aper:180198_1

Análise da Popularidade de Tuítes com Base em Características Extraídas de seu Conteúdo

Lucas L. de Oliveira[1], **Sérgio L. S. Mergen**[2]

[1]Centro de Tecnologia – Universidade Federal de Santa Maria (UFSM)
97105-900 -– Santa Maria -– RS — Brasil

[2]Departamento de Linguagens e Sistemas de Computação
Universidade Federal de Santa Maria (UFSM)

{loliveira,mergen}@inf.ufsm.br

Abstract. *Twitter is a powerful platform for opinions diffusion. Knowing this, companies and influential personalities use the platform as a way to connect with their audience, with the goal of boosting their popularity. On Twitter, the tools which are able to measure popularity are the retweets and likes that each post receives. In this paper, the purpose is to analyze the content of the messages transmitted through the platform and correlate them with their popularity. Characteristics of the messages such as the feeling, its size in characters, banality and the use of URLs or hashtags are evaluated. It was possible to identify, in a general way, a preference for tweets with a good feeling and median banality, while the message's size had a different influence on retweets and likes.*

Resumo. *O Twitter é uma plataforma poderosa para a difusão de opiniões. Sabendo disso, empresas e personalidades influentes usam a plataforma como um meio de se conectar com seu público com o objetivo de alavancar a popularidade. No Twitter, os mecanismos capazes de medir a popularidade são os retuítes e curtidas que cada publicação recebe. Neste trabalho, o objetivo é analisar o conteúdo das mensagens veiculadas através da plataforma e correlacioná-las com sua popularidade. Foram avaliadas características da mensagem como o sentimento, tamanho em caracteres, banalidade e o uso de URLs ou hashtags. Foi possível identificar, de maneira geral, uma preferência do público por tuítes com sentimento positivo e banalidade mediana, enquanto o tamanho das mensagens influenciou de maneira diferente nos retuítes e curtidas.*

1. Introdução

O Twitter é um meio de veiculação de mensagens que se destaca por sua simplicidade e objetividade. Segundo o portal de estatísticas Statista [1], é a 11º rede social mais utilizada no mundo, usado por mais de 330 milhões de usuários diariamente.

Uma das preocupações de usuários do Twitter é alavancar sua popularidade, através do aumento no número de seguidores. Essa preocupação é fundamental para empresas e personalidades públicas que utilizam suas imagens para fins monetários. Nesses casos, o uso das redes sociais deve ser planejado e monitorado. Quando isso é realizado da maneira correta, a marca e/ou a pessoa ficam muito mais próximos de seus fãs e seguidores, o que consequentemente, faz sua popularidade e influência aumentar.

[1]*Statista:* https://www.statista.com/topics/737/twitter/

Como pode ser observado no trabalho de [Cha et al. 2010], um dos fatores que mede a influência de um usuário do Twitter é a quantidade de retuítes que ele recebe. Levando isso em consideração, pode-se afirmar empiricamente que o aumento na quantidade de retuítes leva a um aumento na quantidade de seguidores, devido a propagação exponencial daquele conteúdo.

Como afirma [Suh et al. 2010], a propagação de um tuíte está diretamente ligada ao conteúdo e valor informativo contido nele. Nesse sentido, os autores avaliaram um conjunto de características extraídas das mensagens. Os resultados mostraram que a presença de *hashtags* e URLs são os fatores que mais ajudam a impulsionar uma publicação. Apesar de ser um resultado relevante, o trabalho não realizou uma análise exaustiva das características que se pode extrair das mensagens.

Nesse contexto, este trabalho realiza uma análise para verificar a influência de determinadas características sobre a popularidade dos tuítes, das quais três delas não foram contempladas pelo estudo de [Suh et al. 2010]: o tamanho em caracteres, o sentimento (que mede a emoção transmitida) e a banalidade (que mede a relevância da mensagem). Para fins de comparação, a presença de *hashtags* e URLs também foi avaliada.

Este artigo está estruturado nas seguintes seções. A seção 2 apresenta os trabalhos relacionados. A seção 3 apresenta a arquitetura de extração de tuítes usada, que realiza desde a coleta até a preparação dos dados para análise. A seção 4 apresenta as análises realizados a partir dos dados coletados. A seção 5 apresenta as considerações finais.

