas disponíveis nos perfis destes usuários. Os atributos extraídos inicialmente foram:

- Username: define o nome de login do usuário;
- Bio: Descreve um "*status*" ou uma mensagem para descrição da conta;
- Website: Permite que o usuário insira uma url de contato;
- Profile picture: Fornece o caminho para acesso a imagem do perfil do usuário;
- Full name: Nome de apresentação do usuário;
- Media: Define a quantidade de fotografias que o usuário publicou;

[2]Instagram Desenvolvedor: `https://instagram.com/developer/`

[3]Oauth 2.0: `http://tools.ietf.org/html/draft-ietf-oauth-v2-12`

[4]Lista das 50 marcas mais valiosas do mundo: `http://exame.abril.com.br/marketing/noticias/as-50-marcas-mais-valiosas-do-mundo-em-2013`

Tabela 1. Quantidades de dados coletados separados em usuários comercial e comuns

Classe	Número de usuários	Porcentagem
Comum	761	55%
Comercial	628	45%
Total	1389	100%

- Followed by: Mostra a quantidade de seguidores que o usuário tem;
- Follows: Mostra a quantidade de seguidos que o usuário tem.

Através das informações coletadas do perfil dos usuários, foi feita uma análise manual de cada usuário, classificando-o entre usuário comum ou comercial. Uma sumarização dos usuários classificados pode ser vista na Tabela 1. A partir desta classificação manual, foi possível identificar que a base de dados obtida é composta tanto por usuários comerciais como por usuários comuns, mesmo a coleta tendo partido de uma busca por usuários relacionados com conteúdo comercial.

4. Caracterização de usuários

Os usuários do Instagram tem diferentes objetivos ao utilizar o sistema, e portanto, comportam-se de maneira diferente para alcança-los. Sendo assim, nesta seção serão analisados os vários atributos que refletem o comportamento do usuário na rede, para identificar as características que distinguem usuários comuns de usuários comerciais. Foram considerados três grupos de atributos: atributos de conteúdo, atributos de imagem e atributos de usuário. Esse grupos serão discutidos a seguir.

4.1. Atributos de conteúdo

Atributos de conteúdo são propriedades do texto contido na descrição de perfil feita por cada usuário através do atributo bio. Os seguintes atributos foram gerados para avaliação de cada usuário.

- número de palavras ou expressões que estão contidas em uma lista popular de conteúdo *spam*;
- número de caracteres maiúsculos;
- número de caracteres númericos;
- número de URL's no texto;
- número de endereços de e-mail;
- número de telefones;
- número de informações de contato no texto (que representa a soma do número de endereços de e-mail, telefone, e palavras Facebook e Twitter);
- número de palavras;
- número de palavras maiúsculas.

Para calcular o **número de palavras ou expressões que representam *spam***, foi utilizado uma lista com termos *spam* identificada no trabalho de Costa et al. [2013]. Esta lista está disponível em uma página do GitHub[5], que apesar de apresentar termos relacionados a *spam*, percebeu-se a semelhanças nessas palavras ao tentar atrair consumidores/-clientes. Isso permitiu avaliar este atributo para a identificação desses termos na bio dos usuários.

