Compreendendo Mecanismos de Influência no Twitter através do Comportamento dos Usuários

Davi Zanotto, Carlos Kamienski

Universidade Federal do ABC (UFABC)
{davi.zanotto,cak}@ufabc.edu.br

Abstract. *Online social networks are increasingly used for analyzes that purport to understand how human relationships occur in these environments. In particular, users are influenced by others and show it in different ways, such as shares, comments and likes. Organizations would like to understand how is the mechanism of influence in social networks, however, it is not easy to characterize. This paper presents an extensive analysis of how the contents are broadcast on Twitter, what are the characteristics of influential users and the characteristics of messages that become viral. In all, more than 62 million tweets were collected between 2014 and 2015. The results shows that you can better use Twitter to get your broadcast content and also point out directions for future research.*

Resumo. *As redes sociais online cada vez mais são usadas para análises que se propõem a compreender como ocorrem as relações humanas nesses ambientes. Particularmente, usuários são influenciados por outros e demonstram isso de diferentes maneiras, como compartilhamentos, comentários e curtidas. Organizações de naturezas variadas gostariam de compreender como ocorre o mecanismo de influência nas redes sociais, que no entanto, não é simples de caracterizar. Neste trabalho é apresentada uma análise extensa de como os conteúdos são difundidos no Twitter, quais as características dos usuários denominados influentes e as características das mensagens que se tornam virais. Ao todo, foram coletados mais de 62 milhões de tuítes entre 2014 e 2015. Os resultados mostram que é possível usar melhor o Twitter para ter seu conteúdo difundido e também apontam direções para pesquisas futuras.*

1 Introdução

As primeiras iniciativas de marketing digital tiveram início no fim da década de 80 com a veiculação de banners nos primeiros serviços de assinatura de internet nos EUA, denominados *banner ad*. Atualmente, campanhas de marketing digital são também por redes sociais online (OSN – em inglês), como o Twitter[1]. A estratégia é criar artificialmente uma propaganda boca a boca entre os clientes potenciais, fazendo com que a marca seja divulgada de forma exponencial a partir de uma pessoa influente e de confiança, proporcionando aumento de credibilidade dessa propaganda. Tal estratégia é conhecida por marketing viral. Neste artigo, a utilização dos conceitos de marketing viral são utilizados com foco na difusão de informações e conteúdo de forma exponencial.

Este trabalho apresenta a implementação de um coletor de tuítes em *streaming* e a análise dos tuítes coletados por contagem e retuítes e menções, análise, através de histogramas, do comportamento das mensagens de acordo com os diferentes temas.

[1] https://twitter.com/

Almeja-se responder se é possível criar uma mensagem que atraia o interesse de usuários formadores de opinião em divulgá-la e provocar um comportamento viral e, para responder essa questão, este trabalho se propõe a estudar a OSN Twitter, realizar a coleta de mensagens trocadas entre os usuários, entender como as informações são distribuídas entre os usuários e identificar quais são os principais responsáveis pela disseminação de conteúdo em determinados assuntos, baseado nas funcionalidades desta rede social que têm o objetivo de espalhar informações. São elas: Seguidores, Retuítes e Menções.

Dentre os resultados obtidos, foram coletados mais de 62 milhões de tuítes durante o período entre 03/02/2014 e 13/01/2015 para 7 diferentes temas. Durante os experimentos de contagem dos tuítes, foi possível constatar que os usuários mais mencionados são pessoas públicas e celebridades e os usuários mais retuitados são blogueiros e pessoas comuns.

Na sequência do artigo, a seção 2 apresenta conceitos básicos para entendimento do artigo, com as características do ambiente de análise de influência em redes sociais online. Na seção 3 a metodologia utilizada para avaliar a proposta é descrita assim como os experimentos realizados e métricas utilizadas. Na seção 4 são apresentados os resultados obtidos pelos experimentos e na seção 5 é realizada a discussão sobre os resultados obtidos e lições aprendidas.

2 Conceitos Básicos

2.1 Marketing Viral

Uma das primeiras definições de marketing viral surgiu no boletim informativo do Nestcape, em 1997, como “rede boca a boca aprimorada”. Segundo (Jurvetson), a inspiração para o termo “marketing viral” surgiu originalmente a partir do padrão de anúncio adotado pelo Hotmail[2] que conseguiu aumentar sua rede de usuários de forma exponencial. O Hotmail incluiu um campo promocional com um link (URL clicável) em cada mensagem de e-mail enviada por um usuário de sua rede e, assim, cada cliente tornou-se um vendedor involuntário simplesmente usando o produto, fazendo com que a informação divulgada alcançasse o maior número de pessoas contidas em uma rede possível. Ainda, (Hill) define que o termo marketing viral está relacionado a qualquer estratégia que encoraja indivíduos a transmitir uma mensagem de marketing para outros, criando o potencial de crescimento exponencial da exposição e influência da mensagem.

Recentemente, o marketing viral está sendo vastamente explorado em OSN’s, por conta da concentração de usuários e distinção de vários nichos de mercado e interesses contidos nessas redes.

2.2 Redes Sociais Online

Dentre as definições de sites de redes sociais, (Ellison) define que são serviços baseados na web que permitem aos indivíduos construir um perfil público ou semi-público dentro de um sistema limitado; articular uma lista de outros usuários com quem eles compartilham uma conexão; e ver e percorrer a sua lista de conexões e aquelas feitas por outros dentro do sistema. A natureza e nomenclatura dessas conexões podem variar de site para site. Ainda segundo o autor, o que torna uma rede social única não é o fato de permitir que usuários conheçam estranhos, mas sim o fato de permitir que os

[2] http://www.live.com/

usuários possam se pronunciar e tornar visíveis em suas redes. E isso faz com que usuários conheçam outros a partir da troca de interesses em comum.

Neste projeto, a rede social online escolhida para coleta de informações e estudo dos comportamentos dos usuários foi o Twitter. Informações divulgadas pela empresa, referente aos dados de outubro de 2013, apontam que atualmente existem mais de 904 milhões de usuários cadastrados, porém apenas 232 milhões são usuários ativos, e um número em torno de 500 milhões de mensagens são enviadas diariamente na rede[3]. Ainda, 24% do total de usuários do Twitter são usuários Norte-Americano e o Brasil ocupa a quinta colocação com 4,3% deste total, o que corresponde a aproximadamente 10 milhões de usuários.