2. Trabalhos Relacionados

Vários autores já apresentaram em seus trabalhos razões pelas quais empresas tornam-se cada vez mais interessadas na utilização de mídias sociais, como Twitter e Facebook. O interesse visa melhorar a comunicação com o consumidor, ou público alvo. Porém, a simples utilização destes serviços online não é suficiente para agregar valor de mercado ao negócio [Culnan et al. 2010]. É preciso de estratégia para que este tipo de intervenção gere bons resultados. Tratando-se da utilização do Twitter, a influência de uma conta pode estar diretamente ligada à relevância do conteúdo por ela disseminado [Valiati et al. 2012].

Nesse sentido, o trabalho de [Suh et al. 2010] realiza uma análise em larga escala (com mais de 74 milhões de registros coletados) sobre fatores que impactam no índice de redistribuição de tuítes. Foram considerados fatores como a utilização de URLs e *hashtags* distintas, quantidade de seguidores e amigos (contas sendo seguidas), o tempo de existência da conta e a quantidade de tuítes antigos do autor. Através de análises observou-se que, com exceção da quantidade de tuítes antigos, os demais fatores interferem na probabilidade de redistribuição. De acordo com os estudos, os fatores que exercem maior influência são o uso de URLs (que pode variar dependendo do domínio), o uso de *hashtags* e a quantidade de seguidores da conta.

O trabalho de [Suh et al. 2010] também propôs a criação de um modelo de predição, elaborado usando a Análise de Componentes Principais (PCA, sigla em inglês para *Principal Components Analysis*) e Modelagem Linear Generalizada (GLM, sigla em inglês para *Generalized Linear Modeling*). O resultado foi um conjunto de coeficientes aplicados em uma equação para prever a taxa de redistribuição dos tuítes. Esse modelo

corroborou com a descoberta sobre a influência das características dos tuítes sobre a taxa de propagação. Uma das características não analisadas pelo trabalho de [Suh et al. 2010] é o sentimento do tuítes. Os trabalhos de [Bigonha et al. 2012] e [Mehta et al. 2012] propõem a detecção do poder de influência do usuário, através de um modelo de cálculo que considera também a análise linguística dos dados coletados e a conexão entre os usuários. Estes estudos mostram que o sentimento pode sim ter uma ligação direta com o poder de influência de um determinado usuário, o que reforça e incentiva a realização do presente trabalho.

Já os trabalhos de [Agarwal et al. 2011] e [Lakshmi et al. 2017] propõem a elaboração de modelos capazes de classificar o sentimento de um tuíte em positivo, negativo ou neutro. Ambos os casos apresentam técnicas em que realizam a coleta, pré-processamento e classificação dos dados. Na etapa de pré-processamento, além de palavras, são também considerandos *emoticons*, acrônimos e letras repetidas, o que torna a classificação ainda mais precisa. Já a etapa de classificação foi baseada em modelos diversos, como *Naive Bayes* e *Tree Kernel*. Apesar de não usarem o sentimento para nenhum tipo de medição, os modelos são relevantes como uma forma alternativa de extração de características.

Outros trabalhos baseados em análises usando dados do Twitter são os de [Engel 2016] e [Tumasjan et al. 2010]. Ambos tem o intuito de relacionar a opinião dos usuários no Twitter com as eleições em seus países (Estados Unidos e Alemanha, repesctivamente). O trabalho desenvolvido por [Engel 2016] busca distinguir e exibir em tempo real o sentimento da população, baseada na sua localização, quanto aos candidatos a presidência. Já o trabalho de [Tumasjan et al. 2010] analisa a influência que a plataforma tem sobre as eleições. Neste trabalho realiza-se uma análise das mensagens considerando 12 dimensões do sentimento político. Esta análise é realizada através do *software* LIWC2007 (*Linguistic Inquiry and Word Count*), que avalia componentes emocionais, cognitivos e estruturais de amostras de texto. Tumasjan pôde concluir que estudos com dados do Twitter de fato servem como preditores do resultado de eleições, chegando perto até mesmo das pesquisas tradicionais.

3. Proposta

O propósito deste trabalho é correlacionar a popularidade dos tuítes em função de um conjunto de características. Para atingir esse objetivo, é necessário coletar os tuítes e extrair suas características. Esta seção apresenta a arquitetura de coleta e extração utilizada neste trabalho.