[5]Lista de palavras ofensivas: `github.com/spam-detection/badwords-pt-br`

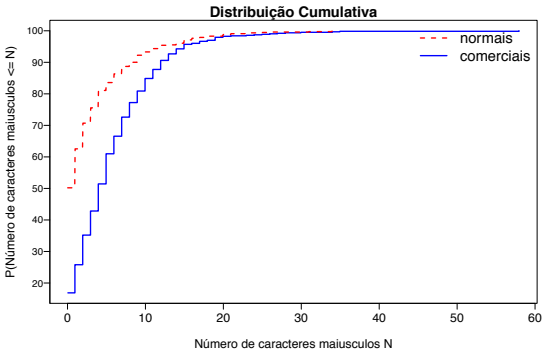

Figura 1. Número de Caracteres Maiúsculos na Bio

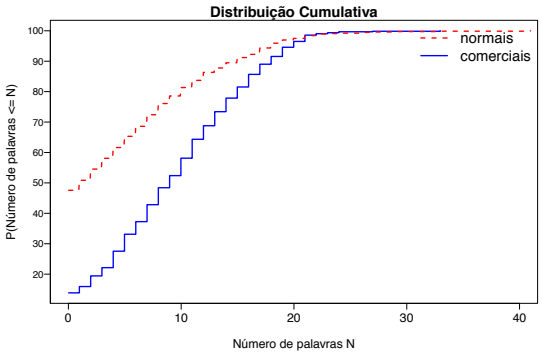

Figura 2. Número de Palavras na Bio

Para ilustrar o poder discriminativo dos atributos gerados, na Figura 1 temos a CDF (*Cumulative Distribution Function*) do atributo número de caracteres maiúsculos. Pode-se notar que os perfis comerciais possuem mais caracteres maiúsculos nas informações descritas na bio. Este fato acontece porque esta classe busca chamar mais a atenção dos outros usuários. Desta forma, percebe-se que aproximadamente 80.0% dos usuários comerciais tem até dez caracteres maiúsculos presentes nas informações da bio, enquanto que na classe de usuários comuns, cerca de 50.0% não possui nenhum carácter maiúsculo na bio.

Outra descoberta relacionada com os atributos do conteúdo é que usuários comuns utilizam menos os recursos disponíveis para a divulgação de suas informações. A Figura 2 mostra a CDF do atributo número de palavras. Observa-se que quase 50% dos usuários comuns não possuem nenhuma palavra na bio, enquanto que apenas 10.0 % aproximadamente dos usuários comerciais não possuem nenhuma palavra na bio.

4.2. Atributos de usuário

O segundo grupo de atributos consiste de propriedades específicas do comportamento do usuário na rede. Sendo assim, foram considerados os seguintes atributos de usuário:

- número de fotos postadas;
- número de seguidos;
- número de seguidores;
- fração do número de seguidores por seguidos;
- se possui informação na bio, sendo 1 para sim e 0 para não;
- se possui informação no website, sendo 1 para sim e 0 para não.

O número de seguidores se mostrou muito importante, como observado na Figura 3. Os usuários pertencentes a classe comercial têm mais seguidores do que a classe de usuários comuns. Isso acontece porque é comum os perfis comerciais serem mais expostos dos que os perfis comuns. Desta forma, nota-se que 60.0% da classe comercial tem até 10^{10} seguidores, enquanto que 60.0% da classe comum possui até 10^5 seguidores. Nesse sentido, ao invés de simplesmente medir o número de seguidores e seguidos, também foi calculada a fração de seguidores por seguidos, onde usuários comerciais mostraram um valor menor da fração. Isto mostra que os perfis comerciais produzem mais interações no