No Twitter, usuários podem enviar mensagens de até 140 caracteres. São mensagens curtas e objetivas, muitas vezes com link para o conteúdo citado de forma completa. É muito eficiente para a difusão de informações visto que as mensagens, chamadas *tuítes*, são exibidas para todos os seguidores do usuário remetente. A funcionalidade seguir (*follow*) é utilizada por um usuário A quando este deseja ser informado dos tuítes enviados por um usuário B e também pode ser utilizada como uma forma de expressar amizade entre duas pessoas.

Outras funcionalidades importantes do Twitter são: retuíte e menção. Quando um usuário lê um tuíte o qual ele se interessa e deseja que seus seguidores também o vejam, ele pode retuitar a mensagem e fazer com que ela seja espalhada na rede dos seus seguidores. A menção, por sua vez, é utilizada quando um usuário A deseja citar um usuário B em seu tuíte. Essas duas funcionalidades são essenciais para a difusão de informações nesta rede social e serão analisadas neste trabalho.

2.3 Análise de Influência em Redes Sociais Online

(Sun) afirmam que a influência social é a mudança de comportamento de uma pessoa por causa da relação percebida com outras pessoas, organizações e sociedade em geral. Ainda, (Liu) cita duas hipóteses para conceituar influência: 1) Usuários com interesses similares possuem forte influência uns sobre os outros; 2) Usuários os quais as ações frequentemente se correlacionam também possuem forte influência uns sobre os outros.

Conforme (Newman), as redes também têm sido estudadas extensivamente nas ciências sociais. Questões típicas que buscam ser respondidas em redes sociais são relacionadas à centralidade (quais indivíduos são mais ligados a outros ou tem mais influência) e conectividade (como os indivíduos estão ligados uns aos outros através da rede) no âmbito de redes complexas.

3 Metodologia

A presente seção descreve a metodologia utilizada neste trabalho para realização da coleta de dados na rede social online Twitter, a análise de influência através das contagens de retuítes e menções dos usuários envolvidos na amostra coletada e a identificação das características das mensagens e dos usuários envolvidos na amostra a fim de entender melhor como ocorre a propagação de conteúdo nesta rede social.

[3] http://www.mediabistro.com/alltwitter/twitter-ipo-filing_b50130 acessado em 09 de dezembro de 2013.

3.1 Extração de Dados do Twitter

O primeiro passo é a extração de dados do Twitter. A extração será feita através do protocolo HTTP, o qual fará requisições de dados ao servidor do Twitter e terá como resultado um conjunto de dados no formato JSON. Entretanto, o Twitter exige que seja enviado na requisição uma chave de acesso e uma chave secreta. Para obtenção dessas duas chaves, é necessário o registro de um novo aplicativo por seu usuário da rede social[4].

Um algoritmo foi desenvolvido para fazer requisições HTTP e armazenar os dados em arquivos. Este algoritmo, desenvolvido na linguagem Python, utiliza a biblioteca Python-Twitter[5] que é responsável por encapsular os métodos HTTP da API do Twitter. A Streaming API será utilizada para fazer coleta em tempo real de assuntos específicos, dado que este tipo de conexão é ativa com o servidor do Twitter e, utilizando um filtro de palavras-chave, é possível coletar todos os tuítes enviados desde a criação dessa conexão e que contenham essas palavras. Nesse caso, apenas os tuítes enviados a partir da hora em que foi estabelecida a conexão serão coletados.

3.2 Análise de Influência por contagem de Retuítes e Menções

Baseado no trabalho de (Cha), os seguintes dados serão utilizadas: a) quantidade de seguidores; b) quantidade de retuítes; c) quantidade de menções. (Cha) e (Bakshy) afirmam que a quantidade de seguidores representam a audiência de determinado usuário. Isto porque, no Twitter, quando um usuário envia um tuíte, todos os seus seguidores irão receber essa mensagem.

A segunda métrica definida, quantidade de *retuítes*, segundo (Cha), representa o valor do conteúdo de um tuíte. Quando um usuário lê um tuíte e se identifica com este conteúdo, ele tende a retuitá-lo para que os seus seguidores também vejam este mesmo tuíte. Esta funcionalidade é muito poderosa porque é a responsável pela difusão exponencial de conteúdos na rede. Já a quantidade de menções, representa o valor de nome de determinado usuário (Cha), ou seja, o poder de engajamento de determinado usuário perante os outros.

Baseado nesses estudos, a Figura 1 exibe a arquitetura construída neste trabalho para coleta, tratamento e análise dos dados:

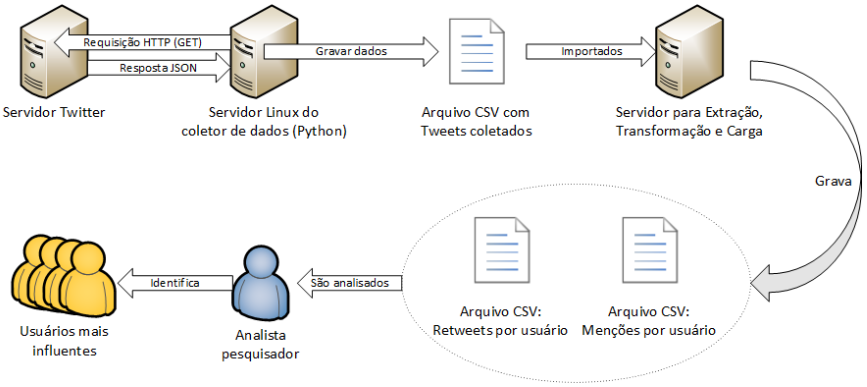


Figura 1- Arquitetura para descoberta dos usuários influentes

[4] https://dev.twitter.com/apps/new

[5] https://github.com/bear/python-twitter

Esta arquitetura identifica as etapas necessárias para o identificação dos usuários mais influentes. Entretanto, ela pode ser dividida em dois momentos: 1) Coleta dos dados; e 2) Tratamento e análise.

