

# Uma Visão do Mercado Brasileiro de Ações a partir de Dados do Twitter

Hugo S. Santos<sup>1</sup>, Alberto H. F. Laender<sup>1</sup>, Adriano C. M. Pereira<sup>1</sup>

<sup>1</sup>Departamento de Ciência da Computação  
Universidade Federal de Minas Gerais  
31270-010 – Belo Horizonte – MG – Brasil

{hugo, laender, adrianoc}@dcc.ufmg.br

**Resumo.** Neste artigo apresentamos uma visão do mercado de ações brasileiro a partir da caracterização e análise de dados coletados do Twitter no período de julho de 2013 a julho de 2014. Este trabalho tem como principal diferencial ser o primeiro a apontar as ações da BOVESPA que podem ter seu volume financeiro e número de ordens de compra e venda monitorados pelo Twitter. Além disso, nas nossas análises observamos que os eventos e notícias sobre o mercado de ações são capazes de gerar picos de postagens pelos usuários do Twitter e que a frequência de postagens acompanha o início das negociações e mantém-se por cerca de três horas após o fechamento do mercado de ações. Além disso, verificamos que 10% dos usuários são responsáveis por mais de 90% das postagens no Twitter. Finalmente, observamos que o montante financeiro e o volume de ordens de compra e venda são positivamente correlacionados para 66% das ações mencionadas no Twitter, ao passo que as dimensões de oscilação e oscilação máxima não estão correlacionadas.

**Abstract.** In this paper we present a view of the Brazilian Stock Exchange Market based on the characterization and analysis of data collected from Twitter from July 2013 to July 2014. This work has as its main differential being the first one to show which BOVESPA stocks can have its financial volume and number of orders monitored through Twitter. In addition, in our analysis we have observed that events and news about the stock market are capable of generating peaks of publications by Twitter users and that the frequency of posts follows the starting of the exchange trading day and maintains for about three hours after the stock market closing hour. Moreover, we have found that 10% of the users are responsible for more than 90% of the Twitter posts. Finally, we have observed that the financial volume and the number of orders are positively correlated for 66% of the stocks mentioned on Twitter, whereas the oscillation and maximum oscillation dimensions are not correlated.

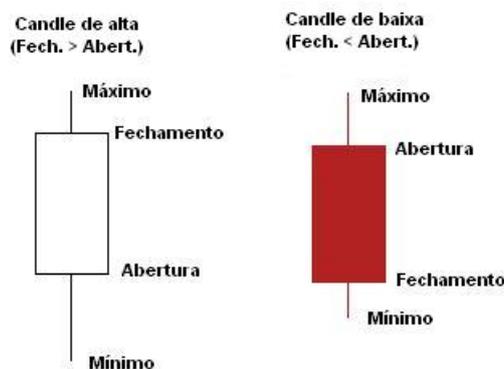
## 1. Introdução

Compreender com precisão comportamentos e tendências futuras do mercado de ações ainda é um grande desafio [Atsalakis and Valavanis 2009, Martinez et al. 2009]. Muitos estudos com o objetivo de mapear o comportamento de ativos já foram elaborados [Fama et al. 1969], mas com o surgimento de novas tecnologias computacionais e, sobretudo, das redes sociais *online*, como o Twitter<sup>1</sup> e o Facebook<sup>2</sup>, novas linhas de pes-

---

<sup>1</sup><http://www.twitter.com>

<sup>2</sup><http://www.facebook.com>



**Figura 1. Representação *candlestick***

quisa se tornaram viáveis [Sprenger et al. 2013] a partir da vasta quantidade de dados disponíveis nessas redes que, de certa forma, refletem comportamentos e opiniões de seus usuários [Maia et al. 2008].

Neste contexto, o microblog Twitter emergiu como o meio de propagação mais popular de acontecimentos recentes [Kwak et al. 2010]. O limite de postagem de 140 caracteres é um direcionador para os usuários realizarem postagens de forma rápida e sucinta. Por exemplo, no dia 24 de junho de 2014, uma nova informação divulgada com exclusividade no Twitter explicitou a capacidade do serviço influenciar as negociações na BOVESPA, a Bolsa de Valores, Mercadorias e Futuros de São Paulo. Nesse caso, um tweet postado foi relevante para reverter na mesma hora a tendência de alta das ações da empresa COPEL<sup>3</sup> durante o pregão de negociação<sup>4</sup>.

Neste artigo apresentamos uma visão do mercado de ações brasileiro a partir da análise e caracterização em larga escala de mais de três milhões de *tweets* e aproximadamente 600 mil *candlesticks* oficiais obtidos a partir da BOVESPA. Esses *candlesticks* descrevem estatisticamente as oscilações dos ativos a cada 15 minutos, como ilustrado na Figura 1.