A arquitetura, exibida na Figura 1, tem os seguintes módulos: (a) **Coleta** dos tuítes publicados por cada uma das contas acompanhadas; (b) **Extração** das características de cada tuíte; e (c) **Atualização** periódica dos dados coletados.

3.1. Coleta de tuítes

O módulo de coleta é responsável por extrair tuítes de usuários específicos. A extração ocorre de forma contínua, usando recursos de *streaming* disponibilizados pela API do Twitter[2]. São coletados todos tuítes publicados a partir do momento que o *streaming* entra em execução.

[2]*Twitter Developer Platform:* `https://dev.twitter.com`

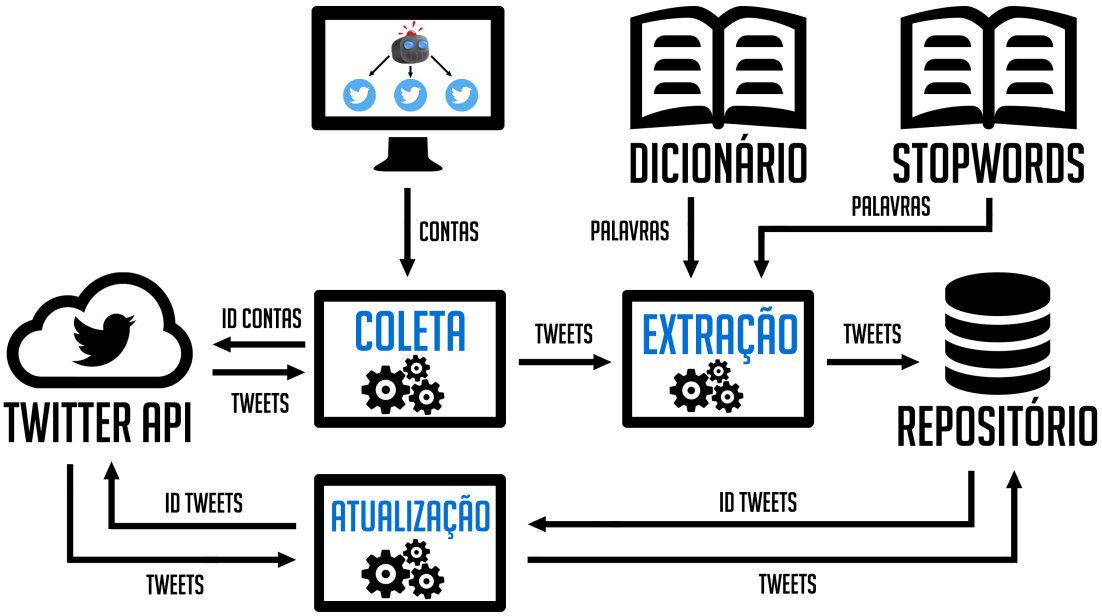


Figura 1. Arquitetura adotada no desenvolvimento do projeto

A especificação das contas a serem seguidas é feita a partir de uma conta raiz. São extraídos os tuítes de todos usuários seguidos por essa conta raiz. Essa estratégia permite que novas contas sejam adicionadas à lista sem que haja interrupções. O módulo também conta com tratamento de exceções para que a coleta não seja interrompida devido à problemas temporários de acesso aos dados, como indisponibilidade do serviço ou extrapolação do limite de requisições permitido por instante de tempo.

Na Tabela 1 podem ser visualizadas as informações extraídas de cada tuíte através da API. O campo "mensagem" é usado para a extração das características. Já os "retuítes" e "curtidas" são usados como forma de medir a popularidade do tuíte. Por sua vez, os campos "identificação" e "data/hora" são usados pelo módulo de atualização.

Tabela 1. Dados coletados para cada tuíte

Informação	Conteúdo
autor	código e nome da conta que originou o tuíte
identificação	código do tuíte (permite a consulta posterior)
mensagem	texto de no máximo 280 caracteres
data e hora	data e hora da publicação do tuíte em seu país de origem
retuítes	quantidade de retuíte que a mensagem recebeu
curtidas	quantidade de vezes que o tuíte foi favoritado

3.2. Extração de características de tuíte

Esta etapa corresponde a extração das características de cada um dos tuítes coletados. A extração ocorre imediatamente após a coleta. Os itens abaixo mostram como cada característica foi extraída.

