



TESE DE DOUTORADO

Uso de técnicas de Computação Social para tomada de decisão
de compra e venda de ações no mercado brasileiro
de bolsa de valores

Deborah Silva Alves

Brasília, Outubro de 2015

UNIVERSIDADE DE BRASÍLIA

FACULDADE DE TECNOLOGIA

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

USO DE TÉCNICAS DE COMPUTAÇÃO SOCIAL PARA TOMADA
DE DECISÃO DE COMPRA E VENDA DE AÇÕES NO MERCADO
BRASILEIRO DE BOLSA DE VALORES

DEBORAH SILVA ALVES

TESE DE DOUTORADO SUBMETIDA AO DEPARTAMENTO DE ENGENHARIA ELÉTRICA DA
FACULDADE DE TECNOLOGIA DA UNIVERSIDADE DE BRASÍLIA, COMO PARTE DOS
REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR.

APROVADA POR:




GEOVANY ARAÚJO BORGES, Dr., ENE/UnB
(ORIENTADOR)



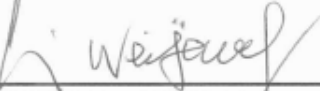
HENRIQUE CÉZAR FERREIRA, Dr., ENE/UnB
(EXAMINADOR INTERNO)



FRANCISCO RAMOS DE MELO, Dr., SI/UEG
(EXAMINADOR EXTERNO)



DANIEL OLIVEIRA CAJUEIRO, Dr., ECO/UnB
(EXAMINADOR EXTERNO)



LI WEIGANG, Dr., CIC/UnB
(EXAMINADOR EXTERNO)

Brasília, 23 de outubro de 2015.

FICHA CATALOGRÁFICA

ALVES, DEBORAH SILVA

Uso de técnicas de computação social para tomada de decisão de compra e venda de ações no mercado brasileiro de bolsa de valores,

[Distrito Federal] 2015.

x, 132p., 297 mm (ENE/FT/UnB, Doutor, Engenharia de Sistemas Eletrônicos e Automação, 2015).
Tese de Doutorado – Universidade de Brasília - Faculdade de Tecnologia.

1. Redes sociais

2. Análise de sentimentos

3. Sistema de apoio a decisão

I. ENE/PGEA/FT/UnB

II. Título (Série)

REFERÊNCIA BIBLIOGRÁFICA

ALVES, D. S. , (2015). Uso de técnicas de computação social para tomada de decisão de compra e venda de ações no mercado brasileiro de bolsa de valores. Tese de Doutorado em Engenharia de Sistemas Eletrônicos e Automação, Publicação FT.PGEA-*n*°102/2015, Departamento de Engenharia Elétrica, Faculdade de Tecnologia, Universidade de Brasília, Brasília, DF, 133p.

CESSÃO DE DIREITOS

AUTOR: Deborah Silva Alves

TÍTULO DA TESE DE DOUTORADO: Uso de técnicas de computação social para tomada de decisão de compra e venda de ações no mercado brasileiro de bolsa de valores.

GRAU: DOUTOR

ANO: 2015

É concedida à Universidade de Brasília permissão para reproduzir cópias desta Tese de Doutorado e para emprestar ou vender tais cópias somente para propósitos acadêmicos e científicos. O autor reserva outros direitos de publicação e nenhuma parte dessa Tese de Doutorado pode ser reproduzida sem autorização por escrito do autor.

Deborah Silva Alves

Instituto de Informática - Universidade Federal de Goiás - Goiânia - GO.

Dedicatória

Ao meu esposo e minha querida filha.

Deborah Silva Alves

Agradecimentos

Ao ser maior que me sustenta todos os dias, autor e consumidor de minha fé.

Ao meu esposo e filha por todo amor, apoio e compreensão.

Aos meus pais, irmão, irmã, sogros cunhados e cunhadas por todo o apoio.

Ao meu orientador pelo apoio, conhecimento, auxílio e dedicação.

Ao especialista da área de compra e venda de ações consultado.

Deborah Silva Alves

RESUMO

O rastreamento do sentimento público para predição de indicadores do mercado financeiro tem ganhado atenção tanto da academia quanto do mundo dos negócios. Entretanto, há várias questões em relação à precisão e significância de modelos que necessitam ser aprimorados. Nesse sentido, este trabalho propõe analisar o relacionamento entre dados obtidos da rede social Twitter em português e do mercado de ações brasileiro através de um sistema de auxílio a tomada de decisão que realiza compra e venda de ações. Para isso, foram coletadas mensagens postadas de agosto de 2013 a abril de 2015 que continham palavras relacionadas às ações de nove empresas brasileiras expressivas no mercado de ações, e dados de volume e preço dessas na Bovespa. Sobre os dados advindos do Twitter, foram aplicadas técnicas para análise de sentimento e tendência para obtenção de indicadores que inicialmente foram relacionados estatisticamente com os da Bovespa e, posteriormente, usados no sistema simulador. Os resultados obtidos demonstraram que o investimento nessa área é promissor apesar dos grandes desafios que esta impõe.

Palavras Chave: Redes sociais, análise de sentimentos, mercado de ações, sistema de apoio a decisão.

ABSTRACT

The tracking of public sentiment indicators to predict the financial market has gained much attention from academia and the business world. However, there are several issues regarding the accuracy and significance of models that need to be improved. Thus, this work aims to analyze the relationship between data in Portuguese language obtained from the social network Twitter and Brazilian stock market through a decision aid system which performs purchase and sale of shares. In order that, messages posted from August 2013 to April 2014 that contained words related to the actions of nine important Brazilian companies in the stock market, and Bovespa data as volume and price were collected. Techniques for sentiment analysis and trend were applied in the data to obtain indicators that were initially associated statistically with the Bovespa and subsequently, they were used in the simulator system. The results showed that investment in this area is promising despite the great challenges it imposes.

Keywords: Social networks, Sentiment analysis, stock marketing, crowd analysis indicators, decision aid system.

SUMÁRIO

1	INTRODUÇÃO	1
1.1	CONTEXTUALIZAÇÃO	1
1.2	DEFINIÇÃO DO PROBLEMA	3
1.3	OBJETIVO DO PROJETO	3
1.4	RESULTADOS OBTIDOS	3
1.5	APRESENTAÇÃO DO MANUSCRITO	4
2	REVISÃO BIBLIOGRÁFICA	5
2.1	REDES SOCIAIS	5
2.2	ANÁLISE DE SENTIMENTOS E OPINIÕES	9
2.2.1	REPERCUSSÃO E OPINIÃO NAS REDES SOCIAIS	12
2.3	ASPECTOS DA COMUNICAÇÃO SOCIAL	14
2.4	ESTIMAÇÃO E REDES SOCIAIS	17
2.4.1	ESTIMAÇÃO NO MERCADO DE AÇÕES	17
2.4.2	ESTIMAÇÃO NA ÁREA DA SAÚDE	21
2.4.3	ESTIMAÇÃO DE POPULARIDADE E REPERCUSÃO EM REDES SOCIAIS	22
2.5	PESQUISAS NO BRASIL	23
3	COLETA DE DADOS E FERRAMENTAS	26
3.1	INTRODUÇÃO	26
3.2	ARQUITETURA DO SISTEMA	26
3.3	COLETA	26
3.3.1	TWITTER	27
3.3.2	DOMÍNIO DE DADOS	27
3.3.3	O COLETOR	28
3.4	PERSISTÊNCIA	28
3.5	PRÉ-PROCESSAMENTO	32
3.6	CLASSIFICAÇÃO - ANÁLISE DE TENDÊNCIA E SENTIMENTO	33
3.6.1	JANELA DE ANÁLISE	35
3.6.2	VOLUME	35
3.6.3	CONTADOR DE PALAVRAS	36
3.6.4	POLARIZAÇÃO DE TWEETS	37

4	ANÁLISE ESTATÍSTICA INICIAL	46
4.1	INTRODUÇÃO	46
4.2	AMOSTRAS	46
4.2.1	DADOS DO TWITTER	47
4.2.2	DADOS DA BOLSA DE VALORES	47
4.3	MODELOS ADOTADOS	49
4.3.1	MODELO DE REGRESSÃO SIMPLES	49
4.3.2	MEDIDAS DE QUALIDADE	51
4.3.3	TESTE DE SIGNIFICÂNCIA	52
4.4	AMBIENTE COMPUTACIONAL	53
4.5	RESULTADOS DA ANÁLISE ESTATÍSTICA INICIAL	54
4.6	COMENTÁRIOS	59
5	TOMADA DE DECISÃO PARA COMPRA E VENDA DE AÇÕES	61
5.1	INTRODUÇÃO	61
5.2	ARQUITETURA DO SIMULADOR	61
5.3	DADOS E INDICADORES	61
5.3.1	DADOS DO TWITTER - JANELA DE DADOS, PRÉ-PROCESSAMENTO E CLASSIFICAÇÃO	62
5.3.2	DADOS DA BOLSA E INDICADORES DE ANÁLISE TÉCNICA	65
5.4	SIMULADOR DE COMPRA E VENDA	69
5.4.1	ESTRATÉGIAS DE ANÁLISE TÉCNICA	71
5.4.2	ESTRATÉGIAS DE ANÁLISE DA MULTIDÃO	72
5.4.3	MÓDULO DE DECISÃO	73
5.5	SAÍDAS DO SIMULADOR	75
6	RESULTADOS	80
6.1	INTRODUÇÃO	80
6.2	DADOS E FERRAMENTAS	80
6.3	RESULTADOS PARA A SIMULAÇÃO DE ANÁLISE TÉCNICA	81
6.4	RESULTADOS PARA A SIMULAÇÃO DE ANÁLISE DA MULTIDÃO	84
6.5	RESULTADOS PARA A SIMULAÇÃO DE ANÁLISE TÉCNICA COM ANÁLISE DA MULTIDÃO	87
6.5.1	TWITTER COM CONVERGÊNCIA/DIVERGÊNCIA DE MÉDIAS MÓVEIS - MACD	87
6.5.2	TWITTER COM CRUZAMENTO DE MÉDIAS MÓVEIS EXPONENCIAIS - MME	91
6.6	COMENTÁRIOS SOBRE OS DADOS	91
6.7	COMENTÁRIOS SOBRE TRANSAÇÕES NO MERCADO DE BOLSA DE VALORES	96
7	CONCLUSÃO	98
7.1	INTRODUÇÃO	98
7.2	CONCLUSÕES	98
7.2.1	COMPARATIVO SIMPLIFICADO DE RENDIMENTOS DA SIMULAÇÃO COM POU- PANÇA E CDI	100

7.3	TRABALHOS FUTUROS	103
7.4	COMENTÁRIOS FINAIS	105
REFERÊNCIAS BIBLIOGRÁFICAS		106

LISTA DE FIGURAS

2.1	Tempo gasto por americanos entre 13 e 64 anos com atividades <i>online</i> [1].	8
2.2	(a) Percentual da população com acesso à internet segundo pesquisa do IBGE [2]. (b) Forma de acesso à internet pelo brasileiro segundo pesquisa PNAD - IBGE [3].	9
2.3	Gráficos de valores de volatilidade ajustados e alvo, o eixo horizontal apresenta os valores observados [4].	18
2.4	Pontuação de bilheteria versus a predita usando dados do Twitter e do Hollywood Stock Exchange obtida por[5].	18
3.1	Arquitetura do Sistema.	27
3.2	Exemplo de um tweet.	28
3.3	Total de Tweets coletados: (a) por empresas (b) por ações das empresas.	31
3.4	Conjuntos de dados e local onde serão utilizados na arquitetura do sistema.	31
3.5	Gráfico de volume de tweets coletados durante o período de 8 meses para cada empresa.	34
3.6	Quantidade de palavras relacionadas à alta e baixa durante os oito meses de captação de tweets sendo (a)PETR4, (b)VALE5, (c)BBAS3 e (d)OGXP3.	38
3.7	Quantidade de palavras com iniciais "compra", "vend" e "alug" por dia para duas janelas de tempo de duas semanas para (a)PETR4, (b)VALE5, (c)BBAS3 e (d)OGXP3.	39
3.8	Amostra de dados coletados do Twitter sendo polarizados manualmente como positivos, negativos e selecionados para avaliação.	44
4.1	Janelas de dados escolhidas para experimento baseadas na quantidade de tweets postados para cada ação.	48
4.2	Dados coletados para a janela de tempo de 9 semanas definido na Seção 4.2 para as ações (a) PETR4, (b) VALE5, (c) BBAS3 e (d) OGXP3.	50
4.3	Retas Ajustadas (a) PETR4 e (b) VALE5, ambas para a janela de amostra de 9 semanas com variáveis Volume de Negociação e Burburinho.	57
5.1	Arquitetura do simulador de compra e venda de ações.	62
5.2	Volume de tweets coletados para PETR4 e VALE5 entre 13 de agosto de 2013 e 04 de maio de 2015.	63
5.3	Volume de tweets para PETR4 com e sem limpeza.	64
5.4	Volume de tweets para VALE5 com e sem limpeza	65
5.5	Linhas de Média Móvel Exponencial de 5, 20 e 200 dias para o preço de fechamento ajustado da ação VALE5 de 13/08/2013 a 04/05/2015.	68

5.6	(a)Linha MACD e Linha Sinal gerada a partir do preço de fechamento, (b) diário da ação PETR4 de 13/08/2103 a 04/05/2015.	70
5.7	(a)Estratégia baseada no Twitter, (b) formação das regras para compra e venda de ações, (c) significado dos símbolos utilizados nas regras e (d) algoritmo da estratégia Twitter.....	74
6.1	Janela de saídas gráficas do simulador para PETR4.	86
6.2	Janela de saídas gráficas do simulador para VALE5.	88
6.3	Janelas de saídas numéricas do sistema para análise da multidão PETR4 com limiar 40 e objetivo de lucro de 10%, (a) saídas por operação e (b) <i>sharpe ratio</i> de toda a simulação.	89
7.1	Cinco maiores lucros acumulados obtidos com as simulações realizadas, em (a) PETR4 e em (b) VALE5.....	99
7.2	Cinco maiores valores de <i>sharpe ratio</i> alcançados com as simulações realizadas, em (a) PETR4 e em (b) VALE5.....	99
7.3	Cinco maiores lucros acumulados com e sem limiar, em (a) PETR4 e em (b) VALE5.	101
7.4	Cálculo do rendimento do valor de uma ação no CDI durante o período simulação adotado, (a) PETR4 e (b) VALE5.	102
7.5	Cálculo do rendimento do valor de uma ação na poupança durante o período simulação adotado, (a) PETR4 e (b) VALE5.....	102

