Predizendo Influenciadores no Twitter por meio de Traços de Personalidade

Renê Gadelha[1], Ricardo Prudêncio[1], Rinaldo Lima[1], Cleyton Souza[2]

[1]Universidade Federal de Pernambuco (CIn-UFPE)

[2]Universidade Federal de Campina Grande (COPIN-UFCG)

{rnsg,rbcp,rjl2}@cin.ufpe.br, cleyton.caetano.souza@gmail.com

Abstract. *The personality traits are directly related to all actions and thoughts of a person, whether in real life or online social networks. On Twitter, a user's popularity is related to his number of followers, but this measure may not reflect on their ability to influence others. This paper presents an experimental study to predict influencers on Twitter, based on personality traits. Two types of users were investigated, celebrity followers and regular ones. The experimental results showed that regression models for influence indicators can be learned with acceptable accuracy in both cases. Several insights about the relationship between personality traits and twitter usage are also reported.*

Resumo. *Os traços de personalidade estão diretamente relacionados a todas as ações e pensamentos de um indivíduo, seja na vida real ou nas redes sociais online. No Twitter, a popularidade de um usuário está relacionada ao seu número de seguidores, mas esta medida pode não ser a única capaz de avaliar sua capacidade de influenciar os outros. Este trabalho apresenta um estudo experimental para predizer influenciadores no Twitter, com base em traços de personalidade. Dois tipos de usuários foram investigados, os seguidores de celebridades e os usuários comuns. Resultados experimentais evidenciaram que modelos de regressão para indicadores de influência podem ser aprendidos com acurácia aceitável em ambos os casos. Percepções diversas sobre a relação entre traços de personalidade e uso do Twitter também são relatados.*

1. Introdução

Nos humanos, todas suas atitudes, pensamentos e preferências são reflexos de seus traços de personalidade. Os traços de personalidade consistem em características únicas que identificam um indivíduo. Neste contexto, alguns trabalhos relatam a identificação da personalidade de um usuário através da forma como este usa as mídias sociais [Adali *et al*. 2012, Golbeck *et al*. 2011a, Golbeck *et al*. 2011b, Quercia *et al*. 2011]. Em outras palavras, traços de personalidade podem ser identificados, com relativa acurácia, por meio da análise de dados públicos compartilhados pelos usuários nas conhecidas Redes Sociais Online (RSO) [Quercia *et al* 2011].

O Twitter, um dos mais conhecidos RSOs, é um serviço voltado ao compartilhamento de conteúdo com mais de 200 milhões de usuários pelo mundo, produzindo atualmente cerca de 400 milhões de mensagens por dia. Os usuários enviam mensagens de até 140 caracteres para seus assinantes, conhecidos como “seguidores”

em português. As regras sociais no Twitter permitem aos usuários serem seguidos, como também seguir outros usuários (amigos). Os usuários recebem notificações quando seus amigos enviam mensagens, sendo estas nomeadas *tweets*. O serviço pode também ser visto como uma ferramenta para discussão de tópicos de interesse comum, além de possibilitar o compartilhamento de opiniões, experiências e sugestões.

Além disso, o Twitter possibilita a comunicação entre usuários através da retransmissão de um *tweet* (retuite). *Tweets* podem conter *hiperlink* para alguma página da *Internet,* assim como caracteres seguidos pelo símbolo "#" para indicar um tópico especial ou a idéia principal deste *tweet* (*hahstag*). Ainda é possível agrupar seus seguidores fazendo uso de listas, ou até mesmo marcar um *tweet* recebido como favorito, podendo este ser facilmente recuperado mais tarde. O Twitter permite aos usuários interagirem com outros através de nomes de usuários, isto é, usando o padrão "@nomedousuário" é possível mencionar ou conversar com outros usuários. Ainda sobre interação no Twitter, o serviço informa o usuário quando seus *tweets* são marcados como favoritos.

A maioria dos usuários do Twitter compartilham mensagens com opiniões sobre pessoas, produtos, marcas, serviços, companhias, etc. Esta informação se torna um recurso valioso para os respectivos interessados (empresas, políticos, etc.), principalmente para captação de *feedback.* No entanto, a grande quantidade de informação disponível torna de difícil processamento.

Assim, é de grande valia identificar usuários que representam o pensamento coletivo nesta RSO, sendo estes conhecidos como *influenciadores*, indivíduos capazes de exercer influência em outros. Credibilidade, perícia e entusiasmo são habilidades naturais de influenciadores, permitindo a estes induzirem idéias e ações sobre grande quantidade de pessoas [Bakshy *et al*. 2011]. A existência e importância de influenciadores nas RSO têm sido amplamente debatidas, em especial no Twitter [Berry e Keller 2003]. Estes estudos apontam a influência nas RSO como equivalente a influência cotidiana (não virtual), como também seu número de seguidores e amigos no Twitter pode estar relacionado ao tamanho da rede social real de um usuário.