Para a coleta dos dados, será utilizada a Streaming API do Twitter, descrita anteriormente. Todos os tuítes que forem coletados deverão ser armazenados em um arquivo de saída. Apesar de todos os campos recebidos do Twitter serem armazenados, os campos utilizados nessa etapa da metodologia serão: identificador único do tuíte, data e hora de criação, mensagem, usuário que criou o tuíte, usuários retuitados (se houver), usuários mencionados (se houver) para cada tuíte coletado. Dessa forma, separando em outros 2 arquivos, foi possível diminuir o tamanho do arquivo em 10 vezes comparado ao tamanho total de cada JSON do tuíte. Esse processo de coleta pode durar horas, dias, meses, dependendo apenas da estratégia definida pelo analista.

O segundo momento é o tratamento e análise dos dados. Uma vez que milhares de tuítes foram coletados e armazenados em arquivo, é necessário a criação de um algoritmo para tratamento desses dados. Esse tratamento deverá ler os dados e organizá-los em rankings de quantidade de retuítes por usuários e quantidade de menções por usuário, através de um algoritmo desenvolvido para este fim. Os rankings de retuítes e menções foram gravados em arquivos diferentes e os dados no formato CSV.

3.3 Identificação das características presentes em conteúdos virais

Esse passo da metodologia tem o objetivo de identificar as características que podem ser utilizadas para compreender o comportamento dos usuários e dos tuítes na amostra coletada neste trabalho. Através de histogramas será possível entender em quais ocasiões as interações entre os usuários ocorrem. Sabendo-se que as interações entre os usuários são realizadas através de tuítes, retuítes, menções, *hashtags*, dentre outras funcionalidades fornecidas pelo Twitter, estudar o momento em que estas mais ocorrem pode fornecer uma visão macro da amostra. Sendo assim, a análise foi realizada baseado na coleta de tuítes original, com todos os metadados fornecidos pelo Twitter. Em seguida, os metadados relevantes foram separados para serem agrupados e sumarizados de acordo com os retuítes de um tuíte origem.

4 Resultados

4.1 Extração de dados do Twitter

O primeiro experimento coletou dados sobre o tema “Fórmula 1” e o segundo tema escolhido foi o “Black Friday”. Ao todo, aproximadamente 3 milhões de tuítes foram coletados durante nove dias, ocorridos entre os dias 22/11/2013 e 01/12/2013. Consequentemente, dado o sucesso da coleta experimental, a coleta de tuítes de outros cinco temas ocorreram entre os dias 03/02/2014 e 13/01/2015. Como pode ser visto na Tabela 1 de forma detalhada, foram coletados ao todo aproximadamente 62.140.000 tuítes.

Tabela 1 - Tabela com os temas das coletas, período e quantidade de tuítes coletados

Tema	Início da coleta	Término da coleta	Total coletado
Fórmula 1	22/11/2013	26/11/2013	204.041
Black Friday	29/11/2013	01/12/2013	2.639.109
Copa do Mundo e FIFA	03/02/2014	13/01/2015	41.070.000
Dilma Rousseff	03/02/2014	13/01/2015	4.940.000

SuperBowl	03/02/2014	13/01/2015	3.940.000
Big Brother Brasil	11/02/2014	13/01/2015	10.350.000
Eleições	10/03/2014	13/01/2015	1.840.000

O próximo passo é a análise de influência dos usuários cujo tuítes foram coletados, baseado em sua audiência, na contagem dos retuítes e na contagem das menções.

4.2 Análise de influência por contagem de retuítes e menções

A primeira análise consistiu em observar os 20 usuários mais retuitados (top 20) da amostra coletada sobre os temas Fórmula 1 e Black Friday. Percebeu-se que os primeiros usuários são responsáveis por grande parte dos conteúdos mais retuitados, como mostra a Figura 2. Este gráfico também evidencia que a audiência do usuário que envia o tuíte não tem relação direta com a propagação deste. É possível verificar usuários com grande audiência e menos retuítes do que outros.

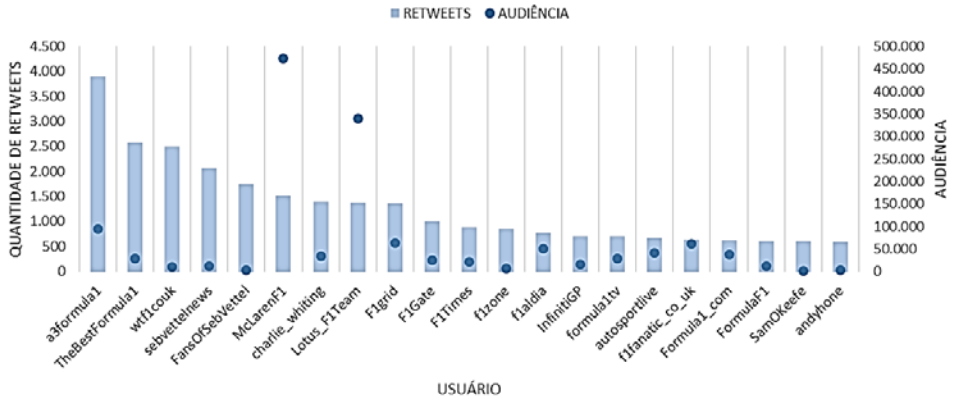


Figura 2 - Ranking dos 20 usuários mais retuitados e sua audiência referente ao tema Fórmula 1.

Um fator curioso é a diferença de retuítes entre o usuário mais retuitado (top 1) e o 20º usuário, o usuário *a3formula1* possui aproximadamente 4.000 retuítes e o usuário *andyhone* possui pouco menos de 1.000 retuítes. E essa quantidade tende a ser linear no gráfico para o restante dos usuários. Ainda, o mesmo comportamento ocorre para o tema Black Friday, onde também foi possível perceber que os primeiros usuários são responsáveis por grande parte dos conteúdos mais retuitados e a audiência do usuário não tem relação direta com a quantidade de retuítes.