LISTA DE TABELAS

2.1	Amostra de redes sociais populares. As informações sobre quantidade de usuários é um valor aproximado fornecido pelo site de informação da rede social comentada.	6
2.2	Pesquisas recentes realizadas no Brasil	24
3.1	Empresas rastreadas no Twitter.	29
3.2	Palavras e expressões utilizadas para limpeza do banco de mensagens.	33
3.3	Volume total de tweets coletados separados por (a) empresas e (b) por ações.....	36
3.4	Ferramentas para análise de sentimentos em textos.	42
4.1	Medições para PETR4, sendo $B; H; E; S$ indicadores obtidos do Twitter, Cor o Coeficiente de Correlação e p-value o valor do teste t de significância.	55
4.2	Medições para VALE5, sendo $B; H, E$ e S indicadores obtidos do Twitter, Cor o Coeficiente de Correlação e p-value o valor do teste t de significância.	55
4.3	Medições para BBAS3, sendo $B; H$ e E indicadores obtidos do Twitter, Cor o Coeficiente de Correlação e p-value o valor do teste t de significância.	56
4.4	Medições para OGXP3, sendo $B; H$ e E indicadores obtidos do Twitter, Cor o Coeficiente de Correlação e p-value o valor do teste t de significância.	56
5.1	Índice CDI mensal (http://www.cetip.com.br).....	79
6.1	Resultados da simulação de compra e venda da PETR4 por análise técnica.	82
6.2	Resultados da simulação de compra e venda da VALE5 por análise técnica.	83
6.3	Resultado da simulação por análise da multidão - Twitter para PETR4 com e sem limiar de tweets.	90
6.4	Resultado da simulação por análise da multidão - Twitter para VALE5 com e sem limiar de tweets.	90
6.5	Resultado da simulação MACD com Twitter para PETR4.	92
6.6	Resultado da simulação MACD com Twitter para VALE5	92
6.7	Resultado da simulação MME de 5 e 20 períodos com Twitter para PETR4.	93
6.8	Resultado da simulação MME de 5 e 20 períodos para VALE5.....	93
6.9	Resultados da simulação para PETR4 por análise da multidão com limpeza severa de tweets.	95
6.10	Resultado da simulação para PETR4 por Twitter com MACD com limpeza severa de tweets.	95

6.11 Resultado da simulação para PETR4 por Twitter com MME 5 e 20 com limpeza severa de tweets.	96
6.12 Custos com corretora para negociações na bolsa de valores (valores pesquisados em outubro de 2015).....	97

LISTA DE SÍMBOLOS

Símbolos Gregos

σ	Desvio padrão
ρ	Coefficiente de Correlação
ϵ	Desvios do modelo de regressão
β_0, β_1	Parâmetros do modelo de regressão a serem ajustados
α	Nível de significância mínimo adotado

Símbolos Latinos

B	Burburinho, quantidade de tweets coletados
H	Humor - otimismo ou pessimismo
E	Tendência do mercado - alta ou baixa
S	Sentimento positivo ou negativo
R^2	Coefficiente de determinação
T_0	Estatística de Teste
$Cov_{(x,y)}$	Covariância entre x e y
$Cor_{(x,y)}$	Correlação entre x e y
P	Valor ajustado do preço do ativo
v	Porcentagem de volume de negociação da ação na bolsa de valores
p	Porcentagem de evolução do preço da ação na bolsa de valores
n	Quantidade de amostras
s^2	Estimativa da variância
s	Estimativa do desvio padrão
r	Retornos
SR	Índice de <i>Sharpe Ratio</i>
p-value	Nível de significância

Subscritos

t	Tempo em dias
i, j	Contadores
a	Indica que o valor está acumulado
d	Indica que o valor está seguindo a tendência do mercado
cd	Indica que o valor está seguindo contra a tendência do mercado
c	Atuando COMPRADO
v	Atuando VENDIDO
cv	Atuando COMPRADO/VENDIDO
f	Indica um valor referencial

Sobrescritos

$\hat{}$	Valor Ajustado
$\bar{}$	Valor médio

Siglas

RAE	do inglês <i>Relative Absolute Error</i> - Erro relativo absoluto
MME	Média Móvel Exponencial
MACD	do inglês <i>Moving Average Convergence Divergence</i> - Convergência/Divergência de médias móveis -

Chapter 1

Introdução

1.1 Contextualização

O rastreamento de sentimento público para predição de indicadores do mercado financeiro tem ganhado atenção tanto da academia quanto do mundo dos negócios. Dados gerados por comunidades formadas a partir de redes sociais *online* vêm gradualmente obtendo credibilidade como fonte válida para análise do mercado de ações [6]. Vários autores reconhecem essa tendência [7, 6, 8], porém, em seus trabalhos eles identificam algumas questões em relação à precisão e significância de modelos e reportam que há muito para ser feito a fim de alcançar predições realmente efetivas.

Para [7], o fator humano tem significativo impacto no movimento do mercado de ações. Em seu artigo, ele comenta vários trabalhos de predição para mercado de ações, especialmente alguns da década de 60 baseados em '*Random Walk Theory*' (teoria do passeio aleatório) e '*Efficient Market Hypothesis*' - Hipótese do Mercado Eficiente (EMH, do inglês) . A EMH afirma que a valorização do mercado financeiro incorpora quaisquer novas notícias e informações [9, 10], ou seja, os preços do mercado de ações são, em grande parte, impulsionados por novas informações e não por preços do presente e do passado. Como uma nova notícia ou informação é algo imprevisível, de acordo com essa teoria, os preços seguiriam um padrão de passeio randômico e não poderiam ser preditos com precisão superior a 50%.

Entretanto, essa teoria é desafiada por vários pesquisadores que, baseados nas perspectivas da teoria de finança socioeconômica e enfatizando a importância de fatores comportamentais e emocionais, incluindo o humor social [11], a criticam e afirmam que os preços nem sempre seguem um passeio aleatório [9], e podem, até certo ponto, serem preditos. Pesquisas recentes sugerem que apesar de a notícia ser algo imprevisível, muitos indicadores precoces podem ser extraídos da mídia social *online* para estimar mudanças em vários indicadores econômicos e comerciais, e que esse também pode ser o caso do mercado de ações [7].

Em seus trabalhos, autores comentam sobre a economia comportamental, que afirma o fato de as emoções poderem afetar profundamente o comportamento individual e a tomada de decisão [11, 7]. Se a emoção do indivíduo investidor pode afetar a forma como ele reage às novas informações, é provável que o sentimento coletivo dos investidores possa influenciar a dinâmica do mercado

de ações[4]. Como consequência, medir o humor social tornou-se uma questão fundamental na pesquisa de previsão financeira [10]. Pesquisas recentes têm explorado uma variedade de métodos para o cálculo de indicadores de sentimento e estados de humor do público gerados a partir de uma grande quantidade de dados *online* disponível. Computar o sentimento da multidão mostra-se mais efetivo, rápido e com menor custo do que o acesso físico às pessoas através de institutos de pesquisa.

Três classes distintas de fonte de dados *online* são definidas em [10] e têm sido investigadas para predição financeira, produzindo diferentes indicadores:

- Notícias: Fator que molda o sentimento dos investidores. Em [10] é comentado que um alto nível de pessimismo em Wall Street precede baixos retornos no mercado no dia seguinte e que pesquisas mostram que a adição de características textuais de notícias em um sistema de predição de ações pode melhorar a precisão da previsão;
- Dados de busca na *web*: Vários trabalhos têm mostrado o valor destes dados para inferir o interesse do investidor. Dados de busca podem ser relacionados às flutuações do mercado e volumes de negociação, e também podem ser preditivos destes [6, 8];
- Dados de redes sociais *online*: Fonte para apoiar a extração e medição do humor social e do investidor. Estes dados têm sido amplamente estudados como geradores de indicadores que possam ser utilizados em sistemas preditivos do mercado de ações.

Nesse trabalho, será investigado o uso de mensagens de microblog, em língua portuguesa, da rede social *online* Twitter para estimar a dinâmica do mercado de ações brasileiro.

Para [4], o uso de microblog para captar o sentimento dos investidores é interessante por diversos aspectos: a quantidade de pessoas que utilizam esses serviços para comunicar suas ideias a respeito do mercado tem crescido; os dados compartilhados nessas mídias estão disponíveis a baixo custo e são sempre atuais; o tamanho da mensagem também é um diferencial em relação aos textos de blogs comuns, são somente cento e quarenta caracteres o que reduz quantidade de processamento e ruído; e as postagens são realizadas em tempo real e com alta frequência.

Vários pesquisadores analisam o que leva uma pessoa a postar e outra a ler mensagens *online* sobre o mercado de ações. Em[12] há um comentário sobre autores que, baseados em teoria da comunicação, afirmam que as pessoas valorizam as opiniões daquelas com as quais elas conversam, e esse tipo de processo de formação de crença é útil na formação de agentes de influência. De maneira geral, agentes desejam saber o que outros influentes pensam a respeito, desde que o que dizem afete o mercado. De outro modo, [13] teoriza que a conversação entre subconjuntos de participantes do mercado pode ter efeitos no equilíbrio. Um investidor à margem pode decidir por participar de negociações ao passo que percebe pensamentos semelhantes entre si e comentários de outros investidores. Se as conversações na Internet permitem esse tipo de comportamento, então é possível que uma postagem de mensagem seja seguida de uma negociação[12].

Hoje, segundo [10] existe apoio considerável para alegar que indicadores de humor e sentimento resultantes de análises de dados de redes sociais online são medidas de opinião pública realmente

válidas, para prever uma variedade de fenômenos socioeconômicos. No Capítulo 2 serão referenciados vários trabalhos. Com relação ao mercado financeiro, várias pesquisas recentes têm apresentado resultados promissores e incentivadores de estudo na área [14, 10, 4, 8, 5].

A pesquisa a ser relatada nesta tese, aborda o problema do uso de dados obtidos através de redes sociais *online* como fonte de informação para previsão de comportamento do mercado de ações brasileiro. Os resultados a serem apresentados demonstram que empregar esses dados em sistemas de auxílio à tomada de decisão de compra e venda de ações é promissor.

Têm-se um conjunto de dados coletados da rede social Twitter com mais de oito milhões de tweets. Desses dados, obtêm-se o sentimento do povo em relação às ações de oito empresas brasileiras selecionadas para a pesquisa. A partir de dados históricos de evolução de preço e volume de negociação das ações dessas empresas, buscou-se, inicialmente, realizar uma análise estatística do relacionamento dos dados. Posteriormente, foi desenvolvido um sistema simulador que efetivou operações de compra e venda de ações baseado em dados do Twitter e preços obtidos da bolsa de valores brasileira - Bovespa.

Com o refinamento dos dados, optou-se por realizar simulações com duas ações de empresas importantes no cenário de bolsa de valores do mercado brasileiro; por possuírem quantidade suficiente de dados diários para o experimento. Com a avaliação dos resultados obtidos a partir do processamento dos dados oriundos da rede social e do simulador, buscou-se refletir sobre a influência das redes sociais para compra e venda de ações no mercado de bolsa de valores.

1.2 Definição do problema

A pesquisa a ser relatada nesta tese, aborda o problema do uso de dados obtidos através de redes sociais *online* como fonte de informação para previsão de comportamento do mercado de ações brasileiro.

1.3 Objetivo do projeto

A mídia social também pode ser interpretada como uma forma de sabedoria coletiva [5]. Nesse sentido, o objetivo desta pesquisa é investigar e entender se as características dessa sabedoria coletiva, adquirida através das redes sociais *online*, contribuem para melhorar predições para o mercado de ações brasileiro.

1.4 Resultados Obtidos

A pesquisa a ser descrita nesse documento apresenta uma arquitetura de sistema que adota técnicas de computação social para auxílio na tomada de decisão de compra e venda de ações no mercado brasileiro de bolsa de valores. Inicialmente são apresentados os dados utilizados e a forma de obtenção dos mesmos, dados oriundos da rede social Twitter e histórico de preços e volume

de negociação de ações de empresas brasileiras na bolsa de valores de São Paulo - BOVESPA. Posteriormente é realizada uma análise estatística para verificar o relacionamento entre esses dados. Foram efetuados testes e medições para obter os resultados. Dentre eles, resultados interessantes foram obtidos para as ações da Petrobrás e Vale S.A., as quais possuíam maior volume de dados postados no Twitter, especialmente quando o relacionamento foi medido em menor tempo (no caso, em menor quantidade de dias).

Após a análise estatística foi desenvolvido um simulador para compra e venda de ações baseado nos dados do Twitter e em análise técnica. Valores interessantes de lucro acumulado foram obtidos para a compra e venda de ações entre agosto de 2013 e abril de 2015, período total de coleta de dados. Para as ações PETR4 e VALE5 das respectivas empresas Petrobrás e Vale S.A., valores significativos foram alcançados com a simulação, sendo que o maior lucro acumulado obtido foi de 277,86% para a primeira e de 224.28% para a segunda, ambos atingidos apenas com o uso de análise da multidão sem considerar custos com impostos e corretagem.

1.5 Apresentação do manuscrito

Esta tese apresenta em sete capítulos o trabalho realizado. No Capítulo 2, comenta-se a respeito do estado da arte no uso de dados de redes sociais *online* para modelagem e estimação do comportamento humano frente às questões sociais, políticas e econômicas. Uma seção em especial trata das questões de estimação para o mercado de ações e outra apresenta alguns trabalhos realizados no Brasil com ênfase na língua portuguesa e inglesa.