Neste sentido, diversos estudos já avaliaram medidas para identificar e prever influenciadores no Twitter [Cha *et al*. 2010, Anger e Kittl 2011, Bakshy *et al*. 2011]. Nestes trabalhos, as ações do usuário (mencionar, favoritar, retuitar) e seus atributos de rede (número de amigos e seguidores) são comumente adotadas como medidas do seu grau de influência. Além dessas, o texto do *tweet* postado pode conter características específicas que exerçam influência sobre os outros.

Presumindo que o comportamento de um indivíduo esteja diretamente relacionado a sua personalidade, faz-se necessário investigar se traços de personalidade podem determinar usuários influentes em RSO, neste caso no Twitter. Com este objetivo, o presente trabalho realizou uma avaliação experimental considerando um conjunto de 6887 usuários e suas mensagens, nos quais a partir destas foi possível identificar os traços de personalidade para cada usuário. Este conjunto de usuários foi dividido em dois grupos a fim de investigar possíveis peculiaridades relativas aos seus traços de personalidade: usuários que seguiam celebridades e aqueles que não seguiam, considerados então como usuários *comuns*.

O principal objetivo deste artigo é apresentar os resultados deste estudo experimental que estabelece como suas duas maiores contribuições:

- a verificação da correlação estatística entre oito indicadores de influência no Twitter e os traços de personalidade do modelo *Big Five* [Mccrae 1992], proporcionando novas perspectivas que facilitem a identificação de potenciais influenciadores, mesmo se estes não estiverem inseridos em uma RSO;
- a avaliação de modelos de regressão utilizados na predição dos indicadores de influência por meio de traços de personalidade.

As demais seções deste artigo estão organizadas conforme descrição a seguir. Na Seção 2 são apresentados alguns trabalhos relacionados. Seção 3 descreve os dois conjuntos de dados (*datasets*) utilizados no estudo experimental, os indicadores de influência e os atributos preditores. Em seguida, na Seção 4, é apresentado o modelo de predição proposto e sua avaliação. Após discutir os resultados obtidos na Seção 5, são delineadas as conclusões e trabalhos futuros na Seção 6.

2. Trabalhos Relacionados

Muitos estudos vêm sendo conduzidos com o objetivo de identificar a influência no Twitter. A seguir, são apresentados trabalhos relacionados aos influenciadores e personalidade.

2.1 Influenciadores

Alguns trabalhos avaliam a influência de um usuário no Twitter através de *indicadores* [e.g., Cha *et al*. 2010, Anger e Kittl 2011, Bakshy *et al*. 2011]. Esses trabalhos apontam alguns indicadores baseados em diferentes aspectos de cada usuário, tais como: características do perfil (número de vezes que o usuário é mencionado, retuitado ou respondido), estatísticas do usuário na rede (número de seguidores, número de amigos, Pagerank) e conteúdo dos *tweets* (opinião positiva e/ou negativa, *hashtag*, links, etc). [Bakshy *et al*. 2011] identificaram o papel dos usuários influentes na propagação de informação sobre eventos importantes e o efeito cascata de *tweets* relacionados a tópicos de interesse comum. Os seguintes atributos foram utilizados para caracterizar os influenciadores: data de criação do perfil, número de seguidores, número de amigos e número de *tweets*.

O algoritmo de Pagerank, utilizado em [Weng *et al*. 2010] para identificar usuários influentes no Twitter, coleta a informação sobre a conexão entre dois usuários como a principal característica preditora. Uma das conclusões desse trabalho é que a semelhança entre os assuntos discutidos e o elevado número de seguidores não estão entre as razões que levam um usuário a seguir outro. Neste estudo é relatada fraca correlação entre usuários com elevado número de seguidores e usuários com elevado número de retuites. Além disso, os autores relataram que o fato do usuário possuir elevado número de seguidores não é necessariamente um indicativo de influência.

Em [Cha *et al*. 2010], os autores investigaram o número de amigos, o número de seguidores e o número de retuites como medidas de influência. Entretanto, os resultados dismistificaram os usuários populares, pois mesmo eles possuindo elevado número de

seguidores não são necessariamente influentes em termos de número de retuites ou menções.