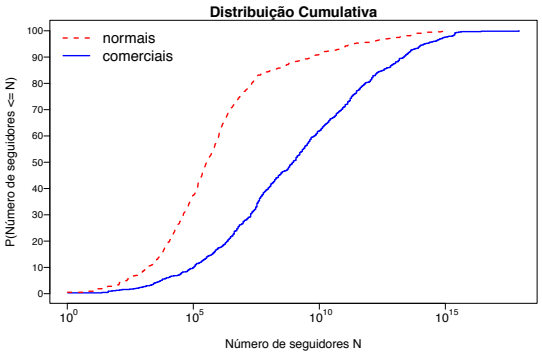

Figura 3. Seguidores

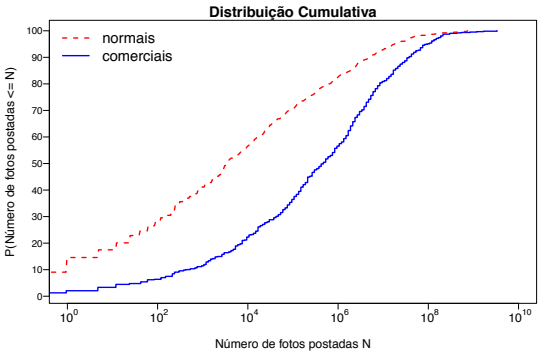

Figura 4. Fotos Publicadas

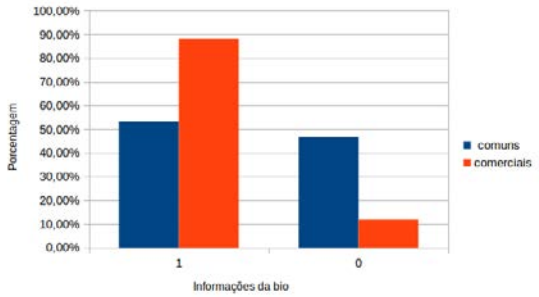

Figura 5. Informações da Bio

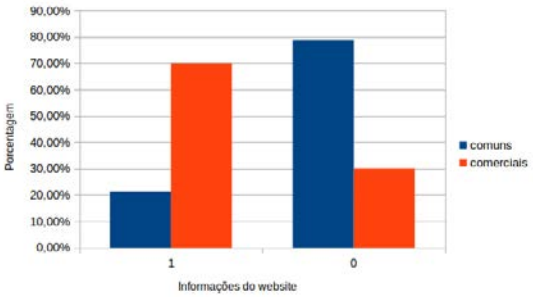

Figura 6. Informações da Website

Instagram, seguindo um número de usuários maior até do que seu número de seguidores. E esta estratégia faz com que eles sejam mais efetivos em conquistar os outros usuários.

A Figura 4 mostra a CDF do número de fotos postadas pelo usuário, onde é verificado que 60% dos usuários comuns postaram aproximadamente até 10^4 fotos, enquanto que apenas 22% dos usuários comerciais postaram a mesma quantidade de foto.

A Figura 5 mostra a distribuição do atributo de informações de bio. Observa-se que quase 90.0% dos usuários pertencentes a classe comercial apresentam textos na bio e apenas um pouco mais de 50.0% da classe comum expõem informações na bio.

Website é um recurso disponível ao usuário do Instagram, para que possam ser inseridas informações de sites ou outras redes sociais para um possível contato. Pode-se observar na Figura 6 que aproximadamente $80.0.\%$ dos usuários comuns não divulgam outros sites ou redes sociais, diferentemente dos usuários comerciais, em que aproximadamente 70.0% destes inserem este tipo de informação no Instagram. Isso mostra que o os usuários comerciais utilizam o máximo possível dos recursos do Instagram para divulgar o seu produto/marca.

4.3. Atributos das imagens

O terceiro tipo de atributo calculado faz referência a popularidade da imagem disponibilizada no perfil do usuário:

- score de popularidade da imagem do perfil.

Para calcular este atributo, utilizou-se um aplicativo que foi desenvolvido por Khosla et al. [2014] com base no estudo proposto por eles. Esse aplicativo está disponível[6] na Web e seu desenvolvimento utilizou recursos de algoritmos de visão computacional, onde foram analisados:

- Essência: analisa a ideia central da imagem;
- Textura: analisa as características visuais da imagem;
- Paleta de cores: analisa as cores das imagens, de acordo com uma paleta de 50 cores pré-definidas;
- Gradiente: analisa a intensidade dos pixels da imagem;
- Aprendizagem profunda: treinou-se um vetor de regressão linear com diferentes imagens relacionadas com objetos populares e assim, realizou as análises das imagens.

O score prevê o número de visualizações que a imagem irá receber, por exemplo, se o score é 5, espera-se que sejam obtidas 2^5=32 visualizações por dia desta imagem. O mais importante é o score relativo. Se A tem score maior que B, provavelmente A terá maior tendência a receber mais curtidas.

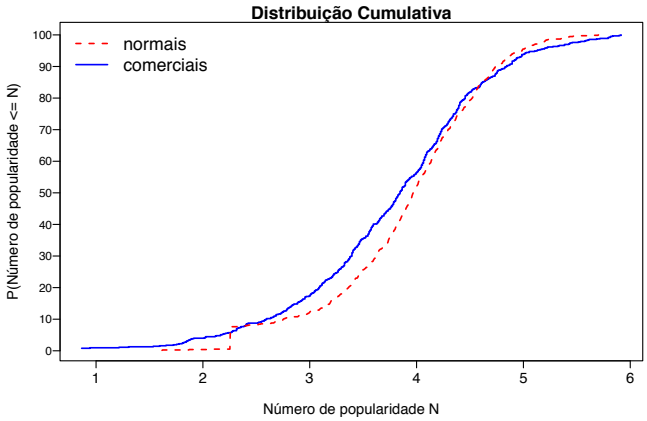

Figura 7. CDF da popularidade das imagens

Analisando a Figura 7, observa-se que os valores mais frequentes estão entre os scores 2.5 e 4, e neste mesmo intervalo é onde há uma maior diferenciação entre as distribuições, com usuários comuns tendo um score de popularidade das imagens de perfil um pouco maior que usuários comerciais.