Podemos listar como as principais contribuições deste artigo:

- Uma caracterização inédita do mercado de ações brasileiro a partir de dados do Twitter, que mapeia um nicho específico de usuários classificados como potenciais investidores e, também, seus padrões de postagens.
- Uma análise de correlação entre a variação das menções aos ativos no Twitter com as respectivas variações dos ativos na bolsa.

O restante deste artigo está organizado da seguinte forma. Na Seção 2 apresentamos uma visão geral da literatura relacionada que aborda estudos de comportamento de usuários no Twitter e no mercado de ações. Na Seção 3 descrevemos a metodologia subjacente à caracterização e análise realizadas e, na Seção 4, os resultados obtidos. Finalmente, na Seção 5 apresentamos as nossas conclusões e alguns direcionamentos para trabalhos futuros.

<sup>3</sup><http://www.copel.com>

<sup>4</sup>Infomoney: Governador do Paraná veta pelo Twitter reajuste da Copel; ações viram para queda - <http://goo.gl/T4YtEt>

## 2. Trabalhos Relacionados

O trabalho realizado neste artigo está relacionado com uma variedade de pesquisas envolvendo redes sociais *online* [Maia et al. 2008, Mislove et al. 2007], o mercado de ações [Cootner 2000, Zhang and Skiena 2010, Ruiz et al. 2012, Gilbert and Karahalios 2010] e, especificamente, o Twitter [Kwak et al. 2010, Sharma et al. 2012, Cha et al. 2012, Sprenger et al. 2013]. Nosso trabalho pode ser posicionado como um ponto de interseção entre essas três frentes de pesquisa, tendo como principal diferencial ser o primeiro a apontar diretamente as ações da BOVESPA que podem ter seu volume financeiro e número de ordens monitorados pelo Twitter.

O Twitter, particularmente, tem sido uma importante fonte de informação em tempo real para inúmeras pesquisas que têm por objetivo extrair padrões de comportamento, influência e tendências de mercado. Sharma et al. [Sharma et al. 2012] desenvolveram um serviço de inferência e caracterização de atributos de usuários do Twitter através de dados coletados a partir de suas listas de interesse (um usuário pode criar uma lista em que outros usuários de mesmo interesse estarão agrupados). Cheong et al. [Cheong and Lee 2009], por sua vez, abordam o Twitter como uma plataforma com potencial para prover inteligência coletiva através da obtenção de opiniões, ideias e fatos úteis para a tomada de decisão.

Bollen et al [Bollen et al. 2011] apresentam um dos trabalhos de base computacional de maior repercussão no mercado financeiro norte americano. Em particular, eles investigaram se as medidas de estado de humor coletivo derivadas da análise em larga escala de dados do Twitter estariam correlacionadas com as variações do valor do índice Dow Jones (DJIA)<sup>5</sup> ao longo do tempo. Como resultado, eles encontraram uma precisão de 87,6% na previsão das variações de fechamento diário do DJIA. Nosso trabalho difere fundamentalmente da investigação realizada por eles por não buscarmos padrões de previsibilidade ou causalidade entre o Twitter e o índice BOVESPA.

Sprenger et al. [Sprenger 2011] basearam-se no Twitter como uma plataforma para a troca de informações sobre o mercado de ações. Eles usaram métodos de linguística computacional para analisar cerca de 250.000 *tweets* cujas mensagens eram relacionadas ao mercado de ações. A partir disso, eles encontraram associações entre o volume de *tweets* e o volume de negociações, e entre o sentimento dos *tweets* e o retorno das ações, além de perceber que os usuários que forneciam conselhos de investimento acima da média eram mais seguidos e tinham um número maior de *retweets*.

Por fim, o principal diferencial deste trabalho é uma visão exclusiva do mercado de ações brasileiro que aponta pela primeira vez as ações da BOVESPA que podem ser diretamente monitoradas pelo Twitter. Essa visão é caracterizada por análises estáticas, temporais e de correlação entre as informações postadas no Twitter e as variações dos ativos na BOVESPA registradas em formato de *candlesticks*.

## 3. Metodologia

Para elaborar uma visão do mercado de ações a partir de dados do Twitter, coletamos por meio da Twitter REST API<sup>6</sup> *tweets* postados por usuários entre julho de 2013 e julho

---

<sup>5</sup>Dow Jones Industrial Average

<sup>6</sup><https://dev.twitter.com/docs/api>

de 2014. A coleta considera termos específicos do contexto de cada empresa (ação), sendo posteriormente processada com técnicas de desambiguação e reconhecimento de entidades. Essa coleta de dados foi feita em parceria com o InWeb (Instituto Nacional de Ciência e Tecnologia para Web), que é um consórcio de grupos de pesquisa que tem como objetivo desenvolver modelos, algoritmos e novas tecnologias para a Web. Os dados referentes à BOVESPA, por sua vez, foram coletados via o *Web Service* oficial<sup>7</sup> em padrão *candlestick* de 15 minutos para cada ação pertencente ao IBOVESPA (o principal índice da BOVESPA cuja cobertura é de aproximadamente 80% do volume financeiro diário), restringindo-se ao mesmo período dos *tweets*.