2.2 Personalidade

Além das características da rede social, o conteúdo produzido pelo usuário pode ser importante para definir sua influência, já que o texto contém informação sobre características únicas relacionadas ao seu autor. Entre essas características estão os trações de personalidade que reúnem os principais aspectos psicológicos que determinam as ações e comportamentos de um indivíduo [Mairesse *et al*. 2007]. Um dos modelos de personalidades mais adotados é o *Big-Five* que caracteriza a personalidade em cinco traços principais: extroversão (*extraversion*), neuroticismo (*neuroticism*), socialização (*agreeableness*), realização (*conscientiousness*) e abertura (*openness*). Uma descrição mais completa está disponível em [Mccrae 1992]. Esses traços podem ser medidos de diversas formas, por exemplo, testes, entrevistas, observação e aspectos linguísticos [Golbeck *et al*. 2011a]. Alguns trabalhos também utilizam características textuais (uso de pronomes, pontuação, tempo verbal, etc.) na identificação dos traços do Big-five em textos em inglês, conforme demonstrado em [Mairesse *et al*. 2007].

Trabalhos anteriores examinaram a relação entre os traços de personalidade e as mídias sociais. [Quercia *et al*. 2011] investigaram como características topológicas podem ser usadas para predizer traços de personalidade. Eles obtiveram resultados promissores na predição de personalidade no Facebook, como também algumas descobertas interessantes, por exemplo, o fato de todos os usuários no experimento apresentarem baixos valores para neuroticismo e elevados para extroversão. [Adali *et al*. 2012] e [Golbeck *et al*. 2011b] encontraram resultados semelhantes.

3. Predizendo Influenciadores

Exercer influência é uma capacidade humana que transcende as redes sociais, isto é, um indivíduo influente pode mobilizar pessoas em qualquer ambiente social, seja ele virtual ou real. No Twitter, em particular, alguns atributos são caracterizados como indicadores de influência, tais atributos apontam padrões para identificar influenciadores nesta RSO.

Sabendo-se que todas as ações de um indivíduo são motivadas pela sua personalidade, este trabalho tem por objetivo relacionar os indicadores de influência com os traços de personalidade no Twitter, possibilitando a identificação de um padrão comportamental comum refletido pelos influenciadores.

Nossa contribuição difere de trabalhos anteriores nos seguintes aspectos: (1) foram utilizados traços de personalidade na predição de usuários influentes; (2) avaliaram-se os traços de personalidade como atributos preditores de cada um dos oito indicadores de influência elencados neste trabalho; (3) o experimento foi realizado em dois grupos distintos de usuários, os usuários comuns e aqueles que seguem celebridades. Acredita-se também que esse é o primeiro estudo relacionando os cinco grandes traços de personalidade como preditores de influência no Twitter.

O problema de predição foi elencado como uma tarefa de regressão linear. Assim, as funções são modeladas de forma a prever o resultado para cada indicador de influência, dados os valores das variáveis independentes - traços de personalidade. Antes de realizar a análise de regressão, investigaram-se quais indicadores poderiam

melhor representar influência no Twitter. Nas próximas subseções, são detalhados os datasets coletados, os indicadores de influência e os atributos escolhidos como preditores.

3.1. Datasets

Um objetivo secundário também verificado neste trabalho foi investigar possíveis características específicas em usuários que seguem celebridades, de forma a diferenciá-los dos outros. Dois *datasets* foram coletados fazendo uso da *Application Programming Interface* (API) do Twitter. Para cada conjunto de dados foram coletados perfis de usuário e seus últimos 200 *tweets*. Em ambos os conjuntos de dados, filtrou-se pelo idioma (Inglês), pela privacidade da conta (pública) e número de *tweets* postados (mais de 200). Ao coletar o conjunto de dados celebridade, foram inseridos os seguidores de Oprah Winfrey, Barack Obama, Mitt Romney, Justin Bieber e Lady Gaga em uma lista. Após o processo de filtragem em todos os usuários desta lista, foi aplicado um método de seleção aleatória para obter 3253 usuários e seus 650600 *tweets* publicados entre julho de 2012 e outubro de 2012.

No desenvolvimento do *dataset* de usuários comuns, escolheu-se um único usuário para obter todos os seus amigos, repetindo-se esta ação iterativamente para cada amigo do amigo, até obter um milhão de usuários (somente os identificadores de cada usuário). Além do filtro comum foram excluídos usuários que seguiam celebridades, sendo estas últimas caracterizadas por número de seguidores maior que quinze mil. Em seguida, através de um método aleatório foram obtidos 3634 usuários e seus 726800 *tweets*, publicados entre março de 2012 e julho de 2012. É calculado o desvio padrão, variância, média, mediana, valor mínimo e valor máximo para todos os preditores em ambos os *datasets* (Tabela 1). É importante ressaltar que nenhum *dataset* público foi encontrado com os requisitos necessários para realizar nossos experimentos, sendo possível obter os dados coletados neste trabalho através de contato com os autores.