O Capítulo 3 apresenta a arquitetura do sistema adotado para a pesquisa com informações sobre coleta, formatação e transformação dos dados obtidos da rede social Twitter. Além disso, também são descritas as ferramentas utilizadas para as atividades desenvolvidas. No Capítulo 4, são expostas as análises estatísticas realizadas. No Capítulo 5, são detalhados o refinamento realizado no processamento de dados e a arquitetura do simulador para auxílio a tomada de decisão de compra e venda de ações desenvolvido para a pesquisa. Os resultados obtidos com as simulações realizadas estão exibidos no Capítulo 6. Conclusões e comentários finais estão disponíveis no Capítulo 7.

Capítulo 2

Revisão Bibliográfica

2.1 Redes Sociais

Geralmente, entende-se a expressão rede social online como sendo um grupo de pessoas que se interagem através de qualquer mídia social, entretanto, uma definição adotada em trabalhos recentes mostra-se mais apropriada. Segundo [15], rede social é "um serviço web que permite indivíduos (1) construir perfis públicos ou semi-públicos dentro de um sistema, (2) articular uma lista de outros usuários com os quais compartilham conexões e (3) visualizar e percorrer suas listas de conexões e outras listas feitas por outros no sistema." Dentro desse contexto, abre-se um leque de oportunidades para, aproveitando-se dessas milhares de opções de conexões entre pessoas, espalhar-se conhecimento, ideias, sentimentos e opiniões.

O conhecimento sobre o que as pessoas pensam a respeito de certo assunto, pessoa ou produto, saber sua opinião sobre fatos do cotidiano ou obter o sentimento negativo ou positivo em relação a alguma informação, sempre foi o desejo de muitos indivíduos e empresas. Segundo [16], opiniões são fundamentais para quase todas as atividades humanas e são as principais influenciadoras de comportamento das pessoas. A forma como as pessoas percebem a realidade, suas crenças e escolhas que fazem, são consideravelmente condicionadas à forma como outros veem e avaliam o mundo. Por isso, na necessidade de tomar uma decisão, o ser humano e até mesmo organizações, sempre que possível, buscam outras opiniões.

Canais para comunicar opiniões e comentários através de mensagens sobre quaisquer domínios estão cada dia mais comuns e disponibilizados nas mídias sociais. Esses têm se tornado fontes importantes para empresas, organizações governamentais ou não e pessoas para controle de difamação, acompanhamento de lançamentos, contato direto com as pessoas, dentre outros. No caso das empresas, além destes, as informações obtidas são fontes relevantes de conhecimento do negócio, representando o impacto de um revisor influente no poder de compras de outros. Assim, a coleta do pensamento do povo postado nas mídias sociais é de grande valor para o planejamento de novos produtos, divulgação, atendimento ao cliente e manutenção da integridade da marca. À medida que uma maior quantidade de público tem acesso às tecnologias, cresce a importância dos dados obtidos através dessas mídias e o desafio de lidar com elas.

Tabela 2.1: Amostra de redes sociais populares. As informações sobre quantidade de usuários é um valor aproximado fornecido pelo site de informação da rede social comentada.

Rede Social	Ano Criação	Usuários aprox.	Comentários
<i>Myspace</i> [®]	2003	Em 2013: 50 milhões	Com ênfase no universo musical, desde 2006 oferece opção de acesso às versões regionais (brasileira, japonesa, etc.) com conteúdo local. Permite criação de perfil de usuário utilizado para relacionamentos [17].
<i>Flickr</i> [®]	2004	87 milhões	Permite armazenar, organizar e compartilhar fotos e vídeos com amigos. Em 2013, registrou mais de 3,5 milhões imagens postadas diariamente. Fotos e vídeos publicados por usuários podem ser acessados sem a necessidade de conta.
<i>Google +</i> [®]	2011	540 milhões em 2013.	Serviço de rede social do Google, é o segundo maior site de redes sociais do mundo com usuários ativos interagindo socialmente com o Gmail (serviço de e-mail), o botão + (outros serviços oferecidos como, por exemplo, agenda) e comentários do Youtube [18].
<i>TripAdvisor</i> [®]	2000	100 milhões de visitantes	É um <i>website</i> que provê informações, fórum de interação e avaliações sobre viagens. Todo o conteúdo é gerado por seus usuários.
<i>LinkedIn</i> [®]	2003	250 milhões	Propõe conectar profissionais mundialmente. Usuários têm acesso às pessoas, vagas, notícias, atualizações e percepções que o ajudam em sua profissão.
<i>Instagram</i> [®]	2010	150 milhões em 2013.	Permite aos usuários compartilharem fotos e vídeos processadas por filtros disponíveis em uma variedade de outras redes sociais incluindo a própria. Muito utilizado por usuários de <i>smartphones</i> [19].
<i>Snapchat</i> [®]	2011	100 milhões em 2014	Aplicativo com foco em relacionamento. Permite vídeo mensagens que ficam disponíveis durante um intervalo de tempo. Está se tornando popular no Brasil.
<i>Twitter</i> [®]	2006	500 milhões em 2014	Rede social <i>online</i> e serviço de microblog que permite aos usuários o envio de mensagens com no máximo 140 caracteres. Usuários registrados podem publicar e ler mensagens, não cadastrados podem apenas ler [20].
<i>Facebook</i> [®]	2004	1,23 bilhões em 2013 (107,7 milhões no Brasil em 2014)	O usuário cria um perfil pessoal e adiciona outros usuários ou grupos como amigos. A cada publicação de mensagem, vídeo ou imagem desse usuário todos os seus amigos são automaticamente avisados. Aproximadamente, possui 180 petabytes(10^{15} bytes) de dados e 9% de usuários fake [21].
<i>Youtube</i> [®]	2005	1 bilhão em 2014 [22]	Permite às pessoas publicarem, procurarem, assistirem, compartilharem e comentarem vídeos.

Com o advento da computação social ou Web 2.0¹, houve uma explosão de sites de redes sociais *online* que podem ser: de contato, *Facebook*[®], *LinkedIn*[®], *Instagram*[®], *Orkut*[®] (criada em 2004, se tornou muito popular no Brasil mas foi desativada em setembro de 2014)²; de conteúdo, *Twitter*[®], *Flirckr*[®], *YouTube*[®], *MySpace*[®]³, descritos na Tabela 2.1, e outros com enfoque em saúde - rede BiBlioSUS⁴ e em tricô - Ravelry⁵, nos quais milhares de indivíduos passaram a trocar diariamente mensagens, imagens, vídeos e outras possibilidades disponibilizadas por cada ferramenta. Diferentes em estilo e popularidade, dependendo do país, todas proporcionam aos usuários experiências de comunicação em tempo real.

Além desses *sites*, existem também outros que permitem às pessoas compartilharem suas opiniões a respeito de produtos, compras, serviços, viagens, etc., como: *mercadolivre.com*, os de leilões como *ebay*, *Olx* e sobre viagens como o *tripadvisor.com*.

Com relação ao mercado de ações, também não é diferente, existem várias opções de redes sociais para os investidores compartilharem e consultarem estratégias de negociação e comentários a respeito do mercado de ações tais como o *StockTwits*⁶ (rede social americana para investidores e negociadores da bolsa que foi criada em 2009 e em 2013 já possuía mais de 200.000 membros ativos), *ZuluTrade*⁷ (rede de compartilhamento de estratégias de investidores) e *eToro*⁸ (rede social de negociação e corretagem criada em Israel em 2007, possui mais de 4 milhões de usuários segundo informações do próprio *site*). De acordo com [24], esses *sites* tem atraído a atenção de brasileiros que não têm condições financeiras suficientes para contratar corretoras e mesmo assim se interessam em aprender estratégias de negociação e investir no mercado de ações.

No Brasil está sendo disponibilizado um portal (portal do investimento⁹) que disponibiliza dados oriundos de fontes da internet processados por técnicas computacionais para auxiliar o investidor na tomada de decisão para negociação no mercado de ações. Neste *site* são disponibilizados dados sobre ações do mercado obtidos de documentos HTML, tweets, blogs e comunidades de redes sociais e análise de sentimento para alguns ativos.

As mídias sociais têm tornado as redes sociais *online* ubíquas, de forma que estas estão fazendo parte do cotidiano das pessoas. Tal fato tem gerado massivas quantidades de dados para análise empírica, sendo fonte de pesquisa em dinâmica social, estrutura de redes e padrões globais de fluxo de informação. As redes sociais são hoje um fator crítico para a disseminação da informação, pesquisa, marketing, descoberta de influência e potencialmente, uma ferramenta para mobilização de pessoas [25].

Segundo [1, 26], as redes sociais online ultrapassaram o e-mail e se tornaram a atividade mais popular online. Em média, um americano adulto passa cerca de três horas online por dia, dessas 37 minutos são em redes sociais e 33 em e-mail [1]. A Internet firmou-se como a segunda mídia mais

¹Termo aplicado aos sistemas que dão suporte ao uso da tecnologias para conectar pessoas [23].

²<http://www.facebook.com>, <http://www.linkedin.com>, <http://www.instagram.com>, <http://orkut.google.com>

³<http://www.twitter.com>, <http://www.flickr.com>, <http://www.youtube.com>, <http://www.myspace.com>

⁴<http://www.cv-redebibliosus.bvs.br>

⁵<http://www.ravelry.com>

⁶<http://www.stocktwits.com>

⁷<http://www.zulutrade.com>

⁸<http://www.etoro.com>

⁹Portal do Investimento: <https://observatoriodoinvestimento.com>

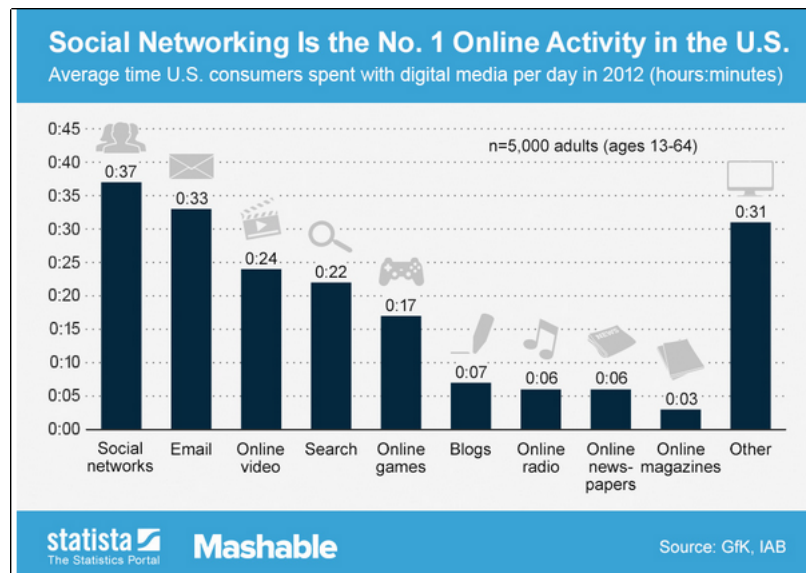


Figura 2.1: Tempo gasto por americanos entre 13 e 64 anos com atividades *online* [1].

importante para as pessoas, perdendo apenas para a televisão. A Figura 2.1 apresenta o tempo médio diário gasto por americanos entre 13 e 64 anos com diferentes atividades *online*.

No Brasil, conforme pesquisa divulgada em 2013 pelo Instituto Brasileiro de Geografia e Estatística (IBGE), o acesso à Internet continua em franco crescimento. De 2005 a 2011, cresceu em 143,8% o acesso na população com 10 anos ou mais de idade, entretanto, o ingresso no mundo digital ainda não alcança 53,5% dos brasileiros nessa faixa etária [2]. Segundo publicação do Jornal Folha de São Paulo [27] sobre pesquisa encomendada pela Secretaria de Comunicação Social da Presidência da República, o brasileiro gasta mais tempo acessando à Internet do que assistindo à televisão ou ouvindo ao rádio. Sobre hábitos de navegação dos brasileiros, um estudo realizado por uma empresa especializada em inteligência de mercado e gestão do relacionamento nas redes sociais - *E.life* - revelou que 98% dos entrevistados passam parte do tempo em redes sociais como Facebook 81%, Google+ (71%) e Instagram com 22%, esse último está em crescimento na preferência [28]. Outro fator interessante apontado por esta pesquisa é que o brasileiro costuma assistir televisão ao mesmo tempo em que acessa à Internet, e muitos entrevistados admitiram pautar suas escolhas de programação baseados nos comentários das redes sociais. A Figura 2.2(a) apresenta a porcentagem de brasileiros com acesso à Internet em relação ao território nacional segundo pesquisa divulgada pelo IBGE [2].