A partir dos *tweets* coletados foram selecionados como relevantes as postagens cujos textos continham menções ao código dos ativos das empresas do IBOVESPA. Como exemplo, o *tweet* “@clubedopairico PETR4 não foi a R\$ 13,65 em 21/11/2008?” faz parte da base de experimentação, pois menciona explicitamente o código “PETR4” que corresponde aos ativos da PETROBRAS. A Tabela 1 apresenta as estatísticas básicas dos *tweets* coletados e da base experimental utilizada.

Total de <i>tweets</i>	2.973.410
Data Inicial (Primeiro <i>Tweet</i> )	01/07/2013 00:22:03
Data Final (Último <i>Tweet</i> )	31/07/2014 23:10:24
<i>candlestick</i> -15min DataSet	619.125
Primeiro <i>candle</i> (Primeiro <i>Tweet</i> )	01/07/2013 10:00
Último <i>candle</i> (Último <i>Tweet</i> )	31/07/2014 17:45

**Tabela 1. Estatísticas básicas da base de dados experimental**

Particularmente, para a análise dos *tweets* coletados definimos as seguintes questões de pesquisa a serem respondidas:

- Q1: Qual o volume de informação compartilhada no Twitter a respeito da bolsa de valores?
- Há um número considerável de menções aos ativos no Twitter?
  - Existe um nicho de usuários engajados no mercado de ações?
  - A partir dos usuários interessados, qual a estrutura de um *tweet* típico publicado em função da presença de URLs, *retweets* e *replies*?
- Q2: Quais padrões temporais e comportamentais emergem das análises dos dados a partir da combinação dos seus atributos ao longo do tempo?
- O comportamento de postagens diárias reflete os dias de operação da bolsa?
  - As postagens durante os dias refletem os horários de pregão da bolsa?
  - Qual a distribuição da participação dos usuários?
- Q3: Há correlação entre o comportamento dos ativos na BOVESPA e suas respectivas menções em postagens no Twitter?
- Como as seguintes dimensões, volume financeiro, número de ordens de compra e venda, oscilação diária e oscilação máxima diária impactam o Twitter?
  - Existem ações que podem ser monitoradas a partir de dados do Twitter em função das quatro dimensões citadas em (a)?

<sup>7</sup><http://goo.gl/1SWujW>

## 4. Resultados

Para responder às questões de pesquisa definidas, realizamos três tipos de análise. Para a primeira questão (Q1), a análise estática, discutida na Seção 4.1, aborda exclusivamente os atributos dos *tweets* coletados, sem considerar aspectos temporais. Para a segunda questão (Q2), a análise temporal, apresentada na Seção 4.2, complementa a análise estática com padrões comportamentais provenientes do estudo da variação e combinação dos atributos ao longo dos 13 meses considerados. Por fim, a terceira questão (Q3), a análise de correlação, apresentada na Subseção 4.3, investiga como o comportamento dos usuários do Twitter se correlaciona com o movimento da BOVESPA.

### 4.1. Análise Estática

Nesta subseção, abordamos a consulta Q1, e os seus desdobramentos, por meio da análise estática dos atributos dos *tweets* pertencentes à base experimental que mencionam explicitamente os códigos dos ativos. Como pode ser observado na Tabela 2, determinamos que apenas 3,1% dos *tweets* podem ser considerados especializados no mercado de ações, uma vez que mencionam os ativos explicitamente. Além disso, verificamos que, de fato, há um nicho específico de 1% dos usuários do Twitter que podem ser considerados investidores passivos ou ativos.

# <i>tweets</i> que mencionam ativos	93.179 (3,1% de 2.9M)
#usuários	4.816 (0,5% de 1.194.771)
# <i>retweets</i>	4.467 (5%)
# <i>replies</i>	2.657 (2,9%)
# <i>tweets</i> com URL	40.293 (80%)