Tabela 1. Análise estatística dos preditores em ambos os *datasets*

		Extro	Neuro	Socia	Reali	Abert
Dataset Celebridade	D.Padrão	0.36	0.479	0.282	0.314	0.357
	Variância	0.13	0.229	0.079	0.098	0.128
	Média	4.78	3.991	4.634	4.761	4.694
	Mediana	4.77	4.019	4.64	4.779	4.721
	Mínimo	3.13	2.002	3.285	3.468	2.133
	Máximo	8.12	5.626	5.625	6.126	6.22
Dataset Comum	D.Padrão	0.39	0.505	0.315	0.343	0.42
	Variância	0.15	0.255	0.099	0.118	0.176
	Média	4.78	3.966	4.643	4.767	4.68
	Mediana	4.77	3.982	4.637	4.78	4.704
	Mínimo	2.94	1.48	2.309	3.042	1.846
	Máximo	7.06	5.847	6.395	6.349	6.092

3.2. Indicadores de Influência

Uma variedade de indicadores tem sido avaliados para detecção de influência no Twitter e, dentre estes, foi possível selecionar os melhores indicadores avaliados nos trabalhos relacionados. Os indicadores apresentados são baseados no perfil do usuário, nas

estatísticas dos seus *tweets* e na sua interação na rede – todos os indicadores assumem valores numéricos.

Como mencionado na Seção 2, alguns trabalhos relatam o número de seguidores, amigos e retuite*s* como fortes medidas de influência. No entanto a quantidade de seguidores e amigos pode indicar apenas a popularidade do usuário, não refletindo assim sua capacidade de exercer influência. Assim, outras medidas de influência foram inseridas na avaliação a fim de distinguir usuários populares e influenciadores. A seguir, são apresentados os demais indicadores avaliados nesta proposta, bem como uma descrição sobre cada um destes.

Menção. O número de menções recebidas por cada usuário foi coletado diretamente de seu perfil, sem análise do texto do *tweet*. O número de menções é dividido pelo número de seguidores, a fim de balancear usuários com as distintas quantidades de seguidores.

Alcance do retuite. A quantidade de retuites pode indicar quão relevante é a informação publicada. No entanto essa afirmação pode não ser verdade para usuários populares, visto que o número de seguidores possa impulsionar a propagação de retuites [Weng *et al*. 2010]. Assim, para medir o alcance de cada retuite, dividiu-se o número de retuite*s* pelo número de seguidores.

Respostas. Outro fator que pode representar a influência de um usuário é a avaliação quantitativa da interação entre usuários do Twitter. O serviço permite responder diretamente o *tweet* de um usuário, dessa forma, se um usuário tem um grande número de respostas para seu *tweet*, isso pode significar que o conteúdo do *tweet* publicado é interessante ou o usuário é popular. As respostas foram adotadas como um indicador, sendo este obtido pelo número de respostas recebidas de todos os *tweets* publicados por um usuário.

Favoritados. Pressupõe-se que quanto maior o número de *tweets favoritados*, maior será a influência deste usuário. Por isso, considerou-se o número de *tweets* marcados como favoritos como outro indicador, sendo este calculado pela razão entre o número de *tweets* favoritados e o número de seguidores, para assim usuários populares não terem vantagem sobre os outros.

Atividade. O conteúdo dos *tweets* e as informações do usuário são medidas importantes para determinar a influência de um perfil de usuário [Naveed *et al*. 2011]. No entanto, os usuários com alta frequência de postagens na rede são mais susceptíveis a atrair novos seguidores. A atividade de um usuário é medida por meio da divisão entre o número total de *tweets* já publicados pelo usuário e a quantidade de meses desde a criação do perfil. Como já mencionado, não foi analisado conteúdo das mensagens para obter os atributos aqui listados.

3.3. Variáveis Independentes

Este trabalho considera os cinco traços de personalidade do modelo Big Five [Mccrae 1992] como atributos do modelo de predição para cada indicador de influência. Para quantificar cada traço de personalidade, foi aplicado o método descrito em [Mairesse *et al*. 2007] que classifica automaticamente personalidade a partir de textos escritos em inglês. Para quantificar cada traço de personalidade disponível na Figura 1, construiu-se modelos de regressão para os cinco traços. Foi utilizado à implementação do algoritmo

SMOreg do WEKA (HALL, 2009), que é uma adequação de Support Vector Machine (SVM) para análise de regressão. Para o algoritmo SMOreg foi escolhido o kernel polinomial com validação cruzada de 10 (*fold*). Os modelos recebem como entrada o conjunto de tuítes de um usuário para pontuar os traços de personalidade do usuário em uma escala entre 0 e 10.