Presença de URLs e *hashtags*: O uso desses recursos na mensagem é detectado pela presença de prefixos específicos no corpo da mensagem. Por exemplo, o prefixo

"http" indica que URLs foram usadas. Já o prefixo "#" denota o uso de *hashtags*.

Tamanho da mensagem: O tamanho é a contagem da quantidade de caracteres usados no corpo do tuíte. A contagem desconsidera caracteres usados em URLs, assumindo que *hiperlinks* não transmitam nenhuma mensagem. A remoção de URLs foi realizada a partir da aplicação de uma expressão regular.

Extração do sentimento: O sentimento de uma mensagem é um valor que classifica o teor da mensagem como positivo ou negativo. Para realizar a extração do sentimento, assim como no trabalho de [Engel 2016], foi utilizada a biblioteca TextBlob da linguagem Python [Loria et al. 2014]. Essa biblioteca permite a obtenção da polaridade e subjetividade de conteúdos textuais na língua inglesa. A API também fornece a possibilidade de tradução do conteúdo de textos escritos em outras linguagens.

A extração do sentimento se baseia em Árvores de Decisão e no modelo de classificação *Naive Bayes*, que também é utilizado no trabalho de [Lakshmi et al. 2017]. A classificação da mensagem retorna um valor decimal entre o intervalo de -1 e 1, onde -1 corresponde a uma mensagem totalmente negativa, 0 corresponde a neutra e 1 corresponde a totalmente positiva.

Extração da banalidade: No contexto deste trabalho, a banalidade corresponde à importância do que foi escrito. A forma adotada para medir a banalidade leva em consideração a presença de palavras que são frequentemente usadas em textos escritos. Quanto maior o número de palavras frequentes, mais banal é a mensagem.

A Equação 1 mostra como computar o valor da banalidade:

$$\frac{\sum_{i=1}^{n}(freq(P_i))}{n} \tag{1}$$

onde o conjunto $\{P_1, ..., P_n\}$ são as palavras da mensagem após a remoção de *stopwords* (preposições e artigos que normalmente são descartados durante o processamento de um texto). Já a função $freq(P)$ retorna 1 caso a palavra P seja frequente e zero caso não seja. A verificação da frequência utiliza um dicionário de palavras previamente construído. Também são removidas as *hashtags* e menções a outros usuários, por entender que não se tratam de palavras que podem ser caracterizadas como banais ou não.

Na Tabela 2 pode ser visto um exemplo dos dados extraídos nesta etapa.

Tabela 2. Dados obtidos na etapa de Extração

Informação	Conteúdo
sentimento	valor entre -1 e 1 correspondente a polaridade do texto
URL	valor 1 se houver URL no texto e 0 se não houver
hashtag	valor 1 se houver *hashtag* no texto e 0 se não houver
tamanho	quantidade de caracteres utilizados na mensagem
banalidade	somatório baseado na no uso de palavras frequentes

3.3. Atualização dos dados de retuítes e curtidas

Como o módulo de coleta funciona por meio de *streaming*, os tuítes são coletados no instante de sua criação. Nesse momento, a quantidade de retuítes e curtidas recebidos têm

o valor zero. Dessa forma, é necessária uma conferência periódica para a obtenção dos dados atualizados.

A atualização é realizada através de um recurso da API do Twitter que obtém informações de um tuíte a partir do seu código de identificação. Para evitar sobrecarga de processamento, apenas os tuítes publicados no intervalo de uma semana são atualizados. Como o status de tuítes mais antigos é raramente modificado, o acesso a eles seria ao mesmo tempo custoso e improdutivo.

4. Resultados

Esta seção apresenta análises que correlacionam características extraídas de tuítes com a sua popularidade. A importância das características é medida por dois indicadores: o número de retuítes e o número de curtidas.

A coleta de tuítes foi realizada tendo como base as contas de personalidades influentes que utilizam o Twitter periodicamente. Ao todo, foram usadas 23 contas de diversas áreas de atuação, como por exemplo Donald J Trump (atual presidente dos Estados Unidos), Jimmy Fallon (famoso apresentador de TV americano) e Katy Perry (cantora detentora da conta com o maior número de seguidores no Twitter). A escolha deve-se ao fato de que a análise do impacto de publicações em redes sociais é mais relevante para esse tipo de usuário. A etapa de coleta permaneceu em execução durante o período entre Novembro de 2017 e Fevereiro de 2018, totalizando cerca de 5500 registros.