Em abril de 2015, o IBGE divulgou como resultado da Pesquisa Nacional por Amostra de Domicílios (PNAD) de 2013 que a Internet chegou a 49,4% da população brasileira, desses 4,1% acessam apenas por dispositivos móveis, Figura 2.2 (b), e deste acesso, cerca de 97,7%, é feito por banda larga, sendo que sua utilização cresce de acordo com a escolaridade, variando de 5,4% para pessoas sem instrução até 89,8 para os que possuem mais de 15 anos de estudo [3, 29]. No final de 2015, a BBC Brasil informou que o país alcançará o quarto lugar em maior população de usuários de Internet no mundo, segundo consultoria realizada por empresa de tecnologia [30]. A informação vem acompanhada da notícia de que tamanha quantidade de acessos foi impulsionada também

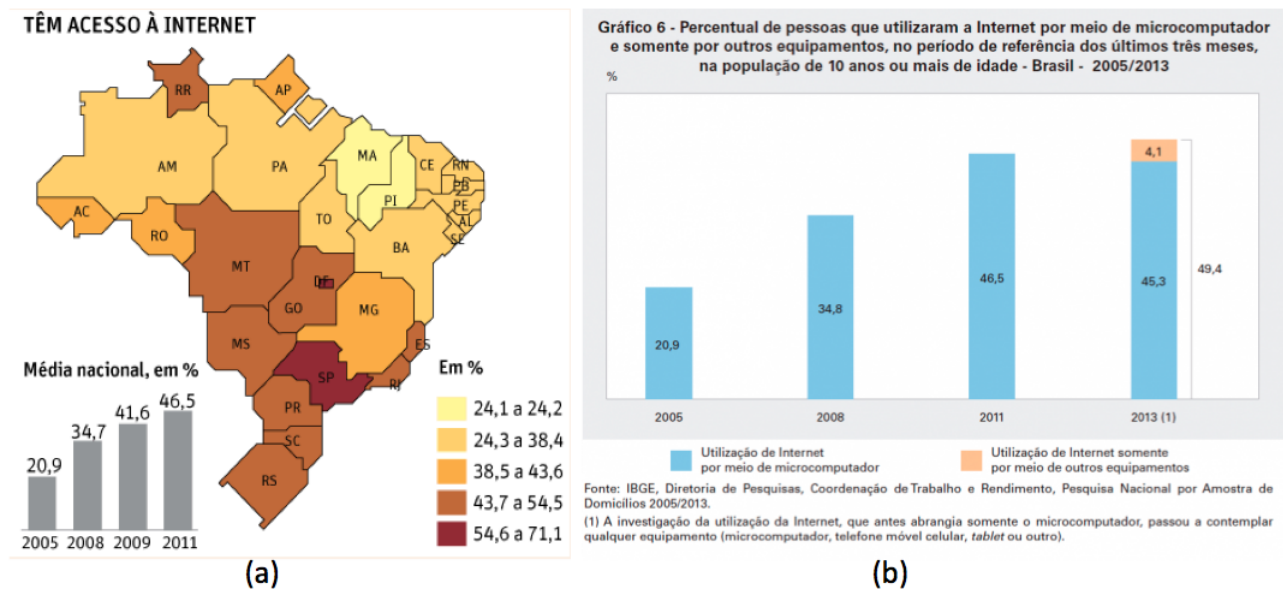


Figura 2.2: (a) Percentual da população com acesso à internet segundo pesquisa do IBGE [2]. (b) Forma de acesso à internet pelo brasileiro segundo pesquisa PNAD - IBGE [3].

pela oferta de dispositivos móveis com conexões de banda larga mais baratas. Há ainda muitas pessoas sem acesso, muitos por possuírem baixa renda mensal e outros por viverem em zonas rurais, entretanto, nas regiões mais distantes dos grandes centros, o acesso tem sido realizado por meio de celulares smartphones [29].

Outro fator interessante é o recente interesse por parte do brasileiro por mercado de capitais e conseqüentemente o uso das redes sociais para expressar sua atuação nesse mercado [24], reforçando assim as possibilidades de pesquisa e estudo nessa área.

2.2 Análise de sentimentos e opiniões

O aumento de plataformas para redes sociais *online* e sua crescente popularização através do uso de Internet em computadores pessoais e celulares smartphones tem permitido ao público registrar: suas impressões através de vídeos e imagens; e seu ponto de vista expressado por descrição de pensamentos, opiniões e sentimentos sobre qualquer assunto, em qualquer lugar, e até mesmo na hora exata em que um fato acontece. Esse comportamento tem resultado na disponibilização de um enorme e crescente repositório de dados com contribuições de usuários sobre uma infinidade de assuntos [14]. A exploração desse conteúdo tem sido alvo de pesquisas e grandes desafios para a mineração de textos e descoberta de conhecimento.

Capturar opiniões com a finalidade de observar a dinâmica do pensamento humano em redes sociais não é uma tarefa fácil, tendo em vista que, a cada dia, mais usuários são inscritos nesses sistemas e conseqüentemente mais mensagens são compartilhadas. Por conseguinte, o volume de dados está sempre em franco crescimento. Esta questão torna a atividade de análise e obtenção de conhecimento um feito praticamente impossível sem o uso de automação. Outro fator desafiador

é a exploração desses dados em tempo hábil, pois tamanha é a diversidade e ausência de estrutura formal nas construções textuais, o que dificulta a extração de informação útil.

Entretanto, desvendar o pensamento da multidão tem se tornado um elemento de importância estratégica para pessoas, empresas, organizações de saúde e até mesmo para agências governamentais [14].

A análise de sentimentos ou mineração de opiniões é a tarefa responsável por encontrar opiniões de autores sobre entidades específicas [31]. Atualmente, ao comprarem um produto, é comum aos usuários de Internet procurarem por comentários escritos por outros sobre este, de forma que as opiniões e pensamentos disponibilizados nas redes sociais e sites passam a compor o processo decisório das pessoas. E essa é uma das razões pelas quais o tema análise de sentimentos e extração de conhecimento das redes sociais *online* está sendo bastante explorado e desejado, tanto na academia, quanto por empresas governamentais ou não.

Blogs, fóruns *online*, sites com painéis de mensagens e redes sociais tais como *Twitter*[®] e *Facebook*[®] disponibilizam os pequenos textos carregados de sentimentos produzidos por seus usuários. Esses textos são de grande importância para empresas e pessoas que desejam monitorar sua reputação e obter, em tempo oportuno, um retorno sobre seus produtos e atuação. Agir de acordo com o que o povo está pensando é o desejo de políticos, marqueteiros, investidores, etc., e a análise de sentimentos habilita a monitoração em tempo real de diferentes mídias sociais, possibilitando extrair a dinâmica do sentimento dos usuários em relação ao domínio pesquisado [31].

As mensagens acessíveis nas redes sociais podem expressar opiniões, sentimentos e até mesmo o estado emocional e o humor do autor sobre si mesmo, amigos, eventos, epidemias, produtos, serviços, programas de televisão, celebridades, política, religião, economia, etc.. Nesse contexto, a análise de sentimentos tem sido empregada para:

- verificar a repercussão de eventos, comportamento de pessoas e ações de promoção de empresas, produtos e serviços;
- análise de opiniões sobre produtos e serviços;
- como variável agregadora em sistemas de predição de dinâmica humana para análise de mercados de ações, bilheteria de filmes, epidemias na área da saúde entre outros.

Extrair informações úteis, sentimento positivo ou negativo, e estados de humor de textos com escrita livre de formalismos e total descompromisso com estruturas linguísticas é o desafio e objetivo de vários pesquisadores da área. Os textos coletados das redes são em sua maioria editados em linguagem coloquial, podem possuir conteúdo carregado de ironia, sarcasmo, sentimentalismo, erros ortográficos, truncagem de palavras, mistura de idiomas, caracteres especiais e emoticons¹⁰. Apesar de haver uma infinidade de publicações científicas na área, sendo esse um problema de processamento de linguagem natural - *Natural Language Processing* (NLP, do inglês), ainda se encontra em aberto.

¹⁰Cadeias de caracteres ou imagens pequenas que traduzem o sentimento de quem está escrevendo a mensagem.

Vários trabalhos atuais exploraram métodos para análise de sentimentos em mensagens de redes sociais [32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42]. As abordagens existentes podem ser agrupadas em três categorias: rotulação de palavras-chave (por exemplo: feliz é positivo, triste é negativo); afinidade léxica (dicionários de palavras chamados de lexicons que pontuam palavras); e métodos estatísticos (algoritmos Bayesianos como *Support Vector Machines* - SVM, do inglês - e outros algoritmos de aprendizado de máquina) [43]. Detalhes sobre tais métodos podem ser encontrados em [44, 45, 16, 46, 31, 47].

Objetivando apresentar estudos comparativos sobre técnicas de análise de sentimentos, [48] e [49] testaram alguns métodos. O primeiro selecionou e experimentou oito e os comparou em termos de cobertura - quantidade de mensagens cujo sentimento é identificado - e concordância - fração de mensagens cujo sentimento identificado é verdadeiro. O segundo comparou métodos tendo como base o uso de dicionários - lexicons, conjuntos de treinamento e dependência de idioma.

Analisar o sentimento humano em relação aos fatos divulgados na mídia brasileira é a questão explorada por [50]. Este avaliou a reação positiva ou negativa de usuários do Twitter em relação às notícias selecionadas. Também, no Brasil, [51] considerou conexões entre tweets para verificar a possibilidade de aprimoramento de sistemas de classificação de sentimento de textos. Para isso, montou uma rede de coocorrência de hashtags¹¹ de tweets e utilizou de características de estrutura de grafos, conceitos de mineração e assortatividade (quando nodos com muitas conexões tendem a se conectar com outros nodos com muitas conexões).

Desenvolver um sistema eficiente de análise de sentimentos de mensagens obtidas do Twitter e de domínio específico é o objetivo de [52]. Para isto, adotou uma abordagem consistindo de quatro etapas: colecionamento de tweets (o usuário passa ao sistema um conjunto inicial de termos e este encarrega automaticamente de coletar os tweets através de um algoritmo para expansão de hashtags proposto); refinamento (remoção de informação espúria através de remoção de spams e outros); criação de um lexicon de sentimentos de domínio específico (foi adotada uma metodologia não supervisionada através de um algoritmo de propagação em grafo); e análise de sentimentos (baseada em aprendizado não supervisionado). Adotaram também quatro lexicons diferentes para averiguar a qualidade do sistema e obtiveram uma precisão de 90%.

Atualmente, existem *sites* que disponibilizam análise de sentimentos para auxiliar pessoas na tomada de decisão em negócios, política, mercado de ações e outros. O Sentdex¹², uma empresa com foco em análise de big data disponibiliza gratuitamente aos usuários cadastrados um gráfico de acompanhamento de sentimento juntamente com o gráfico de evolução de preço de ações do mercado americano. O *site* informa que nem sempre sentimento e preço se encontram perfeitamente alinhados. O sentimento é obtido através de análise de textos de tweets e notícias.

Outra empresa que também oferece software pago para análise de sentimentos do mercado de ações é o The Stock Sonar¹³. Ele apresenta junto ao gráfico de preços o sentimento positivo

¹¹Termo usado para destacar algum assunto importante em textos nas redes sociais *online*. É precedido do símbolo #, por exemplo #redessociasbrasileiras e seu uso torna o tweet facilmente detectável por qualquer pessoa interessada no assunto.

¹²<http://sentdex.com>

¹³<http://www.thestocksonar.com/>

e negativo sobre uma determinada ação ao longo do tempo. O sentimento é obtido através da análise de textos obtidos de fóruns, blogs, tweets, documentos e outros.

Empresas que disponibilizam o histórico de preços de ações e gráficos de evolução diária de preços como Reuters¹⁴ e Yahoo¹⁵ apresentam também análises de opinião para auxiliar a tomada de decisão em seus *sites*.

2.2.1 Repercussão e Opinião nas redes sociais

Canais para comunicar opiniões em redes sociais são fontes importantes de conhecimento para pessoas públicas, empresas, organizações governamentais ou não para:

- controle de difamação;
- acompanhamento de lançamentos;
- contato e atendimento direto ao cliente ou interessados;
- conhecimento do negócio, análise do impacto de um revisor influente no poder de compras de outros;
- planejamento de novos produtos;
- divulgação e manutenção da integridade da marca.

À medida que uma maior quantidade de público tem acesso às tecnologias, cresce a importância de tais opiniões e o desafio de lidar com elas em termos de interpretação contextual; pois, como relatado anteriormente, trata-se de um problema NLP e ainda não existem sistemas robustos para lidar com textos com tamanha variedade de características.

Várias pesquisas têm concentrado esforços em desvendar o pensamento humano expresso nas mensagens de opinião, [53] propõe um sistema que combina análise de sentimentos com dados rotulados manualmente e máquinas de aprendizagem para extrair do conjunto de mensagens, as que contêm opiniões. Em [54], é analisado o uso da plataforma de programação **R**¹⁶ para testar diferentes dicionários léxicos e esquemas de classificação, o pesquisador argumenta que tal ferramenta simplifica a tarefa de análise de sentimentos e mineração de opinião.

Um fato interessante sobre repercussão em redes sociais é o caso da eleição do presidente americano Barack Obama em 2008, que tornou redes como Twitter, Facebook e outras partes integrantes do ferramental das campanhas políticas. Segundo [55], a rede social mantida por Obama nas eleições presidenciais americanas de 2008 o fez bater recordes de mobilização e doações. Uma investigação no contexto da eleição federal alemã é feita por [56] para verificar o uso do Twitter como fórum para deliberação política e se as mensagens *online* serviam como espelho do sentimento político fora da Internet. Em [57], é discutido como a mídia social molda a esfera pública e facilita

¹⁴<http://www.reuters.com/finance/stocks>

¹⁵<http://finance.yahoo.com>

¹⁶Projeto de Computação estatística - <http://www.r-project.org> .

a comunicação entre comunidades com diferentes orientações políticas. Com o uso de algoritmos de rede e rotulação de dados é mostrado que a rede de tweets políticos é uma estrutura partidária altamente segregada e com conectividade extremamente limitada entre os usuários de esquerda e direita política.

Analisando dados do Twitter em relação a quem estava falando para, realizando retweets e falando sobre os motins e tumultos com mortes ocorridos em Londres, em agosto de 2011, [58] procurou identificar e compreender as redes sociais em eventos de crise. Explorando a forma como pessoas reagem em momentos de crise, em termos de reações de limpeza das ruas e orações; mensagens foram coletadas a fim de buscar respostas rápidas para gestão de emergências, comunicação governamental transparente e eficaz, recuperação e prestação de apoio.

Uma análise sobre o papel do Twitter na formação e facilitação de movimentos sociais, especialmente durante protestos é feita por [59]. Analisando dados coletados durante um protesto público em Delhi, capital da Índia, os resultados identificaram o Twitter como um importante canal para difusão de ideias e notícias, desafiando fronteiras geográficas, e também o notável papel dos usuários atuando como "jornalistas-cidadãos" durante os dias de protesto. Os resultados sugeriram que os grandes atores no Twitter foram também os líderes dos protestos nas ruas.

Em [60], são explorados métodos computacionais para medir o impacto das mídias sociais em um movimento social. Analisando dados do Twitter relacionados ao movimento Occupying Wall Street (OWS, do inglês)¹⁷, o pesquisador demonstrou uma correlação entre a vitalidade do movimento e o volume de tweets no tempo. Ao classificarem os usuários com base na quantidade e tempo gasto com tweets relacionados ao OWS, foram capazes de identificar os geradores de "burburinho" e, baseado na quantidade de retweets (retransmissão de tweets), seu poder de influência na rede.