Traço de Personalidade	Valores altos	Valores baixos
Extroversão	Extrovertido	Introvertido
Neuroticismo	Estressado	Emocionalmente estável
Socialização	Compassivo	Competitivo
Realização	Metódico	Desorganizado
Abertura	Criativo	Convencional

Figura 1. Traços de personalidade do modelo *Big Five.*

4. Modelo de Predição e Resultados

Nesta seção, é descrito e avaliado o modelo de regressão nos dois *datasets* descritos na secção 3.1. Inicialmente realizou-se uma correlação individual entre todos os atributos com o objetivo de compreender as relações existentes entre eles. Posteriormente, os algoritmos de regressão são utilizados para aprender os modelos de regressão para cada indicador de influência. A seguir são apresentados os modelos de predição, bem como a avaliação dos resultados da correlação de atributos.

4.1. Análise de Correlação

Neste trabalho, o método *Pearson product-moment* foi aplicado para medir o coeficiente de correlação entre os indicadores de influência e as variáveis independentes. A correlação de Pearson consiste em uma medida da relação linear entre duas variáveis aleatórias e possui intervalo entre -1 e 1. No resultado do método, disponível na Tabela 2, baixos coeficientes de correlação podem ainda ser estatisticamente significativos. Por isso, um teste de significância estatística, P-valor, foi aplicado para assegurar que cada correlação entre os atributos fosse significativa. Assumiu-se para valores de P menores que 0,05 como correlação moderada, enquanto que valores de P inferiores a 0,01 como uma correlação forte.

Extroversão. Os resultados confirmam forte correlação positiva do atributo extroversão com o número de amigos e o número de seguidores no *dataset* de usuários comuns. Este fato é facilmente explicado, pois usuários extrovertidos podem facilmente estabelecer novos relacionamentos, enquanto em introvertidos ocorre o inverso. A correlação negativa entre extroversão e atividade é um resultado surpreendente, já que [Golbeck *et al*. 2011a] relata os extrovertidos como possíveis aficionados pelo Twitter. O traço extroversão está fortemente correlacionado com o indicador favoritados no *dataset* de usuários comuns, porém o mesmo preditor não possui desempenho igual no *dataset* celebridade.

Neuroticismo. No traço de neuroticismo é geralmente o inverso de extroversão, fato este confirmado pela alta correlação negativa em ambos os *datasets* com o número de amigos e o número de seguidores. Outro resultado esperado é a correlação negativa entre neuroticismo e número de respostas, o que significa que as pessoas emocionalmente estáveis (baixos valores em neuroticismo) tendem a interagir mais que

indivíduos estressados (alta em neuroticismo). A relação estatisticamente mais significativa foi identificada entre atividade e neuroticismo (-0,15 e -0,21), o que indica que os neuróticos postam baixa quantidade de mensagens no Twitter. Este fato pode ser explicado pela seguinte cadeia de eventos: um usuário neurótico publica poucos *tweets*, vagamente interage com outros usuários e portanto tem poucos amigos.

Socialização. Indivíduos com valores elevados para este traço de personalidade são carismáticos e amigáveis, características supostamente atreladas a pessoas influentes. No entanto correlações positivas e significativas não foram encontradas entre este traço e os indicadores de influência, de forma unânime nos dois *datasets*.

Realização. Altos valores de correlação negativa foram encontrados entre os atributos de interação (respostas e menções) e o traço realização nos dois *datasets* experimentais. Isso sugere que quanto mais organizado e metódico (altos valores para realização) o usuário é, menos ele usa o Twitter como um serviço de bate papo (interação social). Tal hipótese sugere também que influenciadores apesar de interagirem pouco, exercem influência usando este RSO apenas para propagação de ideias e/ou notícias.

Abertura. Os indicadores respostas e menções possuem correlação positiva significativa com o traço de abertura, implicando que usuários imaginativos/criativos são mais correspondidos que os tradicionais/conservadores. Esse fato embasa o senso comum que influenciadores estão abertos a novas perspectivas e não temem o fracasso ao inovar. O indicador seguidores está correlacionado positivamente com este traço de personalidade em ambos os *datasets*, enquanto que o indicador amigos obteve coeficientes menos expressivos e ambíguos (-0.06 e 0.01). Esta diferença pode ser explicada pelo fato de influenciadores gerarem maior atração social, possuindo assim maior número de seguidores.