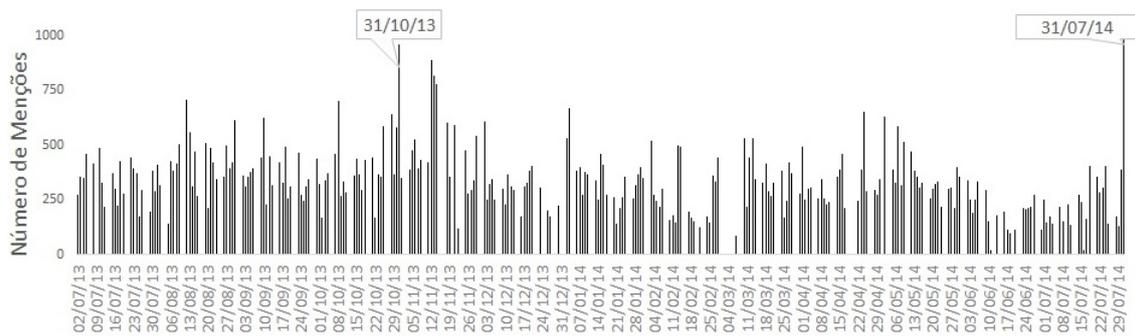
**Tabela 2. Análise estática dos *tweets*: número restrito de *tweets* e usuários especializados, quantidade de *retweets* e *replies*, e percentual de *links* nos *tweets*.**

Considerando o número de *retweets* e *replies*, conforme proposto em [Benevenuto et al. 2010], os nossos resultados sugerem um padrão comportamental caracterizado por postagens de usuários que estão interessados principalmente na divulgação de novas informações (notícias). Tendo em vista que de acordo com a Hipótese do Mercado Eficiente [Fama 1965] os preços das ações são mais influenciados por notícias do que por preços presentes e passados, percebemos que há nesse comportamento um padrão de utilização do Twitter como canal de informações que, de fato, pode influenciar o mercado financeiro desde que as notícias postadas sejam consideradas como relevantes pelos investidores.

### 4.2. Análise Temporal

Na seção anterior, realizamos a análise estática dos *tweets* com base exclusivamente em seus atributos. Nesta seção buscamos responder a pergunta Q2 por meio da análise dos dados da base experimental ao longo do período de 13 meses. Especificamente, analisamos como a quantidade de *tweets* pode variar em função dos dias úteis de negociação da BOVESPA, o número de postagens por hora do dia e, por fim, o percentual de concentração de participação dos usuários dos longo dos meses.

A frequência de postagens dos usuários pode variar em função de suas rotinas diárias, dos dias de operação da bolsa e dos dias da semana, como mostrado na Figura 2



**Figura 2. Número de *tweets* por dia ao longo dos 13 meses**

em que representamos visualmente o número de postagens diárias (eixo y) em função dos dias da nossa análise (eixo x - por restrição de espaço esse eixo mostra intervalos de sete dias). A partir da Figura 2, percebemos como o padrão de postagens no Twitter segue os dias de operação do BOVESPA com um número de postagens constante ao longo dos dias úteis e com valores quase desprezíveis para finais de semana e feriados.

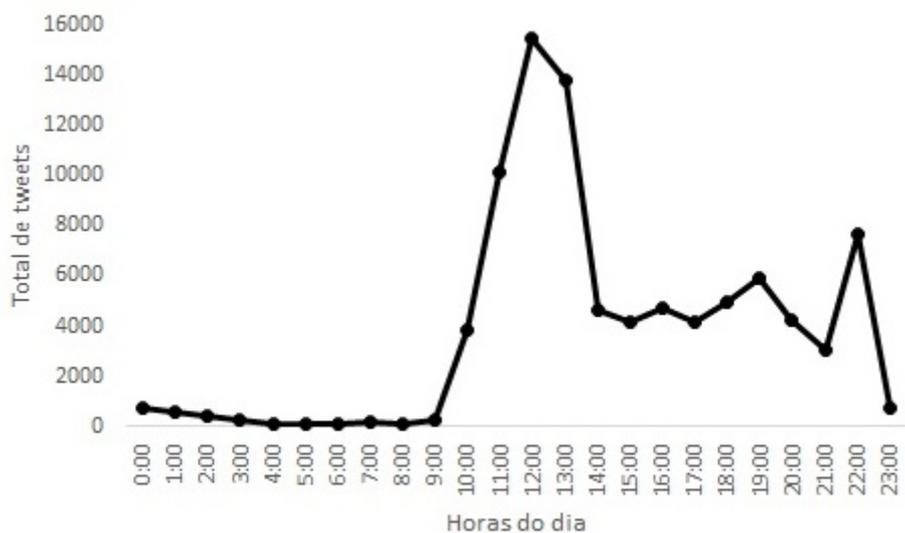
A partir da Figura 2 identificamos para os picos a influência de eventos significativos do mercado financeiro com o número de postagens no Twitter. Por exemplo, o pico em 31 de julho de 2014 deve-se à série de resultados operacionais positivos publicados por empresas brasileiras no segundo trimestre. Resultados como da VALE (VALE5) que registrou 280% de lucro, EMBRAER (EMBR3) que reverteu seu prejuízo em lucro significativo e, também, do aumento de 16% no lucro da AMBEV devido à Copa do Mundo realizada no Brasil. Além disso, o pico das postagens no dia 31 de outubro de 2013 está relacionado à repercussão na mídia da retirada das ações da empresa OGX do índice BOVESPA<sup>8</sup>.