As redes sociais online adicionam uma dimensão extra à dinâmica de celebridade, elas criam suas próprias celebridades que passam a promover causas e interesses. Várias características permeiam esse status na rede, por exemplo, uma celebridade é um influenciador bem conhecido, entretanto, nem todo influenciador e nem toda pessoa bem conhecida é uma celebridade na rede. Buscando identificar celebridades em dados do Twitter, [61] desenvolveu um modelo computacional de pontuação baseado em atributos da psicologia social (fama - pessoas que a seguem, simpatia - pessoas curtem suas mensagens e identificação - quão simpáticas são as pessoas que a curtem) que definem uma pessoa célebre. O modelo obteve bons resultados na identificação de celebridades no Twitter.

O problema de sumarização de opiniões para entidades como celebridades e marcas foi estudado por [62], que desenvolveu um *framework* para resumir opiniões centradas em entidade e baseadas em tópico. Para isto, adotou mineração de tópicos a partir de hashtags, algoritmos de propagação de afinidade para agrupar tópicos similares e algoritmos geradores de paráfrase para resumir tweets com conteúdo expressivo, por fim, aplicaram um analisador de sentimentos, baseado em lexicon, para identificar a opinião expressa no tweet.

¹⁷Movimento que iniciou em 2011 em Manhattan, Nova York, em protesto contra a desigualdade econômica, social, ganância, corrupção e indevida influência das empresas no governo dos EUA.

Identificar formadores de opinião em um tema de interesse, ou seja, pessoas que adotam e espalham novas ideias em redes sociais com sucesso é o objeto de estudo de [63], o qual apresenta uma estratégia para encontrá-los. Combinando atributos temporais de nós e arestas da rede com um algoritmo baseado em classificação de páginas (usado para avaliar a influência de usuários na mídia social), concluíram que usuários no topo da classificação tendem a ser pioneiros a influenciar seus contatos na adoção de uma nova ideia.

Pensando em auxiliar empresas no trato com questões emergentes e em tempo oportuno como: "qual a próxima grande ameaça ou oportunidade para o meu negócio?", [64] desenvolveu um sistema para descoberta e identificação automática de tópicos emergentes associados a produtos de interesse. Em [65], o autor também utilizou de opiniões compartilhadas através do Twitter para ajudar empresas a tomarem decisões sobre suas campanhas publicitárias. Uma abordagem focada no termo verbal existente na mensagem, adotada por [66] e que, segundo esse, trata-se do elemento mais importante para expressar opiniões sobre questões sociais, foi desenvolvida para levantar as diferenças entre análise de sentimentos de mensagens sobre produtos e questões sociais. Os resultados obtidos mostraram uma melhora no desempenho de sistemas de análise de sentimento para opiniões sobre questões sociais. Outro trabalho também focado em questões sociais foi realizado por [67]. Nesse, é proposto um modelo de análise de sentimento de tweets para identificar se o texto expressa uma opinião positiva ou negativa sobre uma entidade, no caso, um político. É baseado em três módulos: um responsável por extrair palavras opinativas das sentenças; outro para associar a opinião com cada entidade relevante; e outro que calcula a pontuação de sentimento para cada entidade. O sistema foi testado com dados coletados sobre as eleições federais australianas de 2010. Os resultados obtidos demonstraram que o sistema desenvolvido obteve bom desempenho, mas precisa ser melhorado.

A análise de comentários e avaliações de clientes em lojas *online* e sua classificação em positivo ou negativo foi o objeto de estudo de [68]. Em seu trabalho, disponibilizou uma ferramenta visual e interativa para mostrar o potencial da técnica baseada em discriminação adotada para extrair termos objetos de um parecer positivo ou negativo em comentários, e um método de ponderação de distâncias para mapear atributos para opiniões positivas e negativas no texto.

Uma epidemia de Dengue¹⁸ pode ser refletida por mensagens postadas no Twitter, e essas podem ser utilizadas por órgãos específicos para a fiscalização de doenças. Exemplo disso é o resultado do trabalho realizado por [69], que se baseou em volume de dados, localização e percepção do público através de análise de sentimento para propor um *site* observatório da doença que permite, como resultado da análise, o acompanhamento da evolução da doença.

2.3 Aspectos da Comunicação Social

Em [70] é proposta uma metodologia para descobrir entre os milhares de usuários do Twitter aqueles que são especialistas em temas. A metodologia utilizada por esse trabalho extrai infor-

¹⁸Doença infecciosa transmitida por picada dos mosquitos (*Aedes Aegypti* e *Aedes Albopictus*), muito comum em regiões tropicais e sub-tropicais.

mações de listas do Twitter construídas com a colaboração dos usuários. Nestas são adicionados os especialistas sobre temas que mais lhes interessam. Baseado na ideia de que um usuário sendo seguido por muitos outros, a respeito de certo tópico, é certamente um especialista nesse tópico, desenvolveram um sistema buscador que captura temas que ocorrem com frequência na lista de metadados e os associa às listas dos usuários.

Redes sociais baseadas em localização permitem aos usuários: compartilharem sua posição geográfica com seus amigos; buscarem por locais de interesse; e postarem dicas sobre locais existentes. Os usuários de tais redes frequentemente lidam com *spam*, os quais acrescentam às localidades mensagens de propagandas não solicitadas. Identificar *spam* na rede social brasileira Apontador¹⁹ foi objeto de estudo de [71]. Esse se baseou em uma coleção de apontadores rotulados e fornecidos pela rede juntamente com informações sobre usuários e localizações para distinguir apontadores de localidade spam ou indefinidos. Os resultados obtidos demonstram a relação que as localizações e atividades do usuário têm com spam na rede. Outros trabalhos tratam da identificação de spam em redes sociais específicas como Twitter [72], Facebook [73] e MySpace [74]. Em [75], é proposto um *framework* para detecção de spam que pode ser utilizado por qualquer *site* de rede social.

A Internet reflete os interesses e valores da sociedade. Ela funciona como um espelho no qual cientistas e pesquisadores podem olhar e analisar comunidades através de um enorme espaço observacional [76]. Dentro desse contexto, estudos sobre o comportamento de homens e mulheres em redes sociais têm sido realizados. Em [76], é investigada a conduta de ambos os gêneros em relação à escolha de uma mesma ou não hashtag ao discutirem sobre um mesmo tópico no Twitter. Em [77], também é apresentado um estudo sobre diferenças de gênero no comportamento dos usuários na rede Twitter.

Analisando comentários e opiniões de usuários, [78] revelou características interessantes sobre perfis de usuários e cultura brasileira. Tal análise foi realizada a partir de dados obtidos de um *site* de compartilhamento de receitas culinárias brasileira - www.tudogostoso.com.br. Através da caracterização de usuários, receitas e ingredientes, descobriu-se que a maioria dos comentaristas são do gênero feminino e que os ingredientes utilizados por boa parte das receitas fornecem indícios de padrões da culinária brasileira.

Link farming (tradução livre - fazenda de ligações) é uma prática que envolve a montagem de uma rede de sites que possuem conexões entre si com a finalidade de aumentar sua relevância quando acionados por algoritmos buscadores. Em seu trabalho, [72] investigou tal prática no Twitter e mecanismos que a desencorajem. Para tal, analisou quarenta mil contas de spammers suspensas pelo Twitter e mostrou que um esquema simples de classificação pode penalizar usuários que se conectam a outros spammers, diminuindo assim a influência desses últimos.

Um estudo sobre como os usuários navegam e interagem quando conectados em redes sociais é apresentado por [26]. Nesse trabalho, foram coletados e analisados dados sobre cliques de usuários que acessaram quatro redes: Orkut; MySpace; Hi5; e LinkedIn, a partir de um *site* agregador brasileiro. O estudo mostrou que a navegação ocupa 92% das atividades dos internautas e que o compartilhamento de conteúdo é feito, geralmente, entre amigos próximos geograficamente. Esse

¹⁹www.apontador.com.br

estudo também discutiu questões para melhoria de interface de sistemas de redes sociais *online*, inserção de propaganda e remodelamento de tráfego de Internet, importante para concepção de sistemas de distribuição de conteúdo futuros.

Pesquisas sobre o comportamento dos usuários em relação as suas interações, tanto com cunho tecnológico quanto comportamental, foram e têm sido realizadas para várias redes tais como Facebook [79, 73, 80, 81, 82, 83], Flickr [84], Twitter [77, 85] e Google+ [86]. A análise da navegação de usuários nessas redes também é estudada por [87]. Uma abordagem nos fatores que levam estudantes a utilizarem as redes *online* e os impactos sociais advindos dessa escolha é adotada por [88].

Em [89], é exposto um estudo sobre os determinantes da participação dos usuários da comunidade online, a partir de uma perspectiva de influência social. Identificar fatores que influenciam na escolha, pelo usuário, de mensagens para responder, dentro do universo de mensagens recebidas, é o objetivo de [90]. Em [91], é apresentado um estudo sobre usuários do Twitter, seu comportamento, padrão de crescimento e tamanho da rede.

Argumentações de que as estruturas das redes sociais não revelam as interações atuais entre pessoas e que o ritmo de vida dessas influencia nos padrões de interação em redes sociais são abordadas por [92].

O desafio para entender como a estrutura da rede afeta a dinâmica do espalhamento da informação e como tal fator é crítico para o uso efetivo da mídia social e desenvolvimento de sistemas é discutido por [25]. Com dados obtidos das redes Twitter e Digg²⁰, a análise realizada mostrou que a informação se propaga mais rapidamente em redes mais densas (com maior interconexão de usuários) como o Digg, entretanto, alcança usuários mais distantes no caso do Twitter.

Estudar a multipolarização gerada por contextos de discussão de temas em redes sociais é o objetivo de [93]. Com uma estratégia proposta para minerar redes sociais em busca de relações de apoio, antagonismo e indiferenças em redes multipolarizadas, entre outros, provou-se que usuários mais próximos na rede nem sempre possuem ideias similares. Para tal estudo, foram utilizadas mensagens coletadas do Twitter relacionadas ao campeonato brasileiro de futebol nos anos 2010, 2011 e 2012.

Comunidades científicas e seus aspectos dinâmicos são outro ponto interessante e atualmente pesquisados pela academia. Saber o papel desempenhado por diferentes membros dessas comunidades na formação e evolução da estrutura da rede é o objetivo de [94]. O estudo mostra que membros centrais de uma comunidade funcionam como pontes para conexão com grupos de pesquisa menores, isto é, comunidades adjacentes. É proposta uma estratégia para inferir o centro da comunidade, ou seja, os líderes de uma dada comunidade científica em um dado período de tempo, e são investigados como aspectos desses centros impactam a estrutura da comunidade adjacente. Foram usados dados obtidos das principais conferências ACM/SIG²¹ através do DBLP *Computer Science Bibliography*[95], *website* com dados de bibliografia de ciências da computação. Outros trabalhos na área de comunidades científicas também podem ser encontrados em [96, 97, 98].

²⁰<http://digg.com> - Agregador de notícias, vídeos e links enviados e avaliados pelos seus usuários.

²¹Association for Computing Machinery / Special Interest Groups - <http://www.acm.org/sigs> .

Outro tema bastante explorado na área de redes sociais e mineração de textos é o das redes sociais acadêmicas. Existem vários grupos de pesquisa desenvolvendo trabalhos, estudando características, o relacionamento entre pesquisadores de áreas, autores e coautores nessas redes, e especialmente no Brasil podem ser citados [99, 100, 101, 102].

2.4 Estimação e redes sociais

2.4.1 Estimação no mercado de ações

O burburinho das comunidades sociais online pode ser usado para realizar previsões quantitativas. Em seu trabalho, [5] considera a possibilidade de prever receita de bilheteria de filmes com dados do Twitter. Na tentativa de provar que um filme, bem comentado na rede, provavelmente será bem sucedido na bilheteria, construíram um modelo de regressão linear para a previsão de receitas de bilheteria de filmes antes de sua estreia. O estudo provou que os resultados obtidos conseguiram superar os do Hollywood Stock Exchange (HSX²², do inglês), e que há uma forte correlação entre a quantidade de atenção dada a um filme nas redes sociais e sua futura classificação. A Figura 2.4 mostra os valores preditos com o uso de tweets e os do HSX.

A análise de dados de microblogs relacionado ao mercado de ações pode revelar novas perspectivas ao prever o sentimento de investidores. Através da coleta de mensagens postadas no Twitter sobre nove empresas de tecnologia como AMD, DELL, e-Bay, Microsoft e IBM, precedidas de "\$" o cashtag (caracter utilizado pela comunidade de investidores dos EUA), o trabalho desenvolvido por [4] criou indicadores que foram investigados na modelagem de variáveis do mercado de ações, tais como: retornos; volume de negociação; e volatilidade. Adotando métodos de análise de sentimento e volume de mensagens, a investigação não produziu evidências de que os indicadores de sentimento possam explicar os retornos do mercado. Porém, o volume de postagens foi usado na modelagem de indicadores financeiros de volume e volatilidade e apresentaram resultados promissores, Figura 2.3.

A correlação entre medidas de estado de humor coletivo (positivo, negativo, calmo, alerta, confiante e outros), derivados de dados do Twitter, com o valor do índice Dow Jones Industrial Average (DJIA, do inglês) ao longo do tempo é investigada por [7]. Para correlacionar os estados de humor com valores DJIA, foi adotada a técnica de análise de causalidade de Granger. Para testar a hipótese de que a precisão de modelos preditores de DJIA pode ser melhorada através da inclusão de medidas de humor do público, utilizou-se de uma rede neural difusa auto-organizada. Os resultados obtidos mostraram que a precisão das previsões DJIA podem ser significativamente melhoradas através da inclusão de dimensões específicas de humor do público. Foi encontrada uma precisão de 86,7% na previsão diária de subida e descida de valores de fechamento do DJIA e uma redução na porcentagem do erro médio em mais de 6%.

Análise de sentimentos e frequência de postagens e comentários de redes sociais *online*, juntamente com análise histórica de preços e volume do mercado de ações são usadas por [8] para

²²<http://www.hsx.com> - líder mundial em mercado de ações de entretenimento.