Uma análise inicial sobre os dados gerou constatações que serviram para orientar o trabalho. Um dado interessante é que o número de curtidas é muito superior ao número de retuítes. Assim, esses dois indicadores são analisados de forma independente.

Outro dado interessante é que algumas características são mais empregadas do que outras. Por exemplo, 78% dos tuítes coletados possuem URLs em seu conteúdo, enquanto apenas 28% possuem *hashtags*. Para evitar distorções causadas por esse desbalanceamento, os indicadores por característica são normalizados. A Equação 2 mostra como normalizar o número de curtidas.

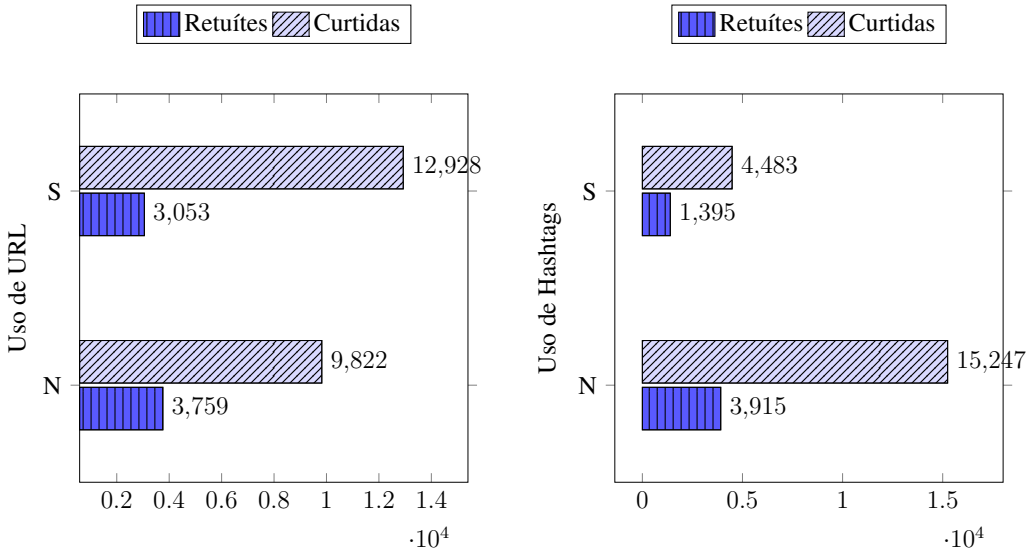
$$\frac{\sum_{i=1}^{n}(count_likes(T_i))}{n} \tag{2}$$

O cálculo usa o conjunto de tweets $\{T_i, ..., T_n\}$ onde a característica apareça. A função $count_likes(T)$ retorna o número de curtidas atribuídos ao tuíte T. Ou seja, o valor final corresponde a soma de todas as curtidas recebidas dividida pela quantidade de tuítes. A normalização do número de retuítes se baseia no mesmo princípio.

4.1. Popularidade x Existência de URL ou *hashtag*

O objetivo deste experimento é verificar se o uso de *hashtags* e/ou URLs contribui (de forma positiva ou negativa) para alavancar a popularidade de um tuíte.

A Figura 2 apresenta os resultados. Como pode-se ver na Figura 2(b), tuítes que não usaram *hashtags* foram mais populares (em retuítes e curtidas) do que aqueles que usaram esse marcador. O mesmo não se pode dizer quanto ao uso de URLs. Nesse caso, Figura 2(a), não há uma diferença significativa no número de retuítes. Já o número de curtidas foi menor para mensagens que não usaram URLs.



(a) Popularidade X uso de URLs (b) Popularidade X uso de *Hashtags*

Figura 2. Análise da taxa de popularidade pelo uso de URLs e *hashtags*. No eixo Y, N corresponde a Não e S corresponde a Sim

Apesar de pequena a diferença, esse resultado reforça o estudo publicado por [Suh et al. 2010], que considera a presença de URLs importante para que um tuíte seja redistribuído. Essa pequena diferença nos resultados pode ser atribuída ao uso de dados diferentes. Enquanto [Suh et al. 2010] usou dados de usuários aleatórios, neste trabalho os usuários foram escolhidos com base na sua popularidade. Assim, é possível que essa característica dependa do perfil do autor da publicação.