4.4. Importância dos atributos

Para avaliar a relevância dos 16 atributos selecionados em discriminar usuários comuns de usuários comerciais, aplicou-se uma métrica de avaliação bem conhecido na literatura: Ganho de Informação [Quinlan, 1986]. Para esse experimento, utilizou-se a implementação do método fornecida pelo Weka, chamada *infoGain*. A Tabela 2 apresenta o ranqueamento dos atributos, mostrando a posição referente a pontuação dos atributos que proporcionam o maior ganho de informação de acordo com o conjunto de cada categoria (conteúdo, usuário e imagem). Dos 10 atributos mais importantes, pode-se notar que 5 pertencem a categoria de atributos de usuários.

[6]Ferramenta de popularidade da imagem: `http://popularity.csail.mit.edu/`

Tabela 2. *Ranking* dos atributos

Categoria	Ranking Ganho de Informação	Descrição
Conteúdo	**5**	Número de caracteres maiúsculos
9 atributos	**7**	Número de palavras
	9	Número de informações de contato no texto
	10	Número de caracteres númericos
	12	Número de endereços de e-mail
	13	Número de URL's no texto
	14	Número de telefones
	15	Número de palavras maiúsculas'
	16	Número termos publicitários
Usuário	**1**	Número de seguidores
6 atributos	**2**	Informações da bio
	3	Fração do número de seguidores por seguidos
	4	Número de fotos postadas
	6	Informações de website
	11	Número de seguidos
Imagem	**8**	score de popularidade da imagem
1 atributo		

5. Detectando diferentes tipos de usuários

Nesta seção, será avaliada a aplicação de um algoritmo de mineração de dados, a fim de diferenciar perfil dos usuários comerciais e comuns. A execução do algoritmo utiliza os atributos descritos na seção anterior, construindo um modelo de classificação por meio da análise do conjunto de instâncias de treinamento (usuários) representadas por um vetor de valores de atributos e um rótulo de classe. Em uma segunda etapa, o modelo de classificação é utilizado para classificar instâncias de teste (usuários) entre as classes: comercial e comum.

O trabalho proposto consiste em um problema de classificação plana, que é a maneira mais simples de lidar com problemas de classificação. Neste tipo de abordagem, um único classificador é treinado a partir da base de dados de treinamento contendo instâncias associadas a uma classe. Então, dada uma nova instância a ser classificada, o classificador atribui a ela uma classe de treinamento, que neste caso são: comum e comercial.

5.1. Métricas de avaliação

Medidas de qualidade de classificação são calculadas a partir de uma matriz que armazena as instâncias que foram classificadas corretamente e incorretamente para cada classe, denominada matriz de confusão. Assim, utilizou-se a seguinte matriz de confusão, mostrada na Figura 8.

		Classe predita	
		Comercial	**Comum**
Classe	**Comercial**	*a*	*b*
verdadeira	**Comum**	*c*	*d*

Figura 8. Matriz de confusão

Analisando a primeira linha da matriz, *a* representa a porcentagem de usuários comerciais que foram classificadas corretamente e *b* representa a porcentagem de usuários comerciais que foram classificados incorretamente como usuários comuns. Na segunda linha, *c* representa a porcentagem de usuários comuns que foram classificados incorretamente como usuários comerciais,e *d* representa a porcentagem de usuários comuns que foram classificados corretamente.

Para avaliar o desempenho do método de classificação proposto, foram consideradas as seguintes métricas, comumente utilizadas nas áreas Aprendizagem de Máquina e Recuperação de Informação [Baeza-Yates and Ribeiro-Neto, 1999]: acurácia, *recall*, Precisão e *F-measure*. *Recall* (R) e Precisão (P) são definidas como:

$$R_i = \frac{TP_i}{TP_i + FN_i},\ P_i = \frac{TP_i}{TP_i + FP_i}, \tag{1}$$

onde TP_i é o percentual de instâncias corretamente classificadas como a classe i, FN_i é o percentual de instâncias que pertencem à classe i, mas que não foram classificadas como i e FP_i é o percentual de instâncias que não pertencem a classe i, mas foram classificadas incorretamente como i pelo classificador.