Tabela 2. Correlação de Pearson entre atributos preditores e indicadores de influência em ambos *datasets* (* = significante para p < 0.05; ** = p < 0.01)

Dataset Usuários Comuns								
Traço	**Amigo**	**Seguid**	**Favori**	**Retuit**	**Respos**	**Mençao**	**Ativiv**	**Alcanc**
Extro	0.13**	0.16**	0.02	0.05**	0.08**	0.1**	-0.03	-0.06**
Neuro	-0.24**	-0.21**	0.07**	-0.17**	0.21**	0.04	-0.15**	0.06**
Socia	0.16**	0.17**	-0.1**	0.12**	-0.08**	-0.039*	0.06**	-0.1**
Reali	0.21**	0.23**	-0.05**	0.13**	-0.16**	-0.05**	0.13**	-0.14**
Abert	-0.06**	0.12	0.16**	0.07**	0.21**	0.17**	-0.06**	0.05*
Dataset Seguidores Celebridade								
Traço	**Amigo**	**Seguid**	**Favori**	**Retuit**	**Respos**	**Mençao**	**Ativiv**	**Alcanc**
Extro	0.09**	0.002	0.17**	-0.03*	0.22**	0.17**	-0.12**	-0.02*
Neuro	-0.09**	-0.23**	0.25**	-0.15**	0.22**	0.01	-0.21**	0.05**
Socia	0.06**	0.09**	-0.08**	0.01	-0.05**	0.007	0.08**	-0.06**
Reali	0.12**	0.17**	-0.003	0.12**	-0.07**	0.02*	0.06**	-0.06**
Abert	0.01*	0.11**	0.06**	0.15**	0.04**	0.1**	-0.03*	0.04**

4.2. Análise de Regressão

Após análise e descoberta de correlações significativas nos dois *datasets* experimentais, decidiu-se aplicar métodos de regressão linear múltipla para modelar a relação entre cada indicador de influência (variáveis dependentes) e os cinco traços de personalidade

(variáveis independentes). Para cada medida de influência, a análise de regressão constrói uma função que pode ser representada pela equação geral (1):

$$Y_i = W_1\text{Extro} + W_2\text{Neuro} + W_3\text{Socia} + W_4\text{Reali} + W_5\text{Abert} + C \quad (1)$$

Onde Y_i é o indicador de influência, C é o valor constante e W_1 até W_5 são os pesos associados aos cinco traços de personalidade. Estes pesos podem ser positivos ou negativos, implicando assim em relação diretamente ou inversamente proporcional, respectivamente.

Quatro métodos de regressão foram utilizadas através da ferramenta de aprendizagem de máquina WEKA [Hall *et al*. 2009], com validação cruzada de 10 partes, repetidas em 10 iterações. As configurações padrão foram usadas para cada um dos seguintes algoritmos: (1) *zeroR*, prevê a classe majoritária, a média para valores numéricos; (2) *SMOreg*, *Support Vector Machine* para análise de regressão com o kernel polinomial; (3) Regressão Linear, algoritmo de regressão linear com critério de seleção do modelo; (4) M5P que gera uma árvore de decisão convencional, com a incorporação de funções de regressão linear para suas folhas.

A função de regressão gerada associa os pesos para cada preditor, em que os valores positivos indicam relação diretamente proporcional, enquanto que valores negativos indicam uma relação inversa.

4.3. Precisão do Modelo de Regressão

No que se segue, a Tabela 3 apresenta o erro quadrático médio (RMSE em inglês), medida que diferencia valores preditos pelo algoritmo dos valores reais observados. Com a verificação do RMSE, deseja-se avaliar a precisão do modelo de predição gerado por cada algoritmo. Aplicou-se o teste t de Student para comparar o algoritmo ZeroR com todos os outros, a fim de verificar se estes realmente estavam aprendendo o modelo melhor que a média (ZeroR). Este teste estatístico calcula a diferença entre os valores de RMSE, utilizando limite de significância de 0,01 (intervalo de confiança de 0,99). Na Tabela 3, os valores em negrito indicam se o RMSE do algoritmo é estatisticamente diferente do RMSE base (ZeroR).