Por outra perspectiva, foi analisado a utilização da funcionalidade retuíte e o impacto que o TOP 20 usuários mais retuitados tem perante os outros. Foi possível observar que, de todos os tuítes coletados na amostra da Fórmula 1, 42% são retuítes, e na amostra do Black Friday, 46% são retuítes. Isso indica que a rede social Twitter é significativamente importante para a difusão de conteúdo, visto que muitos usuários fazem questão de enviar para seus seguidores um conteúdo que eles acharam interessante, um conteúdo de valor. O resto da amostragem de cada tema é formado por tuítes originais e não repetidos. Também foi possível observar que a soma da quantidade de retuítes dos 20 usuários mais retuitados representa 31% e 13%, para os temas Fórmula 1 e Black Friday, respectivamente, de todos os retuítes da amostra, que contém 9.653 (Fórmula 1) e 261.089 (Black Friday) usuários.

Dessa forma, foi possível identificar os usuários que são responsáveis pela maior parte da difusão de conteúdo na rede, representada pela métrica quantidade de Retuíte.

A mesma metodologia foi utilizada para analisar as amostras dos dois temas para o ponto de vista das menções. A hipótese de (Cha) com relação às menções – que são mais utilizadas para usuários que causam engajamento, como celebridades – também foi verdadeira neste trabalho. Os três primeiros usuários mais mencionados são, respectivamente: Mark Webber, Fernando Alonso e Felipe Massa. Mark Webber foi o usuário mais mencionado da amostra, apesar de não ter a maior audiência (representada em logaritmo de 10 na Figura 3 indicando, novamente, que a audiência quando analisada de forma isolada, não tem grande significado.

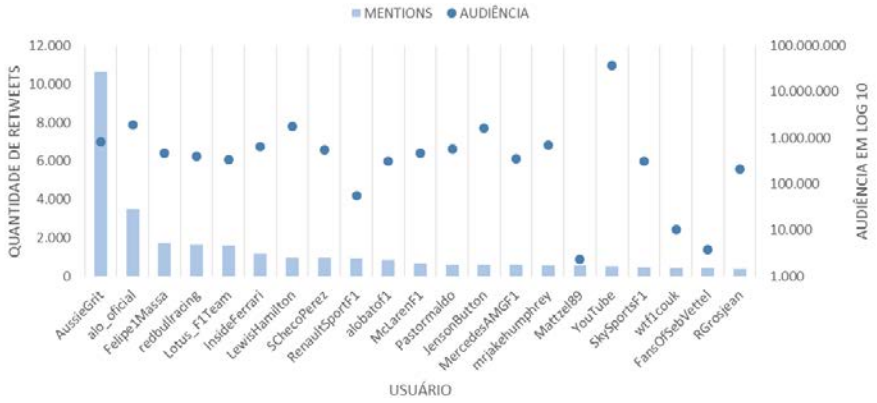


Figura 3 - Ranking dos 20 usuários mais mencionados e sua audiência referente ao tema Fórmula 1

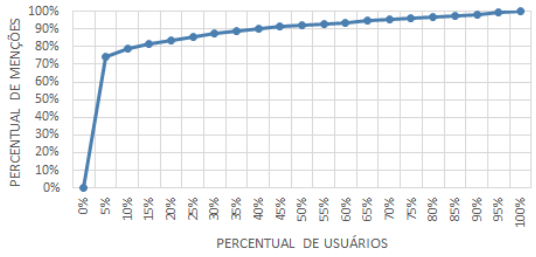
Assim como o gráfico do TOP 20 usuários mais retuitados, este gráfico de menções também tem uma tendência linear após o vigésimo usuário, em que a quantidade de menções vai reduzindo pouco a pouco. Esta métrica não representa a difusão do conteúdo na rede, entretanto, identifica os usuários que causam engajamento sobre determinado assunto e isso pode ser utilizado estrategicamente para este objetivo. Do ponto de vista de marketing, por exemplo, é possível comprovar que é mais valioso patrocinar o piloto Mark Webber do que o Grojean, pois o Mark Webber é o centro das atenções. Porém, para validar esse exemplo, é importante também analisar o conteúdo da mensagem que estão mencionando este piloto para saber se são mensagens positivas ou negativas, o que foge do escopo desta pesquisa.

Ao analisar a amostra das menções dos dois temas do ponto de vista do comportamento da funcionalidade de menções no Twitter, foi possível observar que de todos os tuítes coletados, apenas 25% contém menção, no tema Fórmula 1, e 13% no tema Black Friday. A grande maioria dos tuítes não fazem menção à outro usuário, um comportamento diferente comparado à funcionalidade retuíte que representa quase a metade da amostra. Ainda, com relação à participação dos 20 usuários mais mencionados perante todos os outros usuários mencionados, foi possível observar que 43% das menções realizadas na amostra da Fórmula 1 e 22% na amostra do Black Friday foram para estes usuários. Os outros 57% (Fórmula 1) e 78% (Black Friday) das menções que ocorreram estão divididas para 9.600 e 57.614 usuários, entre Fórmula 1 e Black Friday, respectivamente.

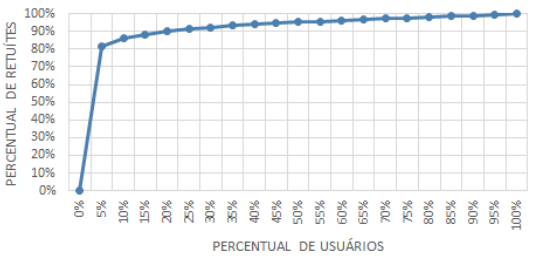
Com esses dados, foi possível identificar os usuários mais importantes do processo de difusão de conteúdo desses temas. Foi possível verificar que poucos usuários são responsáveis por grande espalhamento das informações e que a audiência dos usuários não é uma métrica que deve ser analisada de forma isolada, pois não revela

muita coisa. Também foi possível identificar quais usuários causam maior engajamento na rede, através do ranking de menções por usuário.

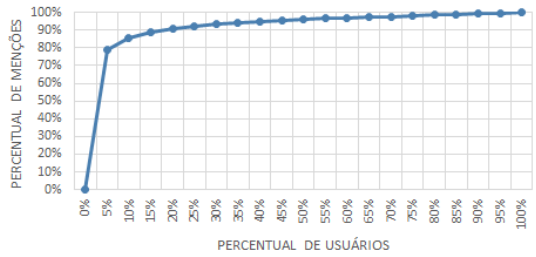
A fim de comparar a relação dos retuítes e menções entre os outros temas (Copa do Mundo e FIFA, Dilma Rouseff, SuperBowl, Big Brother Brasil e Eleições), foram gerados gráficos de distribuição empírica acumulada (DEA). Esse gráfico revelou a relação custo-benefício no que se trata lidar com poucos usuários e gerar grande espalhamento de conteúdo na rede, resultados semelhantes aos demonstrados com os temas Fórmula 1 e Black Friday. Contudo, do ponto de vista de espalhamento de conteúdo por tema, não há novidade. Alguns temas são mais concentrados em poucos usuários do que outros. As figuras Figura 4a até Figura 4j exibem a Distribuição Empírica Acumulada (DEA) de como ocorre a concentração de retuítes e menções sobre cada um dos temas.