Os valores da diagonal principal da matriz de confusão mostrada na Figura 8 representam o *recall* de cada classe.

A métrica *F-measure* é a maneira padrão de sumarizar *recall* e precisão, definida como $F_i = 2 \times \frac{P_i \times R_i}{P_i + R_i}$. Essa métrica infere que quanto mais próximo de 1, melhor é o resultado e, quanto mais próximo de 0, pior é o resultado.

Acurácia é o acerto do classificador considerando a proporção de instâncias corretamente classificadas no total dos registros classificados: $\frac{a+d}{a+b+c+d}$.

5.2. Configuração experimental

O algoritmo de classificação escolhido para nossos experimentos foi o Random Forest (RF), que é o estado da arte em técnicas de classificação. O RF é um classificador formado por uma coleção de árvores de classificação, cada qual construída a partir de um subconjunto aleatório de instâncias da base de treinamento. Para determinar a classe de uma instância, o método combina o resultado de várias árvores de decisão por meio de um mecanismo de votação, ou seja o método percorre cada um das árvores de decisão da floresta, e em seguida cada árvore dá uma classificação (votos) para a instância [Breiman, 2001]. A predição do RF é dada pela classe que recebeu a maioria dos votos entre todas as árvores da floresta.

Com o intuito de encontrar o melhor conjunto de parâmetros do algoritmo RF, foi executado o algoritmo de otimização de parâmetros *GridSearch* para os seguintes parâmetros do RF: *numFeatures* (usado na seleção aleatória de atributos) e *numTrees* (número de árvores a serem geradas). Como resultado, encontrou-se os valores *numFeatures* = **2** e *numTrees* = **144**, que foram adotados nos experimentos.

O desempenho preditivo foi medido usando o método de validação cruzada (CV-*Cross-Validation*) 10-fold. O conjunto de dados foi dividido em 10 partes igualmente distribuídas e a cada experimento, um dos subconjuntos de 10 é utilizado como conjunto de teste e os nove restantes são usados como dados de treinamento. Este processo é então repetido 10 vezes. Sendo assim, nossos resultados são médias de 10 experimentos, onde cada parte do todo foi usada uma vez como teste do classificador treinado com as outras partes.

5.3. Resultados experimentais

Devido ao fato de ter-se gerado e analisado atributos de imagem do perfil do usuário apenas ao final do projeto, optou-se por dividir os experimentos e a análise dos resultados

obtidos em dois testes: 1) conjunto de dados que não levam em consideração o atributo de imagem e 2) conjunto de dados em que é considerado o atributo de imagem. As Figuras 9 e 10 mostram as matrizes de confusão obtidas como resultado de nossos experimentos para a abordagem de classificação usando o classificador RF.

		Classe predita	
		Comercial	**Comum**
Classe verdadeira	**Comercial**	**73,2%**	$26,8\%$
	Comum	21%	**79%**

Figura 9. Classificação plana usando *Random Forest* sem atributo imagem

A Figura 9 mostra os resultados em que não foi considerado o atributo de imagem. O *recall* das classes estão em negrito e indicam que $73,2\%$ dos usuários comerciais e 79% dos usuários comuns foram corretamente classificados pelo RF. Pode-se observar que os resultados obtidos foram bons (*recall* $> 70\%$), porém ainda é necessário melhorar os resultados encontrados, já que $26,8\%$ dos usuários comerciais foram classificados incorretamente como usuários comuns. Para este teste, obtive-se *F-measure* de $73,3\%$ para a classe comercial e $78,6\%$ para a classe comum. A acurácia encontrada para este teste foi de $76,4\%$.