Com o objetivo de melhor compreender os padrões das postagens dos usuários, analisamos a distribuição dos *tweets* em uma menor granularidade de tempo. Assim, a Figura 3 mostra a distribuição dos *tweets* por hora do dia ao longo do nosso período de análise. Sabe-se que a bolsa inicia suas negociações às 10 horas e finaliza às 19 horas. Assim, a partir da Figura 3 identificamos que o período de maior engajamento por parte dos usuários está entre às 11 e 13 horas, além disso as postagens se mantêm após o encerramento do pregão devido a fatores como bolsas externas, noticiários relacionados e, sobretudo, à divulgação de informações estratégicas das empresas que ocorre após o fechamento do mercado, de acordo com regulamentação da CVM<sup>9</sup>.

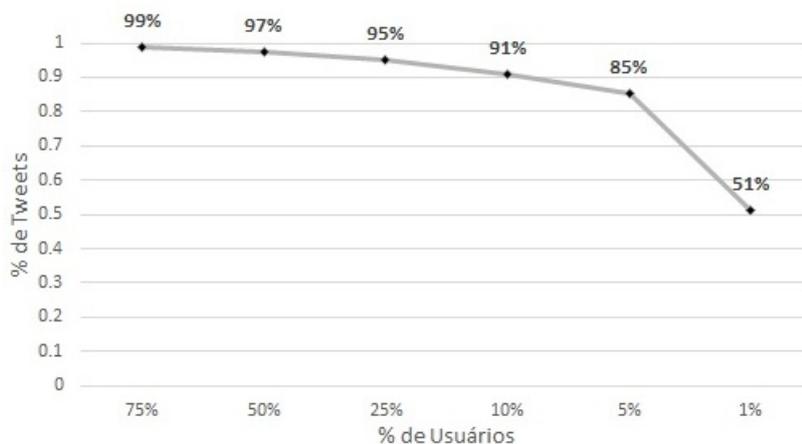
Nesta etapa, concentramos pontualmente nossa análise para evidenciar a distribuição de participação dos usuários. Para isso, calculamos a distribuição das postagens (Figura 4) considerando todo o período de análise. A partir da distribuição acumulada da Figura 4 é possível perceber a presença do padrão de cauda longa típico dos sistemas Web. Ou seja, nesse caso a geração de conteúdo por parte de 10% dos usuários responde por mais de 90% das postagens indicando um grupo de usuários com maior engajamento na geração de conteúdo e que, sobretudo, pode atuar como influenciadores na

<sup>8</sup>Valor Econômico - <http://goo.gl/6UVav9>

<sup>9</sup>Comissão de Valores Monetários - <http://www.cvm.gov.br>



**Figura 3. Postagens acumuladas por hora do dia ao longo dos 13 meses: o valor crescente de postagens acompanha a abertura do pregão de negociação, mas o período de maior interesse dos usuários é entre 11:00 e 13:00 horas, sendo que apesar de o pregão se encerrar às 18:00 horas, as postagens mantêm-se até as 22:00 horas.**



**Figura 4. Distribuição da participação dos usuários: 1% deles publicam metade das informações da rede social.**

disseminação de informação relevante dentro da rede.

### 4.3. Análise de Correlação

Na duas subsecções anteriores analisamos exclusivamente dados do Twitter. Nesta subsecção, motivados pela questão Q3, investigamos o grau de correlação entre o comportamento diário dos ativos na BOVESPA e a frequência de suas menções no Twitter através do coeficiente de correlação de Spearman [Zar 1972] e da escala de Cohen [Cohen et al. 2013].

Estatisticamente, o coeficiente de correlação de Spearman entre duas variáveis X e Y possui escala de -1 a 1. Sendo que -1 e 1 são valores extremos em que as duas

variáveis estão perfeitamente correlacionadas. Ou seja, -1 tem-se que X e Y apresenta uma correlação negativa, enquanto 1, há a correlação positiva. A escala de Cohen está expressa na Tabela 3. Em seguida, definimos a escala de correlação ilustrada na Tabela 4 com o objetivo de mapearmos correlações positivas, negativas ou inexistentes, sem, necessariamente, considerarmos a magnitude da correlação.