Tabela 3. Valor do RMSE dos quatro algoritmos para cada indicador de influência nos dois *datasets*

	***Dataset* Usuários Comuns**				***Dataset* Seguidores Celebridades**			
	ZeroR	SMOreg	L.Reg	M5P	ZeroR	SMOreg	L. Reg	M5P
Ativi	1.43	**1.42**	**1.42**	**1.42**	1.81	**1.75**	**1.75**	**1.75**
Favor	2.14	**2.11**	**2.14**	**2.09**	2.24	**2.15**	**2.16**	**2.12**
Segui	2.12	**2.07**	**2.07**	**2.04**	2.21	**2.15**	**2.19**	**2.09**
Amigo	1.51	**1.45**	**1.45**	**1.45**	1.30	**1.29**	**1.29**	**1.29**
Mença	48.40	**46.17**	**46.29**	**46.14**	39.99	**38.52**	**39.25**	**38.73**
Alcan	1.65	**1.63**	**1.63**	**1.63**	1.62	**1.60**	**1.60**	**1.60**
Respo	48.26	**42.29**	**43.05**	**40.55**	37.37	**36.08**	**36.73**	**35.33**
Retui	1.93	**1.89**	1.92	**1.89**	2.07	2.05	2.07	**1.99**

Nos experimentos, não foram encontradas pontuações de RMSE do ZeroR menor do que os outros algoritmos, mas em alguns casos o teste t indica nenhuma diferença estatística entre eles.

5. Discussão

Foram utilizados dois métodos para investigar a capacidade de traços de personalidade em prever indicadores de influência no Twitter. A correlação de Pearson forneceu algumas suposições para a relação individual entre os medidores de influência e os atributos preditores (indicadores de influência *versus* traços de personalidade). Dentre os resultados promissores encontrados na análise de correlação, alguns desses foram surpreendentes se comparado a literatura (ver Seção 4.1).

Após aplicação do teste t, verificou-se que os algoritmos de regressão produziram modelos com precisão estatística significativa. Os modelos produzidos para sete indicadores possuiam valor de erro (RMSE) estatisticamente menor do que a medida basal (média), exceto para o indicador retuite. Desta forma, os resultados sugerem que o modelo aqui proposto não produz predições de influência aleatórias. Em ambos os conjuntos de dados, pode-se observar melhores resultados nos indicadores respostas, favoritados e menções. Os resultados obtidos por meio do algoritmo de regressão linear do WEKA estão alinhados com os encontrados nos outros dois algoritmos. Como este algoritmo exclui do modelo preditores pouco relevantes (critério de seleção do modelo), decidiu-se utilizar as funções geradas por este algoritmo na discussão dos resultados. Ao relacionar conjuntamente os traços de personalidade com cada indicador de influência, foi possível observar que:

- Extrovertidos são susceptíveis a serem influentes: pesos relativamente altos e positivos foram obtidos nos sete modelos para este preditor (exceto para o indicador retuites no *dataset* usuários comuns), o que parece confirmar que as pessoas comunicativas e espontâneas tem melhor prospecção social. Apenas para o indicador retuite este traço não obteve relevância no modelo gerado, confirmando mesmo resultado na correlação individual.
- Neuroticismo não impacta significativamente no indicador favoritados para os seguidores de celebridades. No entanto, correlação negativa foi observada entre eles (ver Seção 4.1). Assim, pode-se entender que este traço de personalidade é propenso a atuar isoladamente, isto é, se combinado com outros atributos ele provavelmente não terá peso sobre o indicador favoritados. O modelo gerado para os indicadores alcance do retuite, menção e respostas não considerou o fator neuroticismo, assim podemos presumir que este traço não é significante para atributos de interação se analisado conjuntamente. Para os demais indicadores, a relação negativa significante confirma os neuróticos como menos propensos a serem influentes.
- Sete indicadores consideraram o traço socialização no modelo gerado, no qual apenas para o indicador atividade no *dataset* usuários comuns não houve significância. Dentre os sete indicadores, aqueles de interação (menção, respostas e favoritados) obtiveram polaridade negativa e o peso associado relativamente alto, sendo positivo para os demais indicadores. Assim, a

interpretação sugere que pessoas competitivas (baixos valores no traço socialização) tem maior facilidade de interação e assim fazem-se influentes.

- O traço de realização presente nos modelos de regressão obteve peso associado baixo em sete indicadores e polaridade inconstante para o mesmo indicador em ambos *datasets*. Dessa forma não é possível associar este traço aos influenciadores quando este for analisado conjuntamente com os demais traços de personalidade.
- Pessoas imaginativas (altos valores em abertura) são prováveis influenciadores, visto que o peso associado a este traço foi alto e positivo em todas as funções geradas. Em ambos os *datasets* este traço teve peso relativamente maior que os demais para o indicador atividade, implicando na relação direta entre ser criativo e publicar mais conteúdo no Twitter. Estas observações corroboram a tendência que pessoas curiosas e dispostas a enfrentar novas situações serem mais atraentes socialmente.