		Classe predita	
		Comercial	**Comum**
Classe verdadeira	**Comercial**	**82,6%**	$17,4\%$
	Comum	$19,2\%$	**80,8%**

Figura 10. Classificação plana usando *Random Forest* com atributo imagem

Com o intuito de obter melhores resultados, foram executados testes considerando o atributo de imagem. Os resultados são mostrados na Figura 10, onde $82,6\%$ dos usuários comerciais e $80,8\%$ dos usuários comuns foram corretamente classificados pelo RF. Desta vez pode-se observar que os resultados obtidos para ambas as classes alcançaram *recall* $> 80\%$. Os valores da *F-measure* foram $82,8\%$ para a classe comum e $80,3\%$ para a classe comercial. O atributo de imagem se mostrou muito importante devido a melhoria significativa dos resultados.

Quando compara-se os resultados desses dois experimentos, pode-se verificar que a Precisão foi melhor no teste com atributo de imagem, sendo 78% para usuários comerciais e $84,9\%$ para usuários comuns, em comparação aos valores de $74,2\%$ para usuários comerciais e $78,2\%$ para usuários comuns do teste feito sem o atributo imagem. E comparando a acurácia encontrada no primeiro teste, este segundo teste teve um melhor desempenho, alcançando $81,64\%$ de acurácia. Desta forma, a abordagem que utilizou o atributo de imagem gerou um classificador que obteve o melhor desempenho.

6. Conclusão e trabalhos futuros

O trabalho apresentou um método para identificar perfis comerciais em uma rede especializada em publicação de fotos. Através de uma base de dados coleta do Instagram, fez-se uma rotulação manual de seus usuários em duas classes distintas: comum e comercial. Utilizando a base de dados rotulada, fez-se uma caracterização, revelando os aspectos comportamentais que diferenciam essas duas classes.

Utilizou-se a tarefa de mineração de dados, denominada classificação plana, que foi capaz de distinguir de forma eficaz usuários comuns e comerciais. Com a classificação, identificou-se corretamente $82,6\%$ de usuários comuns e $80,8\%$ de usuários comerciais. A imagem de perfil do usuário foi capaz de melhorar o desempenho do nosso classificador gerado em $5,25\%$ em termos de acurácia (sendo $81,64\%$ o melhor resultado).

É esperado que a identificação, caracterização e diferenciação de tipos de usuários apresentadas neste estudo possam ser aplicados em outras OSNs. Além disso, o próprio Instagram poderia disponibilizar aos seus usuários uma ferramenta de identificação de usuários comerciais num mecanismo de busca dentro do próprio sistema.

Quando foram feitos testes envolvendo o atributo de imagem, foi observado que houve uma melhoria na classificação realizada, mostrando que a popularidade das imagens comerciais tende a ser menor. Pode-se então concluir que a imagem do perfil do usuário é um fator crucial para a diferenciação de ambas as classes.

Como trabalhos futuros, pretende-se aumentar nossa base de dados de usuários comerciais e além disso, aplicar algoritmos de mineração de dados para a tarefa de agrupamento com intuito de identificar diferentes tipos de usuários comerciais existentes.

Referências

R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

F. Benevenuto. Redes sociais online: Técnicas de coleta, abordagens de medição e desafios futuros. *Tópicos em Sistemas Colaborativos, Interativs, Multimidi, Web e Banco de Dados*, pages 41–70, 2010.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Helen Costa, Fabricio Benevenuto, and Luiz HC Merschmann. Detecting tip spam in location-based social networks. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, pages 724–729. ACM, 2013.

Estado de Minas. Designer andrés iniesta acusa instagram de "roubar" sua conta e dá-la para jogador homônimo. `http://bit.ly/1oKcThD`. Acessado em abril de 2016.

Yuheng Hu, Lydia Manikonda, Subbarao Kambhampati, et al. What we instagram: A first analysis of instagram photo content and user types. *Proceedings of ICWSM. AAAI*, 2014.

Aditya Khosla, Atish Das Sarma, and Raffay Hamid. What makes an image popular? In *Proceedings of the 23rd international conference on World wide web*, pages 867–876. International World Wide Web Conferences Steering Committee, 2014.

Jure Leskovec, Lada A Adamic, and Bernardo A Huberman. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 1(1):5, 2007.

J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

Thiago H Silva, Pedro OS Vaz de Melo, Jussara M Almeida, and Antonio AF Loureiro. Uma fotografia do instagram: Caracterização e aplicação. *Proc. of XXXII SBRC'13*, 2013.