Coeficiente de Correlação	Descrição
0.0-0.1	trivial, muito pequeno, insubstancial, praticamente zero
0.1-0.3	pequeno, baixo, menor
0.3-0.5	moderado, médio
0.5-0.7	grande, alto, maior
0.7-0.9	muito grande, muito alto, enorme
0.9-1	próximos, praticamente ou quase: perfeito, distinto, infinito

**Tabela 3. Escala de Cohen**

Coeficiente de Spearman	Descrição
$-1 \leq \rho < -0.1$	Negativa
$-0.1 \leq \rho \leq 0.1$	Sem Correlação
$0.1 < \rho \leq 1$	Positiva

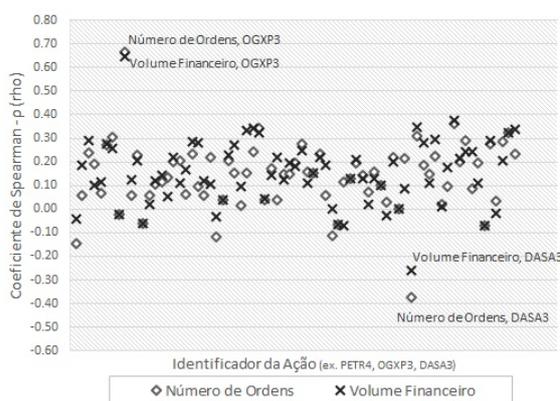
**Tabela 4. Nossa escala Positiva-Negativa baseado no Cohen**

Uma vez definida a escala, construímos uma série para cada uma das quatro dimensões dos ativos para compararmos com o número de suas respectivas menções no Twitter: número de ordens (ou #ordens), volume financeiro, oscilação e oscilação máxima, sendo estas duas últimas dimensões expressas em módulo percentual, ou seja, em dia de desvalorização do ativo a variação em módulo é considerada, já que não é possível ter valores negativos de postagens de *tweets*.

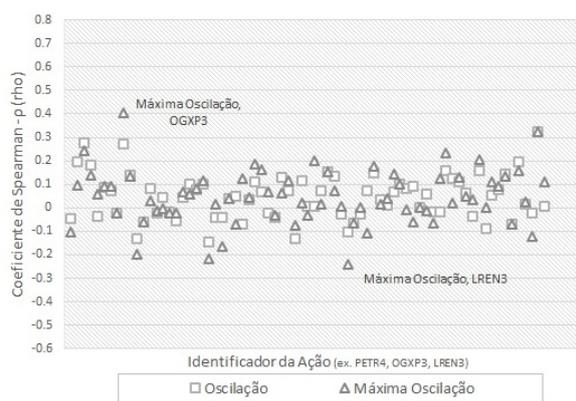
A partir das séries calculamos o coeficiente de Spearman para cada par de séries Twitter-Dimensão\_Y. O resultado está ilustrado nas Figuras 5 e 6. Nelas é possível perceber que determinados ativos, como *OGXP3* e *DASA3*, possuem correlações expressivas e inversas entre o número de menções do ativo no Twitter e as dimensões volume financeiro e número de ordens de compra e venda, respectivamente, sendo que esses dois ativos assumiram valores extremos de  $\rho=0.66$  para *OGXP3* e  $\rho=-0.36$  para *DASA3*.

Por outro lado a Figura 6 ilustra valores extremos de  $\rho$  entre -0.2 e 0.4 aproximadamente para as dimensões oscilação e oscilação máxima, ou seja, uma menor correlação. Ainda a partir da Figura 5 e 6, foi verificado visualmente que a maior parte dos ativos apresentava valor de  $\rho$  fora da escala “Sem Correlação” (Tabela 4) e, sobretudo, que esses valores divergiam em função da dimensão e do ativo.

Com o objetivo de aprofundar a análise dessa observação, identificamos, para cada dimensão, qual o percentual total de correlação determinado por ela sobre todas as



**Figura 5. Coeficiente de Spearman calculado para cada código de ação e as dimensões número de ordens e volume financeiro.**



**Figura 6. Coeficiente de Spearman calculado para cada código de ação e as dimensões oscilação diária e oscilação máxima diária.**

menções aos ativos no Twitter dentro da nossa escala de comparação. Para isso, primeiro consideramos 100% dos dias analisados e, em seguida, selecionamos apenas ativos que possuíssem menções pelo menos a cada dois dias. Os resultados estão expressos nas Tabelas 5 e 6.

	Correlação		
	Negativo	Sem Correlação	Positivo
Núm. de Ordens (#ordens)	5%	29%	66%
Osci. (%)	5%	66%	29%
Osci. Max. (%)	10%	58%	33%
Vol. Financeiro (\$)	1%	25%	74%

**Tabela 5. Dimensão 100% - Percentual de correlação da dimensão sobre as menções aos ativos no Twitter: o valor de 74% indica que o volume financeiro apresentou correlação positiva com 74% dos ativos (todos os dias do período de estudo).**

Os resultados apresentados nas Tabelas 4 e 5 evidenciam que há uma correlação positiva entre o número de ordens, o volume financeiro e o número de menções do ativo no Twitter. Ou seja, à medida que o volume financeiro negociado aumenta espera-se que o número de publicações de *tweets* aumente e vice versa. Por outro lado, ambas as tabelas ilustram que não há uma correlação bidirecional entre a oscilação diária e oscilação máxima diária com o número de *tweets*.