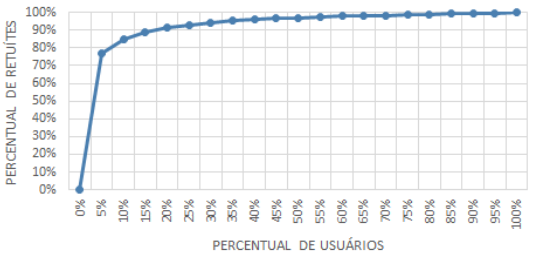
a) DEA das menções (Copa do Mundo e FIFA)



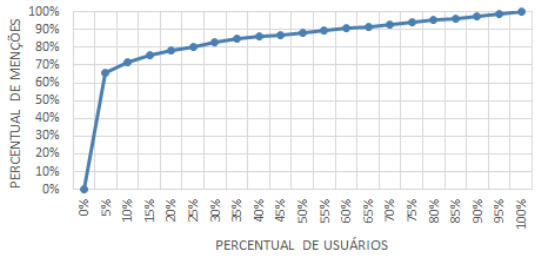
b) DEA dos retuítes (Copa do Mundo e FIFA)



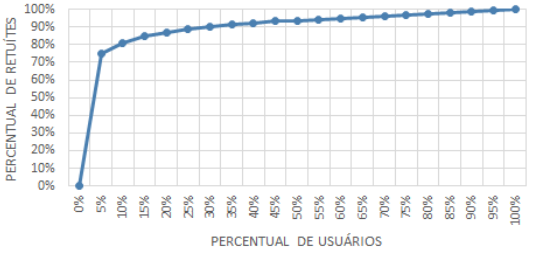
c) DEA das menções (Dilma Rousseff)



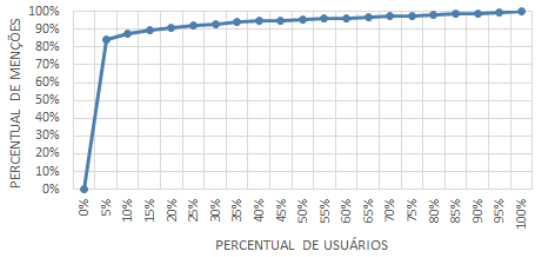
d) DEA dos retuítes (Dilma Rousseff)



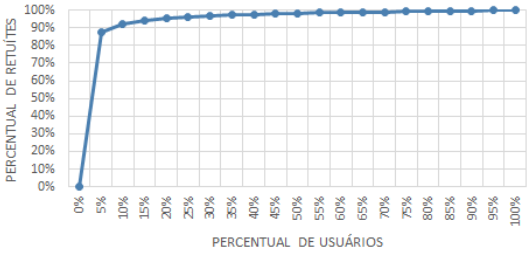
e) DEA das menções (Superbowl)



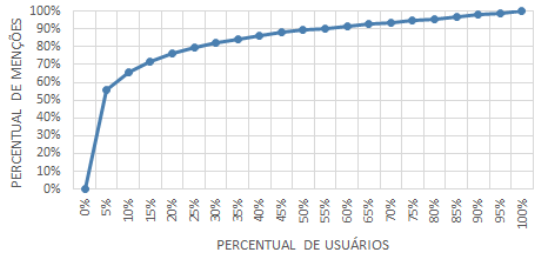
f) DEA dos retuítes (Superbowl)



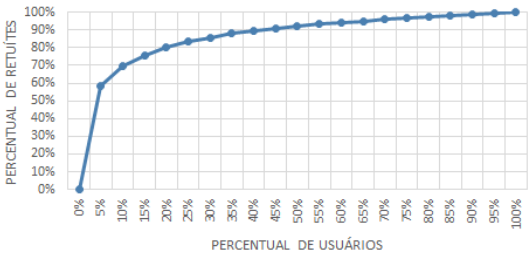
g) DEA das menções (Big Brother Brasil)



h) DEA dos retuítes (Big Brother Brasil)



i) DEA das menções (Eleições)



j) DEA dos retuítes (Eleições)

Figura 4 - Comparativo dos retuítes e menções por Distribuições Empíricas Acumuladas entre os temas

Como se pode observar acima, a ação de retuítes está mais concentrada em menos pessoas do que se forem verificadas as menções. Uma hipótese é a de que a quantidade de usuários que geram conteúdo de relevância dentro da rede social Twitter, consequentemente são mais retuitados, é muito limitada. Primeiro, é preciso ser um pensador; segundo, é necessário estar sempre atualizado; terceiro, é necessário estar gerando conteúdo com alta frequência no Twitter. Enquanto que as menções são distribuídas às celebridades e pessoas famosas que estão relacionadas a cada tema, uma quantidade bem maior visto que essa fama e respeito são obtidos através de outros esforços e não da interatividade dentro do Twitter. Normalmente estes esforços estão relacionados às profissões dos usuários. Também foi possível observar que o tema Big Brother Brasil possui maior concentração de menções do que os outros temas, onde 10% dos usuários mencionados já representam, aproximadamente, 90% do total de menções da amostra, seguido dos temas: Dilma Rousseff, Copa do Mundo e FIFA, SuperBowl e Eleições.

Do ponto de vista de retuítes, os temas mais com difusão de conteúdo mais concentrados em poucos usuários, em ordem, foram: Big Brother Brasil, Copa do Mundo e FIFA, Dilma Rousseff, SuperBowl e Eleições. Como citado, houve uma troca de posições entre o tema Dilma Rousseff e Copa do Mundo e FIFA ao comparar menções e retuítes. Isso significa que, existem mais celebridades relacionadas ao tema Dilma Rousseff do que ao tema Copa do Mundo e FIFA no Twitter. No entanto, a concentração de usuários gerando conteúdos relevantes é maior no tema Copa do Mundo e FIFA.