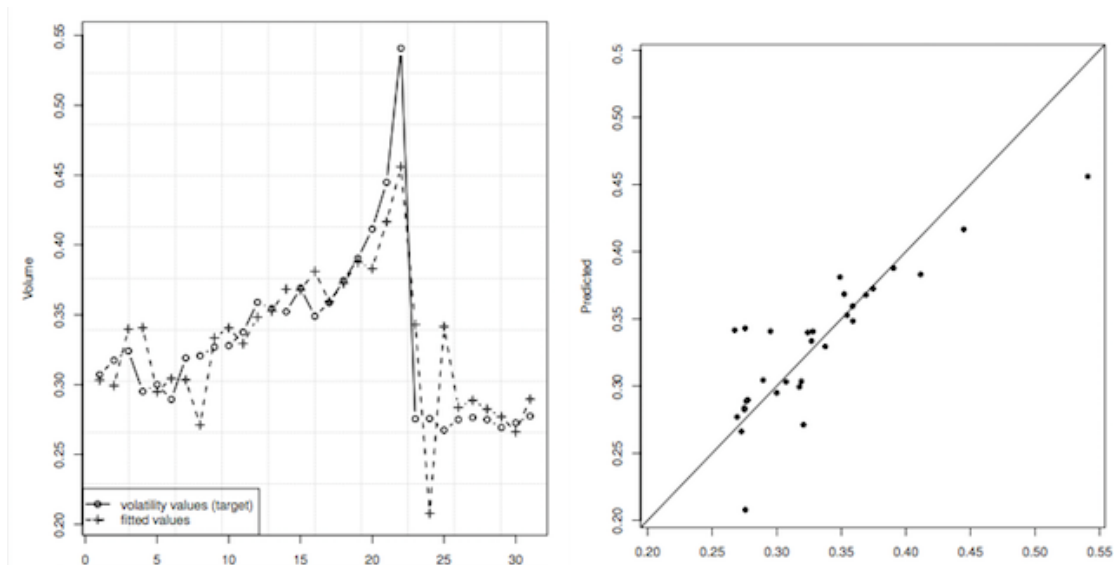


Figura 2.3: Gráficos de valores de volatilidade ajustados e alvo, o eixo horizontal apresenta os valores observados [4].

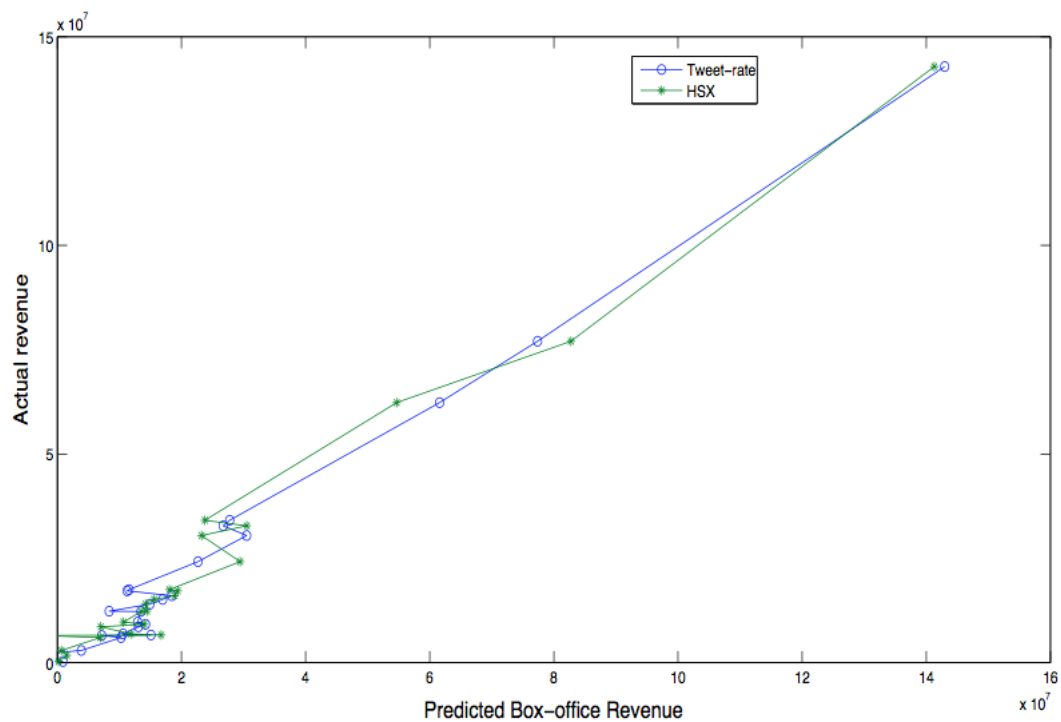


Figura 2.4: Pontuação de bilheteria versus a predita usando dados do Twitter e do Hollywood Stock Exchange obtida por[5].

modelar e estimar os movimentos de preços. A adoção de um framework de regressão com múltiplos kernels de aprendizagem (Multiple Kernel Learning) apresentou resultados que superaram métodos de base como RMSE²³, MAE²⁴ e MAPE²⁵ em termos de magnitude e medidas de predição para ações de três empresas japonesas no mercado de ações americano.

Objetivando prever o comportamento do mercado de ações baseado na coleta de dados de múltiplas fontes (twitter, pesquisas, manchetes de notícias e mecanismos de consultas por termos), [10] definiu uma variedade de indicadores de sentimento e determinou seu valor preditivo sobre uma série de indicadores financeiros tais como DJIA, volumes de negociação, volatilidade do mercado e preço do ouro. Através do uso de testes de causalidade de Granger e análise de correlação entre indicadores, os pesquisadores obtiveram várias contribuições interessantes. Os resultados mostraram que volumes de buscas no Google são bons preditivos de indicadores financeiros. Todos os indicadores de humor estudados exibiram correlação significativa com registros de retornos e volatilidade do mercado, e o volume de dados do Twitter, investigado num período de queda do DJIA, aumentou semanas mais cedo do que os volumes de busca do Google, indicando um ganho de eficiência potencial do primeiro sobre o segundo.

Com um conjunto de mais de 20 milhões de mensagens obtidas do site LiveJournal, o trabalho de [103] demonstrou que estimar emoções, através de weblogs, pode fornecer novas informações sobre futuros preços do mercado de ações. Coletando dados, obtendo rótulos (ansioso, preocupado, nervoso ou temeroso) dos usuários postadores do LiveJournal e usando um quadro causal de Granger, puderam observar o aumento da expressão de "ansiedade" nesses dados, prevendo assim uma pressão sobre o índice S&P 500²⁶, resultado confirmado através de simulação de Monte Carlo.

Dados do serviço de microblog específico para o mercado de ações, o Stocktwits²⁷, foram coletados por [104] durante três meses. Com o uso de um classificador de sentimentos baseado em aprendizado de máquina, foi verificado que os sentimentos capturados possuem grande valor preditivo para futuras direções no mercado de ações.

O emprego de métodos de linguística computacional por [105] possibilitou a estimação do volume de negociação do dia seguinte, isto através de análise de sentimentos de tweets e associação com retornos acionários anormais e volume de mensagens. As descobertas demonstraram que mensagens de usuários com bons conselhos de investimentos são geralmente passadas adiante (retweeted) por outros e que esses usuários possuem maior quantidade de seguidores, o que consequentemente, amplia sua influência em fóruns de microblog.

Avaliar se indicadores de sentimento público, extraídos de mensagens diárias do Twitter, podem melhorar a previsão de indicadores comerciais, econômicos e sociais foi o estudo de [14]. Para tal, coletaram dados de março de 2011 à dezembro de 2013 nos domínios mercado de ações e receitas de bilheteria, que foram utilizados em modelos de previsão. Resultados obtidos mostraram que modelos não lineares funcionam melhor com dados do Twitter ao estimarem índices de volatilidade,

²³RMSE - Root mean square error - raiz quadrada do erro médio quadrático.

²⁴MAE - Mean absolute error - erro absoluto médio.

²⁵Mean Absolute Percentage Error - Erro percentual absoluto médio.

²⁶Satandard & Poor's 500 - índice do mercado de ações baseado nas capitalizações de 500 grandes empresas com ações na bolsa de valores de Nova York ou NASDAQ.

²⁷<http://www.stocktwits.com>

enquanto que os lineares falham ao prever qualquer tipo de série financeira. No caso de previsão de receitas de bilheteria, foi utilizada máquina de suporte vetorial que também obteve bons resultados com dados das redes sociais.

Com um conjunto de dois milhões de dados coletados do Twitter e índices de volume de busca do Google, o trabalho de [106] modelou uma série de relações causais sobre esses dados para títulos de mercado tais como: capital próprio DJIA; Nasdaq-100²⁸; mercado de mercadorias (commodities); óleo; ouro e taxas do Euro Forex²⁹. Os resultados demonstraram que há correlação entre volume de busca e preço do ouro e que os modelos de previsão utilizados apresentaram uma redução significativa do percentual do erro médio.

Um estudo sobre os desafios no uso do Twitter para realizar previsões sobre ações e uma análise sobre várias técnicas de aprendizagem de máquina para analisar o sentimento de tweets é apresentado por [107]. Buscando obter uma correlação entre o sentimento e preços de ações, determinaram através de uma análise da mudança de preço e tweets, quais palavras contidas nesses são correlatas a modificações nos valores das ações.

Utilizando dados de ações de duas companhias líderes de petróleo no mundo, BP América e Saudi Aramco, [108] avalia a variância entre a análise de sentimentos automatizada e a classificação humana. Procurou-se entender como a motivação para postagens de mensagens (feita por usuários do Twitter do Ocidente e Oriente Médio que mencionam tais empresas) afeta a qualidade da classificação. Os resultados apontam para um questionamento sobre a confiabilidade de sistemas analisadores de sentimentos, pois dependendo da mensagem, cultura e relacionamento do usuário com as empresas, as análises feitas pelo sistema e pelo humano produzem resultados significativamente diferentes.

Uma investigação sobre o poder preditivo do tráfego diário de dados não estruturados oriundos de comunidades *online* em relação aos retornos diários de ações é realizado por [6]. No intuito de produzir indicadores para análise do mercado de ações com base em características de tráfego, verificou-se que a qualidade das previsões aumenta quando um nível elevado de tráfego é acoplado a uma baixa volatilidade do mercado, enquanto um nível elevado de tráfego com alta volatilidade gera reações tardias para movimentos violentos do mercado provocando, como consequência, uma previsão ruim.

Em [109], é apresentada uma medida de sentimento do investidor baseado no índice de felicidade nacional bruta do Facebook (Facebook's *Gross National Happiness Index* - GNH, do inglês). Esse índice é calculado através de análise textual das palavras com teor emocional postadas pelos usuários do Facebook. Para comprovar a afirmação de que o sentimento do investidor tem a habilidade de prever mudanças diárias em retornos e volume de negociação do mercado de ações americano, o pesquisador usou de modelos vetoriais autorregressivos para examinar a relação entre o GNH e a atividade diária do mercado de ações.

No Brasil, [110] apresenta um estudo a respeito de mineração de opiniões sobre ativos e propõe uma metodologia para avaliá-las. Neste, textos são obtidos de portais *web* de notícias de finanças

²⁸Índice da bolsa norte-americana NASDAQ que reúne 100 das maiores empresas não financeiras.

²⁹*Foreign exchange* - mercado financeiro destinado a transições de câmbio.

no mercado brasileiro e processados no intuito de extrair a parte relevante, resumindo assim opiniões.

Em [111], é apresentado um relatório de como fazer para que sinais gerados por uma plataforma de negociação automática chamada MetaTrader³⁰, configurados por um usuário, sejam tuitados através do Twitter. Ou seja, ele propõe um sistema de apoio às decisões sociais - SASS, na qual o cérebro humano é o filtro de tomada de decisão para um sistema computacional. Assim, usuários podem ler tweets com as opções de negociação tuitadas por outros usuários, que na verdade foram geradas pelo MetaTrader e decidir se adota ou não a estratégia de negociação exposta.

Uma situação de teste que não está detalhada em artigo científico, mas que é bastante interessante com relação ao contexto da pesquisa a ser descrita neste trabalho, está detalhada em [112], uma postagem publicada na plataforma Quantopian³¹ (permite aos usuários implementarem algoritmos e executá-los utilizando histórico de dados de 13 anos da bolsa de valores). Nesta postagem, o autor explica que realizou testes baseados em dados obtidos do PsychSignal³² (plataforma que possui um histórico de dados da área financeira que é derivado de um mecanismo de processamento de linguagem natural, o qual realiza a rotulação de mensagens em otimista ou pessimista), que refletem o sentimento de mensagens obtidas da rede social de investidores StockTwits. Esses dados foram utilizados para verificar se índices de humor de investidores são capazes de medir o pulso emocional dos mercados. Nos testes realizados, os resultados superaram os do mercado, sinalizando muitas oportunidades de estudo.

2.4.2 Estimação na área da saúde

Identificar e responder rapidamente a uma epidemia de saúde é fundamental para reduzir a perda de vidas. Métodos de pesquisa em hospitais levam semanas para informar resultados, por isso, muito se tem investido em estimação da saúde da população através de informações capturadas da Internet. Um sistema que estima atividades de gripe através da reunião de consultas de pesquisa *online* é o *Google Flu Trends*³³. A descrição de um método para analisar grandes quantidades de consultas de pesquisa do Google, com a finalidade de rastrear doenças com sintomas semelhantes à gripe é encontrada em [113]. Tal método permitiu estimar a atividade do vírus influenza semanalmente em cada região dos Estados Unidos em 2009. Um sistema desenvolvido por [114] busca melhorar os sistemas de vigilância de doenças por redes sociais. Esse analisa mensagens do twitter relacionadas à gripe com mais profundidade, tentando obter as que reportam sobre infecção.

Em um artigo publicado pelo jornal *Science* intitulado *The Parable of Google Flu: Traps in Big Data Analysis*³⁴ [115], e comentado por [116, 117], os autores afirmam que o serviço do Google Flue Trends não só superestimou em mais de 50% o número de casos de gripe nos Estados Unidos entre 2012 e 2013, comparando com os valores reportados pelo Centro de Controle e Prevenção

³⁰www.fxpro.pt

³¹<https://www.quantopian.com/faq>

³²<https://psychsignal.com>

³³<http://www.google.org/flutrends> - Sistema online para explorar tendências da gripe ao redor do mundo.