Os resultados encontrados até o momento sugerem que os usuários do Twitter podem ter no padrão de postagens dos seus usuários um indicador do comportamento do volume financeiro e do número de ordens dos ativos na BOVESPA. Vale ressaltar que essas duas dimensões são importantes para o mercado financeiro pois são utilizadas por investidores como métrica de liquidez dos ativos.

A Tabela 7 mostra as ações que, ao mesmo tempo, apresentaram correlação positiva para as dimensões de volume financeiro e número de ordens. Essa tabela contém 22 ações com correlação positiva que identificamos como relevantes para o investidor moni-

	Correlação		
	Negativo	Sem Correlação	Positivo
Núm. de Ordens (#ordens)	3%	26%	71%
Osci. (%)	3%	58%	39%
Max Osci. (%)	6%	48%	45%
Volume Financeiro (\$)	0%	13%	87%

**Tabela 6. Dimensão 50% - Percentual de correlação da dimensão sobre as menções aos ativos no Twitter: o valor de 87% indica que o volume financeiro apresentou correlação positiva com 87% dos ativos (somente ativos que apresentaram, pelo menos, uma menção no Twitter a cada dois dias).**

Positivo	EMBR3, OGXP3, BBDC4, VIVT4, BVMF3, BBDC3, LLXL3, CRUZ3, LIGT3, BTOW3, GFSA3, USIM5, ITUB4, SBSP3, OIBR4, ALLL3, BBAS3, VALE5, PETR3, PETR4, CSNA3, PDGR3
Sem Correlação	MMXM3, BISA3, MRVE3, NATU3, BRML3

**Tabela 7. O conjunto de ações que, ao mesmo tempo, apresentaram correlação positiva para as duas dimensões: volume financeiro e número de ordens.**

torar via Twitter. Por exemplo, ações de empresas como a Petrobras (PETR4), Usiminas (USIM5) e Vale (VALE5), além de bancos como Bradesco (BBDC4), Banco do Brasil (BBAS3) and Itau Unibanco (ITUB4), apresentaram um padrão de postagens correlacionado com seus comportamentos na BOVESPA. Por outro lado, identificamos que ações como MMXM3, BISA3, MRVE3, NATU3, BRML3, de forma geral, não apresentaram padrões de postagens no Twitter que seja correlacionado com as quatro dimensões que estudamos.

Por fim, a partir dos resultados obtidos nessa seção, vale ressaltar que nossos resultados demonstram que ações mais populares, conhecidas no mercado financeiro como *bluechips*, são mais mencionadas no Twitter e, com isso, tendem a apresentar correlação com seu comportamento na BOVESPA.

## 5. Conclusões

Neste artigo apresentamos uma visão do mercado brasileiro de ações a partir de dados coletados do Twitter entre os meses de julho de 2013 e julho de 2014. Essa visão foi caracterizada inicialmente por padrões estáticos provenientes dos atributos das postagens, por padrões temporais que emergiram da análise da base experimental ao longo do tempo e, por fim, pela análise de correlação entre o número de postagens no Twitter e o número de ordens, a oscilação diária, a oscilação máxima e o volume financeiro dos respectivos ativos citados nos textos postados na rede social.

A partir da análise estática identificamos que apenas 3,1% dos *tweets* postados durante o período julho de 2013 a julho de 2014 podem ser classificados como especializados no mercado de ações por conter menções explícitas aos ativos do Ibovespa em seu texto. Além disso, encontramos que o percentual de usuários que podem ser classificados como investidores potenciais por demonstrarem conhecimento do mercado de ações bra-

sileiro é de 0,5%, medido a partir da nossa base amostral que contém os usuários cujas *tweets* foram coletados. Ao final dos experimentos também verificamos que o principal interesse desse grupo seletivo de usuários é a publicação e o compartilhamento de notícias desse mesmo mercado.

A análise temporal, por sua vez, evidenciou os padrões de comportamento dos usuários e da frequência de postagens de *tweets*. Verificamos que eventos e notícias no mercado de ações são capazes de gerar picos de postagens no Twitter por parte de seus usuários. Além disso, constatamos que a frequência de publicações no Twitter acompanha o início do pregão de negociação da bolsa, mas a frequência de postagens se mantém por cerca de 3h após o fechamento do pregão devido a, por exemplo, noticiários especializados. Por fim, verificamos que 10% dos usuários concentram 90% das postagens na rede social.