Com esse estudo de contagem, foi possível conhecer os usuários que possuíram maior destaque na amostra, seja por retuíte ou por menção. Entretanto, não é possível afirmar com certeza que estes são usuários influentes na rede porque é possível que o

verdadeiro usuário influente tenha retuitado um usuário comum, fazendo com que o conteúdo se tornasse viral.

4.3 Identificação das características presentes em conteúdos virais

Os gráficos gerados para análise do comportamento da amostra foram realizados com duas visões. A primeira visão é com valores absolutos, ou seja, independentemente da quantidade de tuítes escritos, foram somados todos os retuítes da amostra e distribuído os percentuais de participação de cada grupo. Já a segunda visão, é demonstrada de forma relativa, ou seja, o maior ponto no gráfico considera a quantidade de retuítes dividido pela quantidade de tuítes existentes em seu grupo. Com essas duas visões é possível entender melhor a amostra e tomar decisões mais embasadas.

Os dados foram extraídos de uma base de dados que contempla somente o tema “Fifa e Copa do Mundo”, no período de março à junho de 2014. Esta amostra possui exatamente 1.623.882 tuítes e a soma de todos os retuítes recebidos equivale a 11.396.006 ao todo.

Após analisar todos os histogramas, percebeu-se que não é possível tomar decisões baseado apenas na informação gerada pelos gráficos, porém é um bom começo para investigar os cenários que uma amostra possui. A junção de várias informações desse tipo é que pode gerar estratégias mais embasadas para entender como ocorre a difusão das informações de um determinado tema.

4.3.1 Quantidade de retuítes por dia da semana

A Figura 5a exibe um gráfico em que foi analisado a quantidade de retuítes por dia da semana. Verificando o eixo de percentual absoluto dos retuítes, é possível perceber que os dias em que ocorreram mais retuítes, comparando toda a amostra, foram a quarta e quinta-feira com mais de 18% de todos os retuítes em cada dia. De acordo com o percentual relativo de retuítes, verifica-se que a quinta-feira e a segunda-feira são os dias em que mais ocorrem retuítes por tuíte. Analisando a segunda-feira, especificamente, pode-se perceber que não e um dia em que ocorrem muitos retuítes ao todo (percentual absoluto), porém, apesar de ocorrer menos retuítes nesse dia, os conteúdos criados na segunda-feira são mais retuitados do que os tuítes criados na quarta-feira. Portanto, podemos afirmar que os tuítes criados na quinta e segunda-feira tendem a ser mais retuitados, no entanto, a maioria dos retuítes ocorrem entre quarta e quinta-feira.

4.3.2 Quantidade de retuítes por horário

Analisando o gráfico da Figura 5b, é possível perceber que a maioria dos retuítes desta amostra ocorrem durante o período noturno. No entanto, os tuítes criados durante o período da tarde são os mais retuitados, de acordo com o eixo de percentual relativo. Nota-se também que, apesar de ocorrer poucos retuítes durante a manhã, os tuítes criados nesse horário são muito retuitados. O horário da madrugada, por sua vez, indica que os tuítes criados neste período tem menos chance de ser retuitados.

4.3.3 Quantidade de retuítes por quantidade de *hashtags* em um tuíte

O gráfico desta seção compara a quantidade de retuítes com a quantidade de *hashtags* utilizada na mensagem. As *hashtags* são muito utilizadas categorizar as mensagens e não tem limite de quantidade em um tuíte. O limite é o do próprio tuíte (255 caracteres).

A figura Figura 5c exibe um gráfico separando os tuítes com nenhuma até cinco *hashtags* ou mais do que seis. Os tuítes escritos sem a utilização de *hashtags* (valor 0) concentra a maior parte de retuítes de toda a amostra. Contudo, o curioso é o pico do percentual relativo ocorre quando um tuíte possui 4 *hashtags*, indicando que os tuítes que possuem 4 *hashtags* em sua mensagem são os mais retuitados de toda a amostra.

4.3.4 Quantidade de retuítes por quantidade de imagens em um tuíte

Ao comparar a distribuição de retuítes por quantidade de imagens inseridas em um tuíte, percebe-se que aproximadamente 60% ocorrem em tuítes que não possuem imagens anexadas, como pode ser visto na Figura 5d. No entanto, analisando a curva do percentual relativo, percebe-se que os tuítes criados com 1 imagem anexada possui mais chances de ser retuitados do que os que não possuem imagens. Os tuítes que possuem 2 imagens anexadas possuem também ótimas chances de serem retuitados, obtendo 1400% de retuítes comparado a quantidade de tuítes escritos desta forma.

4.3.5 Quantidade de retuítes por usuários que são verificados pelo Twitter (usuários oficiais)

O objetivo deste gráfico é comparar a quantidade de retuítes com a classificação de *verified* que o Twitter possui. Essa classificação existe para garantir que uma pessoa ou personagem é o perfil verdadeiro desta mesma pessoal / personagem do mundo real, visto que é comum a criação de perfis falsos para atrair pessoas com diferentes objetivos.

Pode-se afirmar, de acordo com o gráfico da Figura 5e, que os tuítes criados por perfis verificados tendem a ser mais retuitados do que os perfis que não foram verificados. A diferença percebida é muito significante, os perfis que não possuem verificação somam aproximadamente 500% de retuítes, enquanto que os perfis validados somam pouco mais de 3.500% retuítes. Apesar disso, a maioria dos retuítes existentes na amostra são provenientes de tuítes criados por perfis que não possuem validação do Twitter. Isso se explica porque este tipo de perfil representa a grande maioria do Twitter e produzem mais conteúdo.

4.3.6 Quantidade de retuítes por quantidade de seguidores que o usuário que escreveu o tuíte possui

O gráfico representado pela Figura 5f foi gerado com o objetivo de identificar qual é uma boa quantidade de seguidores que o usuário que escreve um tuíte deve ter para alcançar bom espalhamento de conteúdo pela rede social.