Em uma análise individual das contas, observou-se que esse comportamento pode ser inverso, onde o tuítes sem a presença de URLs recebem um maior número de curtidas e retuítes. Este foi o caso da conta do presidente dos Estados Unidos, Donald J. Trump.

4.2. Popularidade x Tamanho de Mensagem

O objetivo deste experimento é identificar a existência de uma relação entre a popularidade do tuíte e seu tamanho, em caracteres. Dessa forma, é possível saber se existe uma preferência por textos mais extensos ou mais enxutos.

A Figura 3 apresenta os resultados. O resultado mostra claramente que mensagens curtas, marcadas com "X" na Figura 3(b), receberam mais curtidas. A exceção ocorre para textos com tamanhos variando de 120 a 130 caracteres. Curiosamente, o número de curtidas nesse caso foi bastante superior aos demais tamanhos de texto. Também chama a atenção o fato que são mensagens cuja extensão se aproxima do tamanho máximo de texto que a plataforma do Twitter costumava disponibilizar (140).

Já o número de retuítes aparece bem distribuído em todas faixas de tamanho. Mesmo assim, percebe-se uma concentração maior em textos de tamanho próximo aos 140 caracteres, destacados com "X" na Figura 3(a). O motivo pode estar relacionado ao hábito de usar textos dentro desse intervalo. Corrobora para essa tese a constatação de que, mesmo tendo o tamanho máximo aumentado para 280 caracteres, as mensagens mais extensas coletadas não ultrapassaram os 190 caracteres.

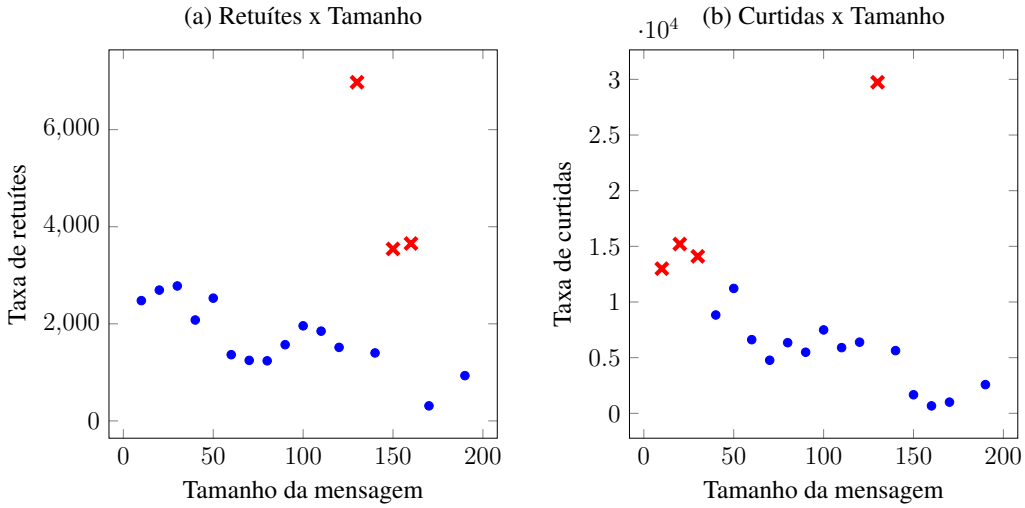


Figura 3. Análise da taxa de popularidade pelo tamanho de mensagens

4.3. Popularidade x Polaridade de Sentimento

Neste experimento, a popularidade do tuíte é confrontada com a polaridade do sentimento. Dessa forma, é possível saber a preferência dos seguidores por um conteúdo mais bem humorado ou mal humorado.

A Figura 4 apresenta os resultados. A polaridade do sentimento foi considerada até uma casa decimal para facilitar a visualização e desfragmentar as faixas de sentimento. Os resultados demonstram que tuítes "mal humorados" são os menos populares, tanto em número de curtidas quanto em número de retuítes. De 20% dos registros com menor popularidade, destacados com "X" nos gráficos, 3/4 deles estão no extremo negativo.