Por fim, na análise de correlação estudamos o grau de correlação entre o comportamento diário dos ativos na BOVESPA e a frequência de menções no Twitter. Inicialmente calculamos o coeficiente de Spearman ( $\rho$ ) para todos os ativos considerando como série o número de menções diárias no Twitter e cada uma das quatro dimensões de variação dos ativos: número de ordens (#ordens), oscilação diária, oscilação máxima e volume financeiro. Nessa etapa identificamos que as dimensões de volume financeiro e número de ordens apresentam correlação positiva com mais de 66% dos ativos, enquanto as dimensões de oscilação diária e máxima oscilação não demonstraram possuir correlação. No final da seção nós identificamos as 22 ações pertencentes ao índice BOVESPA que investidores podem utilizar o Twitter para monitorar seu comportamento na bolsa.

Percebemos uma variedade de aplicações para os resultados obtidos neste trabalho, como a elaboração de indicadores financeiros que se baseiam nas informações do Twitter para oferecer suporte à decisão dos investidores. Além disso, como trabalhos futuros a serem realizados, planejamos utilizar modelagem de grafos e análise de sentimentos para explicitar padrões interessantes sobre os usuários e comparar os sentimentos extraídos dos seus *tweets* com estatísticas de desempenho da bolsa de valores.

## **Agradecimentos**

Esta pesquisa foi apoiada pelo Instituto Nacional de Ciência e Tecnologia para a Web (InWeb - CNPq no. 573871/2008-6) e por projetos específicos financiados por CAPES, CNPq e Fapemig.

## **Referências**

- Atsalakis, G. S. and Valavanis, K. P. (2009). Surveying stock market forecasting techniques—part ii: Soft computing methods. *Expert Systems with Applications*, 36(3):5932–5941.
- Benevenuto, F., Magno, G., Rodrigues, T., and Almeida, V. (2010). Detecting Spammers on Twitter. In *Proc. of the, Electronic messaging, Anti-Abuse and Spam Conference*.
- Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1 – 8.
- Cha, M., Benevenuto, F., Haddadi, H., and Gummadi, P. K. (2012). The World of Connections and Information Flow in Twitter. *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, 42(4):991–998.

- Cheong, M. and Lee, V. (2009). Integrating Web-Based Intelligence Retrieval and Decision-Making from the Twitter Trends Knowledge Base. In *Proceeding of the 2nd ACM Workshop on Social Web Search and Mining*, pages 1–8. ACM.
- Cohen, J., Cohen, P., West, S. G., and Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.
- Cootner, P., editor (2000). *The random character of stock market prices*. Risk classics library. Risk Books, London.
- Fama, E. F. (1965). The Behavior of Stock-Market Prices. *The Journal of Business*, 38(1):34–105.
- Fama, E. F. et al. (1969). The adjustment of stock prices to new information. *International economic review*, 10(1):1 – 21.
- Gilbert, E. and Karahalios, K. (2010). Widespread Worry and the Stock Market. In *ICWSM*, pages 59 – 65.
- Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a Social Network or a News Media? In *Proceedings of the 19th International Conference on World Wide Web*, pages 591–600. ACM.
- Maia, M., Almeida, J., and Almeida, V. (2008). Identifying User Behavior in Online Social Networks. In *Proceedings of the 1st Workshop on Social Network Systems*, pages 1–6. ACM.
- Martinez, L. C., da Hora, D. N., de M Palotti, J. R., Meira, W., and Pappa, G. L. (2009). From an artificial neural network to a stock market day-trading system: A case study on the bm&f bovespa. In *International Joint Conference on Neural Networks*, pages 2006–2013. IEEE.
- Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. (2007). Measurement and Analysis of Online Social Networks. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, pages 29–42. ACM.
- Ruiz, E. J., Hristidis, V., Castillo, C., Gionis, A., and Jaimes, A. (2012). Correlating Financial Time Series with Micro-blogging Activity. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, pages 513–522. ACM.
- Sharma, N. K., Ghosh, S., Benevenuto, F., Ganguly, N., and Gummadi, K. (2012). Inferring Who-is-Who in the Twitter Social Network. *ACM SIGCOMM Computer Communication Review*, 42(4):533 – 538.
- Sprenger, T. O. (2011). Tweettrader.net: Leveraging Crowd Wisdom in a Stock Microblogging Forum. In *In Proc. of the International Conference on Weblogs and Social, ICWSM'11*. The AAAI Press.
- Sprenger, T. O., Tumasjan, A., Sandner, P. G., and Welpe, I. M. (2013). Tweets and trades: the information content of stock microblogs. *European Financial Management*, pages 926 – 957.
- Zar, J. H. (1972). Significance testing of the spearman rank correlation coefficient. *Journal of the American Statistical Association*, 67(339):578–580.
- Zhang, W. and Skiena, S. (2010). Trading strategies to exploit blog and news sentiment. In *Proc. of the Fourth International Conference on Weblogs and Social Media*, pages 375 – 378. The AAAI Press.