Não foi possível chegar em número ideal de seguidores que o usuário deve possuir para ter mais chances de ser retuitado, porém, pode-se perceber que a tendência de ser mais retuitado é de quanto maior o número de seguidores. É interessante distribuir a amostra em números mais equalizados e em grupos menores para descobrir este valor. De acordo com o gráfico, os usuários que possuem entre 30.001 a 50.000 seguidores tendem a ser mais retuitados do que os outros, ignorando o grupo de mais de 50.000 seguidores por ser muito genérico.

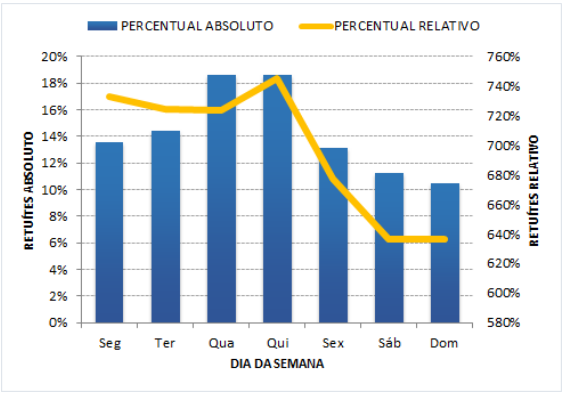
4.3.7 Quantidade de retuítes por quantidade de amigos do usuário que escreveu o tuíte

Este outro gráfico, por sua vez, faz a comparação da distribuição dos retuítes de acordo com a quantidade de amigos que o usuário escritor possui. A relação “amigo” é designada aos usuários que são seguidos pelo escritor do tuíte.

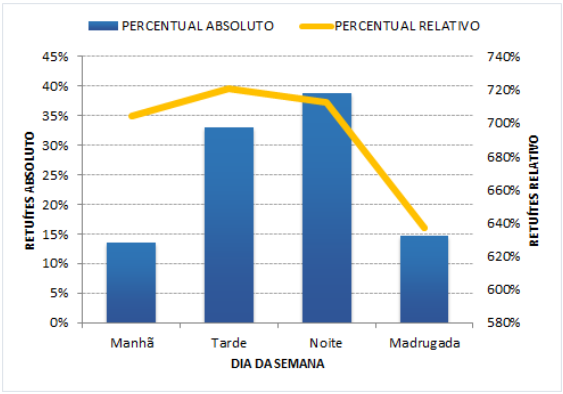
A maioria dos retuítes ocorrem para os usuários que seguem até 1.000 perfis, ou seja, possui no máximo 1.000 amigos na rede social, como pode ser visto na Figura 5g. Também é possível assumir que os usuários que possuem entre 20.000 e 30.000 amigos ou mais que 50.000 amigos tendem a ser mais retuitados do que os outros.

4.3.8 Quantidade de retuítes por quantidade total de tuítes escritos pelo usuário

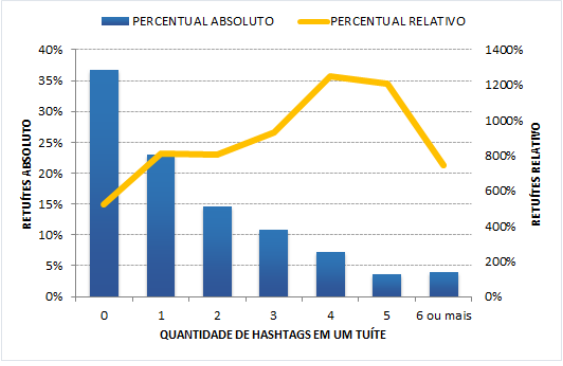
O último histograma analisado compara a distribuição dos retuítes da amostra de acordo com a quantidade de tuítes que o usuário já escreveu em toda sua história dentro da rede social Twitter. Assim, de acordo com a Figura 5h, a maioria dos retuítes estão distribuídos entre os usuários que possuem entre 10.001 e 20.000 tuítes escritos ou mais que 50.000 tuítes. No entanto, de acordo com a curva do percentual relativo, é possível perceber que os tuítes criados por usuários que possuem entre 3.001 a 4.000 tuítes criados foram os mais retuitados da amostra, podendo significar uma tendência.