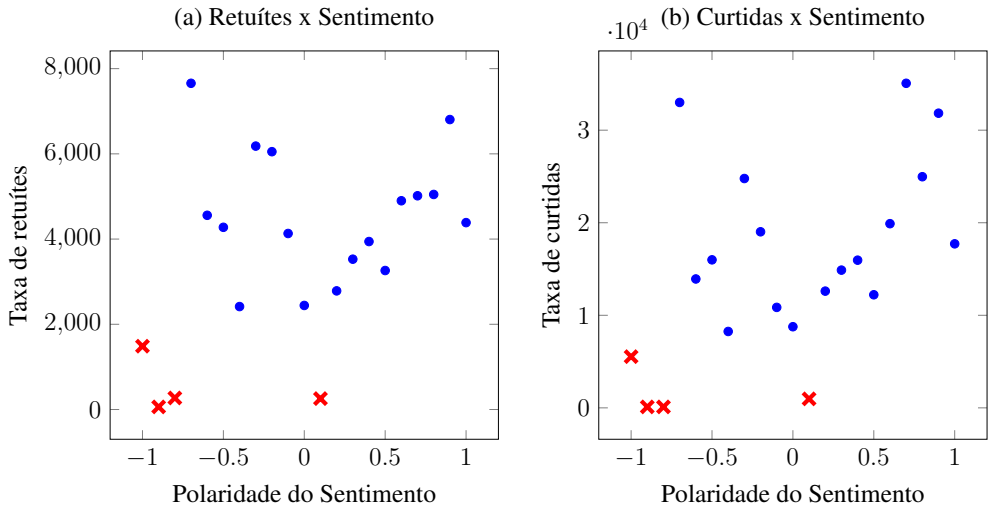


Figura 4. Análise da relação entre popularidade e polaridade sentimento.

Vale destacar, que na análise individual das contas, uma delas apresentou um comportamento no qual 25% dos tuítes com a menor taxa de polaridade encontrava-se nos dois extremos, tanto negativo quanto positivo (que também foi o caso da conta de Donald J. Trump). Apesar de esse resultado individual apontar para outra direção, ainda assim é

possível perceber que o sentimento exerce influência na popularidade, muito embora a correlação possa variar de conta para conta.

4.4. Popularidade x Banalidade da mensagem

Neste experimento o intuito é identificar se tuítes que fazem maior uso de palavras frequentes recebem mais ou menos retuítes e curtidas. Neste processo, foram consideradas apenas as mensagens escritas na língua inglesa. O dicionário usado possui 3.000 palavras frequentes desta linguagem[3].

A Figura 5 apresenta os resultados. Como pode-se ver, tanto os tuítes com taxa baixa e alta de banalidade são os menos populares, os quais também foram destacados com "X" nos gráficos. A impopularidade dos tuítes banais pode estar relacionada à ausência de originalidade das mensagens. Já a impopularidade dos tweets que estejam no outro extremo pode estar associada ao uso de uma linguagem pouco acessível ao público.

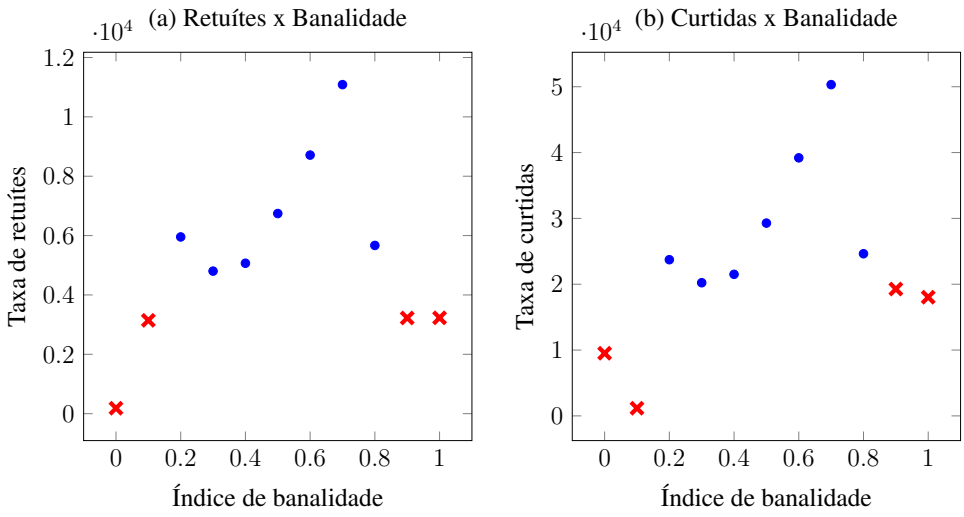


Figura 5. Análise da relação entre popularidade e banalidade da mensagem.