³⁴Livre tradução: A Parábola do *Google Flu*: Armadilhas da Análise do *Big Data*.

de Doenças dos EUA, como também estava com tecnologia desatualizada nos últimos anos. Os autores comentaram também que, após uma atualização disponibilizada pelo serviço, houve uma melhora significativa na estimação, entretanto, ainda superou em 30% os valores do órgão de saúde americano. Mesmo sendo de grande valor, pesquisas como a descrita no parágrafo anterior e refutadas por [115] mostram que a tecnologia é algo em evolução e ainda apresenta grandes desafios.

Outro trabalho que explora a detecção de surto de influenza através do Twitter é realizado por [118]. Em sua pesquisa, coletou quinhentos mil tweets durante dez semanas, e desenvolveu vários modelos de regressão para prever a proporção de pessoas que apresentam sintomas parecidos com gripe baseadas na frequência de mensagens que continham certas palavras-chave. Diante desse cenário, realizaram testes com regressores lineares e concluíram que um simples classificador de palavras melhora os modelos preditores utilizados alcançando uma correlação de 0.78 em relação às estatísticas disponibilizadas pelo Centro de Controle e Prevenção de Doenças dos Estados Unidos. Outro trabalho que também demonstrou o potencial do Twitter para a obtenção de dados de surtos do vírus H1N1 através de coleta e análise de mensagens é reportado em [119].

2.4.3 Estimação de popularidade e repercussão em redes sociais

Estimar se um novo item alcançará popularidade é um fator importante para as empresas que hospedam *sites* de mídia social e seus usuários. Entretanto, a previsão de popularidade em mídia social é desafiadora devido a diversos fatores, entre eles destacam-se: a qualidade do conteúdo; a forma de destaque do conteúdo; e a influência entre os usuários. Modelos estocásticos de comportamento que descrevem matematicamente a dinâmica social dos usuários da rede social Digg são utilizados por [120] para prever a popularidade de uma nova história postada baseada nas reações iniciais de um usuário perante o novo conteúdo. Utilizando observações da evolução do número de votos recebidos por uma história logo após ser postada, é possível prever a quantidade de votos que esta receberá após alguns dias.

O trabalho de [121] discute um modelo para criar genótipos que são resumos de tópicos de interesse do usuário. Os pesquisadores fizeram uso desse modelo para realizarem uma previsão de influência de uma nova propagação de conteúdo.

Em [122], é apresentada uma estratégia para construção de modelos estatísticos da dinâmica da mídia social para estimar a dinâmica do sentimento coletivo. Esse conhecimento pode permitir uma reação proativa contra opiniões e sentimentos negativos do público ou o desenvolvimento de estratégias que dissipem rumores e reverta a situação.

Discussões sobre predição em eleições são realizadas por [123, 56]. Um novo estudo da Universidade de Indiana, nos EUA, comentado por [124] aponta para a predição das eleições americanas usando dados do Twitter. Tal estudo afirma que existe uma relação significativa entre os dados do Twitter e os resultados das eleições norte-americanas. Entretanto, ele também relata sobre os desafios advindos dos dados das redes sociais, pois esses representam apenas uma parcela da população e podem sinalizar um comportamento que não será necessariamente refletivo nas urnas.

Em [125], tweets coletados no âmbito da eleição para reitor da Universidade Tecnológica do Paraná foram polarizados manualmente para realizar a predição desta. Verificou-se após análises e processamento das mensagens capturadas no Twitter que essas refletiram os resultados da eleição.

O Twitter também tem sido utilizado para acompanhar o comportamento das pessoas em desastres naturais. Em [126], é apresentado um método para estimar a localização de um evento através dos dados coletados, para tal, utilizaram comentários acerca de terremotos. Um procedimento para detectar dados relacionados a possíveis terremotos é investigado por [127].

2.5 Pesquisas no Brasil

O Brasil possui alguns grupos de pesquisa trabalhando com redes sociais, seja estudando o comportamento das pessoas em psicologia, antropologia ou sociologia, seja analisando os efeitos da convivência diária com as redes sociais na saúde das pessoas, ou explorando características peculiares relacionadas às tecnologia e sistemas (modelagem de interação entre pessoas, fluxos de dados, influência nas redes, modelagem de conteúdo, sentimentos e outros). Alguns grupos que podem ser identificados através de busca no Google são:

- Rede Social no Instituto Nacional de Ciência e Tecnologia para Web - linha de pesquisa voltada para a caracterização e modelagem topológica de redes sociais para modelagem do comportamento social coletivo, tratamento de informação e desenvolvimento de algoritmos e protocolos para aumentar a eficiência, confiabilidade e segurança de sistemas de informação distribuídos em larga escala conforme relato no *link* <http://www.inweb.org.br/linhas-depesquisa-inweb/redes-sociais-rs/>;
- Grupo de Pesquisa em Interação, Tecnologias Digitais e Sociedade (GTIS) - grupo de pesquisa voltado para a área de comunicação e psicologia - <http://gitsufba.net>;
- Núcleo de Estudos de Redes Sociais - NERDS - grupo com ênfase em sistemas de computação - <http://homepages.dcc.ufmg.br/~fabricio/>;
- Grupo de Pesquisa em Ciberantropologia – GrupCiber - voltado para pesquisas sobre fenômenos sociais engendrados no "ciberespaço" - <http://www.grupciber.net/blog/apresentacao/>;
- Grupo de pesquisa Tecnologias, Culturas, Práticas Interativas e Inovação em Saúde – voltado para o estudo das tecnologias emergentes e redes e os impactos que sua introdução causam na sociedade - <http://wiki.next.icict.fiocruz.br/>.
- Redes de Conhecimento e Informação - Grupo de pesquisa da Ciência da Informação da Universidade Estadual de Londrina - trabalha com identificação, avaliação e seleção de fontes de informação sobre as redes sociais na internet - <http://www.uel.br/grupo-pesquisa/redesconhecimento/fontes>
- Grupo Interdisciplinar de Pesquisa em Análise em Redes Sociais, GIAR - Estuda as diversas metodologias que visam produzir e analisar dados sobre as redes sociais, conforme definição do próprio site - <http://www.giars.ufmg.br>.

- Grupo de Pesquisa Redes, Ambientes Imersivos e Linguagens - GPral - Estuda o comportamento das pessoas usuárias das redes sociais e as consequências geradas com seu uso como em termos de mudanças nos hábitos, criatividade e aprendizagem - <http://www.ufjf.br/redeslinguagens/linhas-de-pesquisa/ambientes-imersivos-colaborativos-redes-sociais-e-semiotica/>.

Tabela 2.2: Pesquisas recentes realizadas no Brasil .

TRABALHO COM MENSAGENS DE REDES SOCIAIS	EM PORTUGÊS	EM INGLÊS
Análise de sentimento de tweets com foco em notícias [50]	x	
Análise de Sentimentos e mineração de <i>Links</i> [51]		x
Comparação e combinação de métodos de análise de sentimentos [48]		x
Análise de rede de ingredientes e receitas [78]	x	
Detecção de spam na rede Social Apontador [71]	x	
Uso do Twitter para pesquisa de opinião em eleições municipais[128]	x	
Sistema Cognos - Busca por especialistas de tópicos[70]		x
Identificando formadores de opiniões em redes sociais[63]		x
Panas-t - medidor de humor com base em escala psicométrica [32]		x
Navegação e interação de usuários em redes sociais[26]		x
Fiscalização de epidemia de dengue [69]	x	
Redes sociais multipolarizadas [93]	x	
Modelagem e caracterização de redes científicas[100]	x	
Identificação de áreas de atuação de pesquisadores[99]	x	
Emoticons e sentimento coletivo [129]		x
Predição de relacionamentos em redes[101]		x
Análise de sentimento de tweets sobre protestos no Brasil[130]	x	
Padrões de relacionamento em comunidade científica[102]		x
Análise de opiniões em comentários de notícias sobre eleição [131]	x	
Análise de sentimento e influência das palavras [132]	x	
Análise de sentimento em tweets sobre eleição para reitor de universidade[125]	x	
Tradução português-ínglês para análise de sentimentos[133]	x	
Análise de sentimento [52]		x

Além dos grupos citados, existem também vários professores espalhados nas universidades brasileiras desenvolvendo projetos em linhas de pesquisa voltadas para o estudo de redes sociais, porém não divulgados, ou com perfis desatualizados na Internet. Na Tabela 2.2 estão expostas algumas referências de trabalhos, em sua maioria comentados anteriormente nesse Capítulo, que foram realizados por pesquisadores brasileiros. Essa Tabela discrimina os trabalhos que utilizaram dados coletados das redes sociais no idioma português e inglês.

A Tabela 2.2 reflete um dos maiores desafios de quem trabalha na área no Brasil, o idioma. Vários trabalhos realizados por brasileiros envolvem pesquisa em caracterização de redes sociais, explorando aspectos de conexão entre usuários, geração e estudo de grafos que representam a interação entre indivíduos, mapeamento de perfis com grande influência na rede, volume de busca,

busca e contagem de palavras, modelagem de redes científicas, etc.. Mas, em se tratando de processamento natural de linguagem para descoberta de sentimento, os trabalhos existentes, ou realizam tradução de conteúdo para o inglês ou realizam uma análise superficial dos dados.

Além do idioma, o Brasil, por ser um país em desenvolvimento e possuir um vasto espaço geográfico, possui áreas nas quais a Internet funciona precariamente e nem todos os cidadãos possuem acesso de qualidade à tecnologia [2], Figura 2.2. Isso acarreta às pesquisas um resultado que reflete apenas parte da população e, geralmente, uma parte que está aglomerada em uma localidade geográfica na qual a tecnologia funciona.

Apesar de as pesquisas esbarrarem nesses grandes desafios, elas possuem um vasto caminho pela frente, pois, sendo o Brasil um país em desenvolvimento, seu envolvimento com tecnologia está em constante evolução, abrindo assim portas para descobertas em vários domínios ainda não explorados por aqui.

Capítulo 3

Coleta de Dados e Ferramentas

3.1 Introdução

Nesse capítulo, serão apresentados o método utilizado para a realização de coleta de dados do Twitter, a arquitetura de sistema, as ferramentas utilizadas e a metodologia de extração de informação dos dados.

3.2 Arquitetura do Sistema

Com o objetivo de verificar se o uso de dados do Twitter podem ajudar a melhorar previsões para o mercado de ações brasileiro, esse trabalho concentra na integração de várias técnicas. Para um melhor entendimento do que foi realizado sobre os dados e as técnicas utilizadas, um desenho de arquitetura de sistema é proposto na Figura 3.1. Nessa é possível ter uma visão geral das transformações pelas quais os dados passaram.

O framework de estudo proposto passa por quatro estágios. O primeiro lida com a coleta, armazenamento e pré-processamento dos dados, o segundo busca obter através dos dados indicadores que possam ser utilizados no próximo e no último estágio. O terceiro trata de uma análise estatística do relacionamento entre dados da bolsa e indicadores obtidos dos dados do Twitter. O quarto e último estágio apresenta o desenvolvimento de um sistema simulador de compra e venda de ações para auxílio na tomada de decisão de atuação no mercado de ações. Cada detalhe e característica dessas fases serão discriminados nas seções e capítulos seguintes.

3.3 Coleta

O objetivo da fase de coleta é obter dados diretamente do Twitter para armazenamento. Os dados são capturados do fluxo contínuo de mensagens publicadas no Twitter e armazenados em tabelas no banco de dados.



Figura 3.1: Arquitetura do Sistema.

3.3.1 Twitter

Criado em 2006, o Twitter é uma plataforma de microblog *online* que permite a criação de redes sociais através das conexões entre seus usuários. Esses publicam mensagens chamadas de tweets que podem conter até no máximo cento e quarenta caracteres que podem ser lidas por todos os seus seguidores. Um tweet pode conter várias características em seu texto, a Figura 3.2 exemplifica algumas e isso constitui um de seus maiores problemas, a falta de padrão nos dados [14].

O Twitter foi selecionado, dentre os demais sites de redes sociais online, por ser um serviço de microblog popular no Brasil, e está entre os cinco principais mercados em termos de usuários no mundo[134]. Outro fator que favoreceu a escolha por essa rede é a disponibilização de uma *Application Programming Interface* (API, do inglês) que permite a coleta de mensagens em tempo real. Entretanto, possui em seus termos de uso e serviço uma regra que proíbe a terceiros a redistribuição de conteúdo coletado sem a aprovação por escrito da companhia [135].

3.3.2 Domínio de dados

O domínio de dados escolhido para coleta nas redes sociais é um grupo de empresas brasileiras que possuem ações para negociação na bolsa de valores. Esse grupo de empresas foi selecionado levando em consideração alguns critérios como solidez, história e tempo de atuação no mercado brasileiro, participação com ações na bolsa de valores e que sejam alvos de comentários em *sites* de notícias, finanças, economia, investimentos e em blogs.

Uma pequena descrição sobre as companhias escolhidas para a atividade empírica deste trabalho está disponibilizada na Tabela 3.1. Todas, com exceção do Grupo X, são empresas com mais de cinquenta anos no mercado brasileiro. O Grupo X foi selecionado, apesar de sua recente fundação, pelo burburinho existente ao redor das perspectivas de sucesso apontadas por seus investidores e pelo dono da empresa Eike Batista em 2012 e 2013. Entretanto, logo após o início da coleta, iniciou-se um processo de decadência de algumas empresas desse grupo, o que acarretou um crescente volume de comentários na rede a seu respeito.