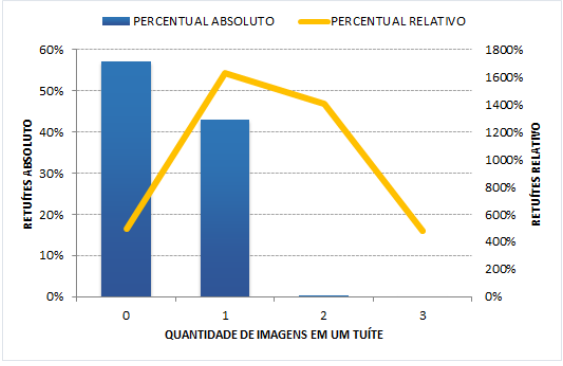
a) Retuítes por dia da semana



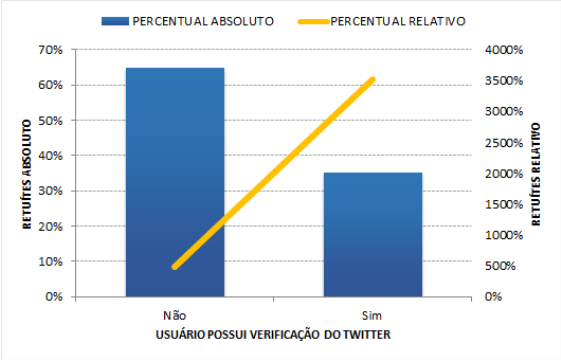
b) Retuítes por horário



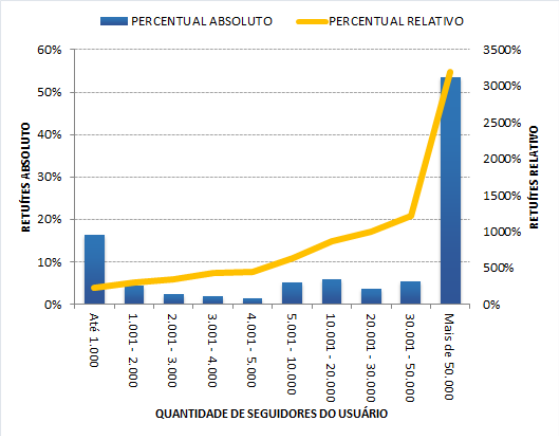
c) Retuítes por *hashtags* em um tuíte



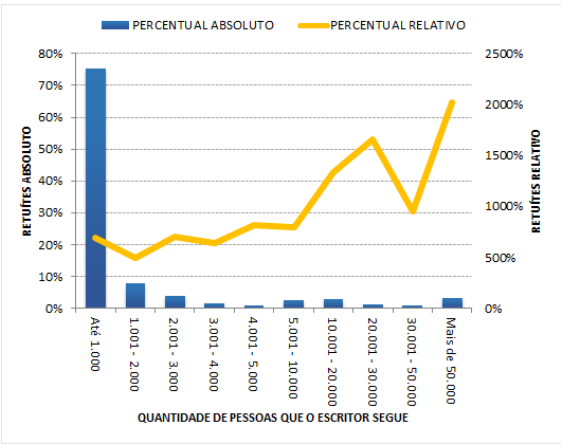
d) Retuítes por imagens em um tuíte



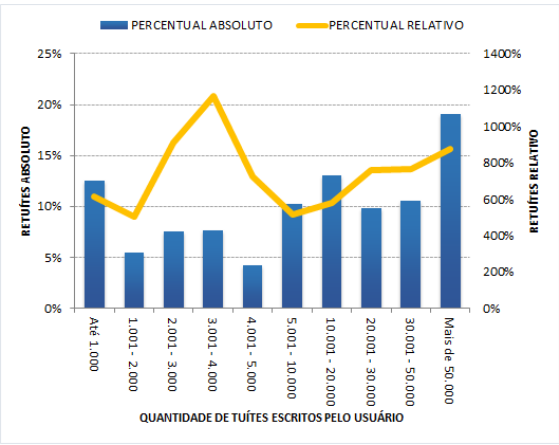
e) Retuítes por usuários que são verificados pelo Twitter



f) Retuítes por seguidores que o usuário que escreveu o tuíte possui



g) Retuítes por amigos do usuário que escreveu o tuíte



h) Retuítes por total de tuítes escritos pelo usuário

Figura 5 - Histogramas do comportamento de usuários e mensagens no Twitter.

5 Discussão

A medida “Seguidores” não tem relação direta com a quantidade de retuítes ou menções dos usuários. Isto ocorre porque independente da audiência do usuário que criou o conteúdo, outros usuários que se interessarem por este conteúdo e retuitarem, esse conteúdo será enviado para seus respectivos seguidores, gerando um espalhamento exponencial pela rede. Assim, foi possível comprovar a hipótese de (Cha) que afirma que os retuítes equivalem ao valor de conteúdo de determinada mensagem e que os usuários mais retuitados são blogueiros e usuários comuns e usuários mais mencionados são pessoas públicas e celebridades. Ainda, as menções identificam quais usuários são responsáveis por causarem engajamento na rede.

Com relação à distribuição das informações, um resultado interessante é que a distribuição de retuítes para diferentes assuntos são similares. Aproximadamente 50% da amostra, de cada um dos temas, trata-se de retuítes e a outra metade trata-se de tuítes originais e únicos, fazendo com que o Twitter apareça com um grande papel no meio

digital: compartilhar e difundir informações. E, de forma generalizada, foi possível constatar o Princípio de Pareto, em que aproximadamente 20% dos usuários mais influentes correspondem a mais de 80% de toda audiência de determinado assunto.

Com a análise de comportamento dos usuários e tuítes foi possível perceber a diferença entre os momentos em que ocorrem mais retuítes *versus* momentos que cada conteúdo é mais retuitado. Sendo assim, é possível ter uma estratégia inicial para difusão de conteúdo, porém deve ser realizado separadamente para cada tema.

6 Conclusão

Foi desenvolvido um coletor de dados do Twitter que funciona 24 horas por dia, 7 dias por semana, coletando todos os tuítes que sejam criados ou retuitados que possuam as palavras-chave especificadas pelo programador. Com este algoritmo, foi possível coletar mais de 62 milhões de tuítes com todos os metadados fornecidos pelo Twitter, composto por dados do próprio tuíte como também dados atuais do usuário que escreveu o tuíte e do usuário que retuitou, possibilitando, assim, que outro pesquisador já possua grande massa de dados para realização de estudos.

Foi possível concluir que a massa de dados obtida é suficiente para uma boa análise de influência em redes sociais online, porém muito trabalho ainda é preciso ser realizado para chegar próximo a um resultado satisfatório, que permita a indução de um conteúdo e obtenha sua viralidade no Twitter. Técnicas de redes complexas e minerações de dados podem ser aplicadas nessa amostragem com o objetivo de entender melhor a amostra a partir de diferentes ângulos. Com os resultados obtidos nessa pesquisa, foi possível compreender melhor como os usuários se comportam na rede social Twitter.

Referências

Bakshy, Eytan and Hofman, Jake M and Mason, Winter A and Watts, Duncan. (2011) "Everyone's an influencer: quantifying influence on twitter." *Proceedings of the fourth ACM international conference on Web search*. ACM. 65-74.

Cha, Meeyoung and Haddadi, Hamed and Benevenuto, Fabricio and Gummadi,. (2010) "Measuring User Influence in Twitter: The Million Follower Fallacy." *ICWSM*: 10-17.

Ellison, Nicole B and others. (2007) "Social network sites: Definition, history, and scholarship." *Journal of Computer-Mediated Communication*: 210-230.

Hill, Shawndra and Provost, Foster and Volinsky, Chris. (2006) "Network-based marketing: Identifying likely adopters via consumer." *Statistical Science*: 256-276.

Jurvetson, Steve. (2000) "What exactly is viral marketing?".

Liu, L. and Tang,J. and Han,J. and Yang,S. (2012) "Learning influence from heterogeneous social networks." *Data Mining and Knowledge Discovery*, 3 ed.: 511-544.

Newman, M. E. J. (2003) "The structure and function of complex networks." *SIAM Review*: 167-256.

Sun, Jimeng and Tang, Jie. (2013) "Models and algorithms for social influence analysis." *Proceedings of the sixth ACM international conference on Web search*. ACM. 775-776.