É interessante também analisar como as características aliadas influenciam na importância atribuída a um tuíte. Por exemplo, a Figura 3 sugere que tuítes curtos são os mais curtidos. Por outro lado, tuítes banais são pouco curtidos. Assim, possivelmente a importância de tuítes curtos seja intensificada caso seu conteúdo não seja banal. A verificação de hipóteses desse tipo depende de uma análise conjunta das características extraídas, o que foge do escopo deste artigo.

5. Considerações Finais

Medir a variação do índice de popularidade pode ser do interesse de administradores de grandes contas do Twitter, pois pode indicar o sucesso ou fracasso de uma determinada campanha realizada ou a dimensão de um escândalo e, tendo consciência disso, ações preventivas ou corretivas podem ser tomadas e sua repercussão pode ser monitorada. Tratando-se de personalidades públicas, é importante identificar quais assuntos e/ou

[3] *3000 most common words in English:* `https://www.ef.com/english-resources/english-vocabulary/top-3000-words/`

abordagens agradam mais o público-alvo para que assim possam ser mantidas ou evitadas. Assim, pode ser relevante entender quais características de um tuíte publicado interferem no interesse do público por aquele conteúdo.

Apesar de incipientes, os resultados alcançados mostram que as características avaliadas parecem exercer influência no nível de interesse, e podem ser levadas em consideração na elaboração de estratégias que visem impulsionar publicações.

Como trabalhos futuros, pretende-se avaliar a taxa de popularidade quando as características são combinadas entre si. Por exemplo, aliar a taxa de banalidade da mensagem com o seu tamanho pode fornecer um indicador de popularidade importante. Além disso, outra possibilidade de trabalho futuro envolve a elaboração de preditores de popularidade baseados nas características estudadas. Com preditores específicos para cada conta, seria possível verificar o alcance de uma determinada mensagem antes mesmo de publicá-la.

Referências

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R. (2011). *Sentiment Analysis of Twitter Data*. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 30–38, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bigonha, C., Cardoso, T. N. C., Moro, M. M., Gonçalves, M. A., and Almeida, V. A. F. (2012). *Sentiment-based influence detection on Twitter*. *Journal of the Brazilian Computer Society*, 18(3):169–183.

Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, P. K. (2010). Measuring user influence in twitter: The million follower fallacy. *Icwsm*, 10(10-17):30.

Culnan, M., J. McHugh, P., and I. Zubillaga, J. (2010). *How Large U.S. Companies Can Use Twitter and Other Social Media to Gain Business Value.* 9:243–259.

Engel, A. (2016). Election 2016 twitter sentiment map.

Lakshmi, V., Harika, K., Bavishya, H., and Harsha, C. S. (2017). *SENTIMENT ANALYSIS OF TWITTER DATA*.

Loria, S., Keen, P., Honnibal, M., Yankovsky, R., Karesh, D., Dempsey, E., et al. (2014). Textblob: simplified text processing. *Secondary TextBlob: Simplified Text Processing*.

Mehta, R., Mehta, D., Chheda, D., Shah, C., and Chawan, P. M. (2012). *Sentiment analysis and influence tracking using twitter*. *International Journal of Advanced Research in Computer Science and Electronics Engineering (IJARCSEE)*, 1(2):pp–72.

Suh, B., Hong, L., Pirolli, P., and Chi, E. H. (2010). *Want to be retweeted? large scale analytics on factors impacting retweet in twitter network*. In *Social computing (socialcom), 2010 ieee second international conference on*, pages 177–184. IEEE.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *Icwsm*, 10(1):178–185.

Valiati, H., Silva, A., Guimaraes, S., and Meira Jr, W. (2012). *Detecç ao de Conte udo Relevante e Usuários Influentes no Twitter*.