6. Conclusão

Este trabalho apresentou medidas para quantificar a influência de um usuário no Twitter, podendo estas serem preditas a partir de traços de personalidade. Os resultados obtidos também ajudaram a entender o comportamento do usuário em um dos serviços de *microblogging* mais populares, o Twitter. Foram apresentadas correlações significativas entre indicadores de influência e traços de personalidade. Os modelos de regressão também foram utilizados para predizer cada indicador de influência, sendo estes avaliados posteriormente. O resultado da avaliação de acurácia foi encorajador para sete dos oito indicadores. Forte correlação entre traços de personalidade e as atributos do Twitter (indicadores) também demonstraram o impacto da personalidade na forma como o usuário geri sua rede social. Além disso, foi possível apontar algumas contribuições para o entendimento do padrão de uso no Twitter, como também associar traços de personalidade a influenciadores. Como o Twitter possui algumas características semelhantes com outras RSO, os resultados aqui podem ser estendidos. Contudo, experimentos adicionais são requeridos a fim de atestar tal fato.

Apesar dos resultados animadores encontrados, o presente trabalho tem algumas limitações. A primeira é que o *dataset* de usuários comuns foi coletado a partir de um usuário específico, escolhido arbitrariamente. Isso pode ter introduzido um viés na seleção para o processo de predição. Em segundo, ocorreu os mesmos inconvenientes relatados em [Mairesse *et al*. 2007], uma vez que apenas foi utilizado o texto para classificar os cinco traços de personalidade. Nos experimentos realizados neste trabalho, foi utilizado o texto de todos os *tweets* do usuário para classificar a personalidade deste, o que pode ter interferido igualmente na precisão dos resultados.

Este trabalho prosseguirá com uma discussão da ampliação da abordagem para outros grupos de usuários do Twitter. Pretende-se também avaliar outro modelo de personalidade com o objetivo de investigar sua relação com indicadores de influência. Finalmente, os resultados e perspectivas aqui apontados são valiosos para o Twitter. No entanto, é necessário examinar se as percepções sobre a influência apontadas neste artigo podem ser extensíveis a outras RSO, como o Facebook. Isso pode ajudar no entendimento do fenômeno de influência como um conceito amplo em redes sociais.

Referências

Adali, S. and Golbeck, J. (2012) Predicting Personality with Social Behavior. In: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 302-309. ASONAM, Istanbul.

Anger, I. and Kittl, C. (2011) Measuring influence on twitter. In: Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies, pp. 31. ACM.

Bakshy, E., Hofman, J. M., Mason, W. A. and Watts, D. J. (2011) Everyone's an influencer: quantifying influence on twitter. In: Proceedings of the fourth ACM international conference on Web search and data mining, pp. 65-74. ACM.

Berry J. and Keller E. (2003) The influentials: one American in ten tells the other nine how to vote, where to eat, and what to buy. Free Press, New York.

Cha M., Haddadi H., Benevenuto F. and Gummadi K.P. (2010) Measuring user influence in Twitter: the million follower fallacy. In: 4th AAAI International Conference on Weblogs and Social Media. ICWSM, Washington.

Golbeck, J., Robles, C., Edmondson, M. and Turner, K. (2011) Predicting Personality from Twitter. In: 3th IEEE International Conference on Social Computing, pp. 149-156. SocialCOM, Boston.

Golbeck, J., Robles, C. and Turner, K. (2011) Predicting Personality with Social Media. In: Proceedings of the 2011 annual Conference on Human Factors in Computing Systems, pp. 253-262. CHI, Vancouver.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. (2009) The WEKA data mining software: An update. In: ACM SIGKDD Explorations Newsletter, pp. 10-18.

Mairesse, F., Walker, M., Mehl, M. and Moore, R. (2007) Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. In: Journal of Artificial Intelligence Research, pp. 457–500.

Mccrae, R.R. (1992) The five-factor model: Issues and applications (Special issue). In: Journal of Personality, pp. 175-215.

Naveed, N., Gottron, T., Kunegis, J. and Alhadi, A.C. (2011) Bad news travel fast: A Content-based Analysis of Interestingness on Twitter. Retrieved from http://www.websci2011.org/fileadmin/websci/Papers/2050_paper.pdf.

Quercia, D., Kosinski, M., Stillwell, D. and Crowcroft, J. (2011) Our Twitter Profiles, Our Selves: Predicting Personality with Twitter. In: 3th IEEE International Conference on Social Computing, pp. 180-185. SocialCOM.

Weng, J., Lim, E. P., Jiang, J. and He, Q. (2010) Twitterrank: finding topic-sensitive influential twitterers. In: Proceedings of the third ACM international conference on Web search and data mining, pp. 261-270.