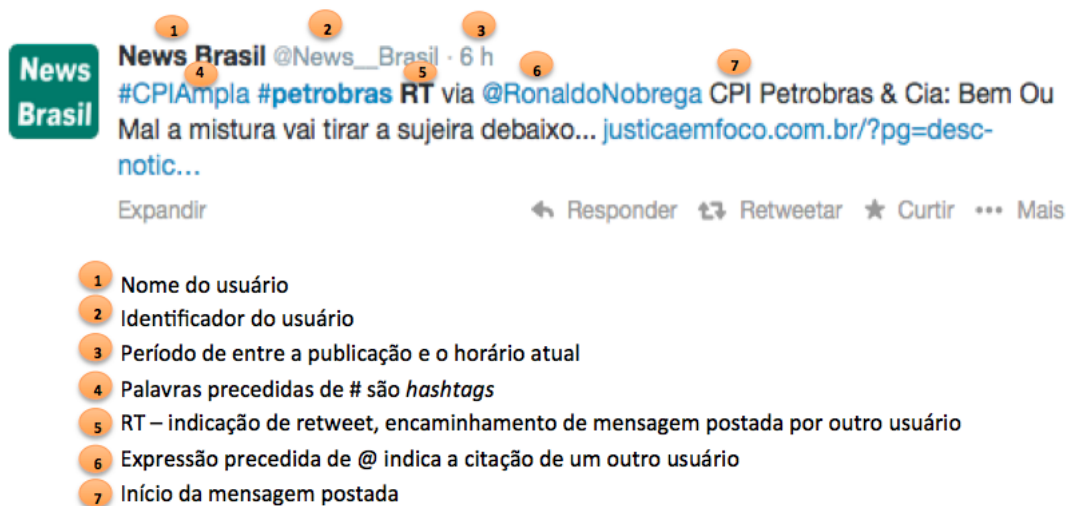


Figura 3.2: Exemplo de um tweet.

3.3.3 O Coletor

Após a escolha das companhias a serem utilizadas na pesquisa, foram selecionados os termos para filtragem da coleta. A captura de uma mensagem fica então atrelada à regra de esta possuir, obrigatoriamente, uma das palavras disposta na primeira e terceira coluna da Tabela 3.1. Tais palavras ou são os nomes das empresas, ou siglas que as representam, ou nomes dados as suas ações para negociação na bolsa de valores.

A atividade de coleta foi realizada através do desenvolvimento de um *script* Java que acessa o Web Service do Twitter e captura mensagens do fluxo contínuo que se adequam aos termos escolhidos para filtro. Essa atividade foi estabelecida no dia 13 de agosto de 2013.

3.4 Persistência

Para armazenar os dados, foram criadas duas tabelas - "tweets" e "usuários" - em um banco PostgreSQL¹. Este é um sistema gerenciador de banco de dados objeto-relacional (SGBD), uma ferramenta *open source* que permite ser utilizada, modificada e distribuída por qualquer pessoa gratuitamente e para qualquer finalidade. Ele suporta grande parte dos padrões SQL² e tem sido utilizada em diferentes pesquisas e aplicações de grandes empresas internacionais, órgãos governamentais de vários países e universidades para: sistema de análise de dados financeiros; monitoramento de desempenho de motores à jato; banco de dados de rastreamento de asteroides; armazenamento de informações médicas; vários sistemas de informação geográfica; e também como ferramenta de ensino [136].

¹PostgreSQL é um sistema gerenciador de banco de dados *open source* - www.postgresql.org.

²*Structured Query Language* - Linguagem de Consulta Estruturada.

Tabela 3.1: Empresas rastreadas no Twitter.

Empresa	Comentário	Ações
Gerdau	Uma das maiores e mais tradicionais empresas siderúrgicas brasileiras. Fundada em 1901, é líder no segmento de aço das Américas e maior reciclador da América Latina.	GGBR3 GGBR4
Vale S.A.	Criada em 1942, no governo de Getúlio Vargas, é hoje uma empresa privada, de capital aberto e uma das maiores mineradoras do mundo. Produz minério de ferro, manganês, cobre, carvão, cobalto e outros.	VALE3 VALE5
Itaú Unibanco	Maior banco privado da América Latina, surgiu da união do Banco Itaú, Banco Itaú Holding e Unibanco em 2008.	ITUB3 ITUB4
Banco do Brasil - BB	Instituição financeira brasileira estatal. Criada em 1808, foi o primeiro banco em território do Império Português[137].	BBAS3
Grupo EBX - Grupo X	Conglomerado de seis empresas com atuação em vários setores. Seu fundador é o empresário brasileiro Eike Batista. Das empresas foram escolhidas para coleta: OGX - Óleo e Gás Participações S.A, fundada em 2007, atua nas áreas de exploração e produção de petróleo e gás natural. MMX - Mineração e Metálicos S.A., fundada em 2005, atua na área de mineração de minério de ferro. MPX Energia S.A. - fundada em 2007 é considerada a maior empresa privada na área de geração de energia do Brasil. Em 2007, mudou sua razão social para ENEVA.	OGXP3 MMXM3 MPXE3
Petrobrás	Empresa estatal instituída em 1953. Atua no segmento de energia, nas áreas de exploração, produção, refino, comercialização e transporte de petróleo e derivados.	PETR4 PETR5
Companhia Siderúrgica Nacional - CSN	É a maior empresa siderúrgica do Brasil e América Latina. Fundada em 1941, foi privatizada em 1993. Possui minas de minério de ferro e outros minerais. Destaca-se na produção de aço bruto e laminados.	CSNA3
Usiminas	Usiminas Empresa fundada em 1956. Destaca-se como líder na América Latina como produtora e comercializadora de aços planos e outros destinados aos setores de bens de capital e bens de consumo da linha branca e indústria automotiva.	USIM5

A tabela "tweets" contém dados dos tweets postados (mensagem, data de publicação, identificador do usuário, identificador do tweet, se é um retweet, quantidade de vezes que foi realizado o seu retweet, *software* de origem, latitude e longitude, cidade de origem, país de origem). O conteúdo

da tabela "usuários" é composta pelos dados de usuários (identificador, nome, texto descritivo de perfil, data de inscrição, idioma, localização). É importante salientar que alguns itens desta tabela só são preenchidos quando o usuário permite a publicação através de seu perfil. O Twitter concede ao usuário a possibilidade de escolha de níveis de permissão de publicação de conteúdo e dados de perfil, dessa forma, o usuário tem liberdade de definir se seus dados são totalmente ou parcialmente públicos.

Os experimentos foram realizados em duas fases, a primeira foi realizada com os dados coletados de 13 de agosto de 2013 a 19 de abril de 2014 e consiste na análise estatística apresentada na etapa III da arquitetura do sistema - Figura 3.1, essa será descrita no próximo capítulo. A segunda fase consiste na realização da etapa IV da Figura 3.1 e será detalhada no Capítulo 5. Para essa segunda fase foram utilizados dados coletados de 13 de agosto de 2013 a 04 de maio de 2015.

Durante o período de coleta explorado na primeira fase, foram armazenados dois milhões setenta e um mil novecentos e setenta e cinco tweets e dados de mais de meio milhão de usuários. Para melhor visualização e estudo dos dados, comandos SQL foram utilizados para separar os dados em tabelas por empresas e por ações. As Figuras 3.5 (a) e (b) apresentam, respectivamente, o volume de tweets coletados por empresas e por ações. Nelas, é possível observar algumas peculiaridades nos dados. As coletas obtidas para as empresas CSNA, Guerdau e Usiminas, comparadas às demais, resultaram em quantidade inexpressiva de dados. Contudo, Itaú e Petrobrás foram as que obtiveram maior quantidade de dados, mais de quinhentos mil tweets cada, Banco do Brasil e Grupo X obtiveram mais de cem mil e Vale mais de cinquenta mil.

Quando os dados são separados em mensagens que contém apenas o nome das ações das companhias, Figura 3.5 (b), observa-se uma redução drástica no volume. Nenhuma das ações possuem mais de dezoito mil tweets e as ações ITUB3, GGBR3 e MPXE3 obtiveram quantidades inexpressíveis de mensagens.

O que se pode entender, a partir de uma análise visual realizada sobre tal amostra dos dados e comparando os volumes de dados por empresas e por ações, Figuras 3.5 (a) e (b), é:

- muito se comenta sobre as empresas, mas em alguns casos, comenta-se quase nada sobre as ações disponibilizadas na bolsa de valores das mesmas;
- na Figura 3.5, o Itaú apresenta grande quantidade de dados, entretanto, na Figura 3.5 (b) seu tamanho reduz drasticamente. O Itaú é um banco que possui várias iniciativas relacionadas à cultura, oferece aos seus correntistas cinema, fornece apoio a shows de música, exposições e a uma vida sustentável, além disto, disponibiliza bicicletas para as pessoas utilizarem na cidade do Rio de Janeiro, possui muita publicidade televisiva expressiva, dentre outros. Durante a análise visual de comentários sobre o Itaú em tweets, é notória a quantidade de mensagens com conteúdo relacionado a tais iniciativas. Por isto, ao comparar as Figuras (a) e (b), percebe-se que há muito comentário sobre o banco, mas pouco sobre suas ações na bolsa;
- a Vale S.A. é outra empresa que apresenta uma quantidade significativa de dados na Figura 3.5(a) e quase nada em (b). Levando em consideração que a palavra "vale" pode ser usada como conjugação do verbo valer na terceira pessoa do singular, ou como substantivo mas-

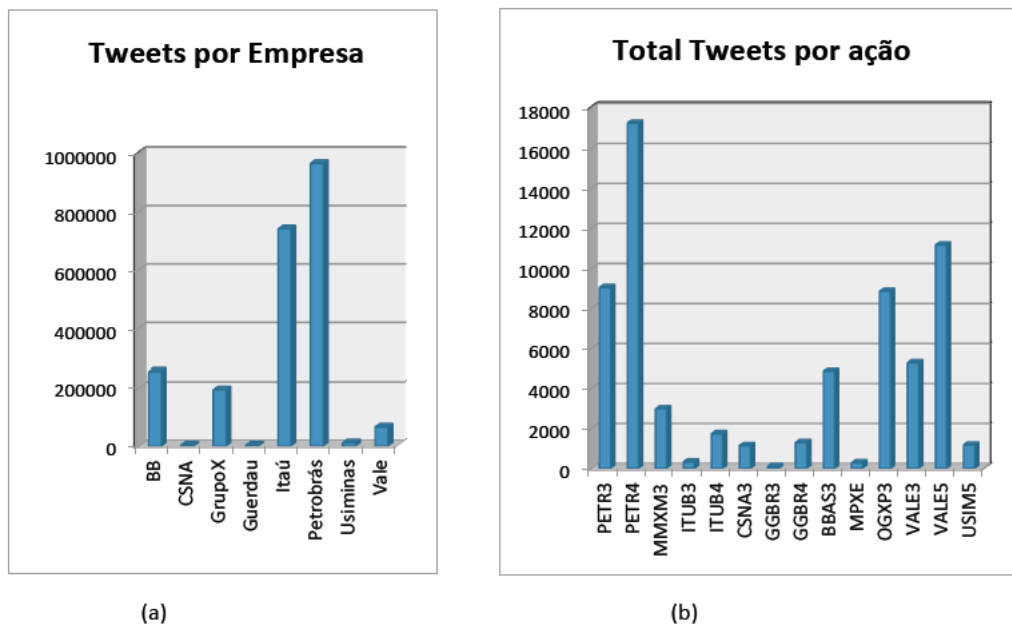


Figura 3.3: Total de Tweets coletados: (a) por empresas (b) por ações das empresas.

culino significando várzea ou planície à beira de um rio, trata-se de uma palavra bastante comum no vocabulário brasileiro. Muitas das mensagens contidas na tabela Vale não têm relação alguma com a empresa;

- com mais de dois mil tweets a respeito de suas ações, em oito meses de monitoramento, apenas as ações PETR3, PETR4, BBAS3, OGXP3, VALE3 e VALE5 se destacaram.

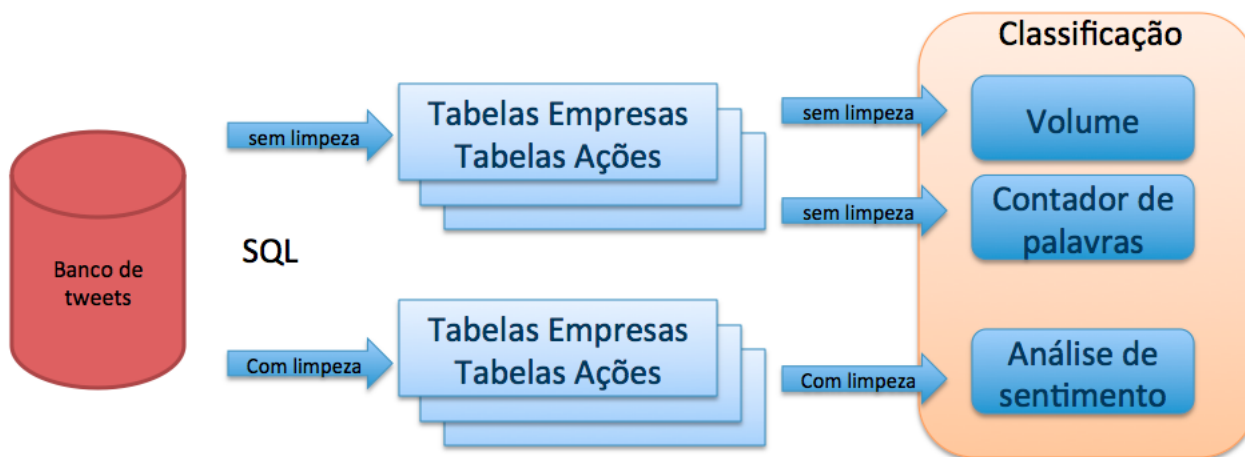


Figura 3.4: Conjuntos de dados e local onde serão utilizados na arquitetura do sistema.

3.5 Pré-processamento

Os tweets coletados na etapa anterior serão utilizados em três processos distintos na próxima fase que é a de classificação. A Figura 3.4 apresenta os dados em dois conjuntos, um "sem limpeza" (dados brutos) e o outro "com limpeza" (dados transformados).

Vários trabalhos tratam do uso de análise de sentimentos em textos de microblog para rastreamento de sentimento em tempo real e modelagem do humor público [10], descobre-se que muitas mensagens são desprovidas de sentimento [6], outras podem tratar de comentários subjetivos ou não a respeito de situações, várias delas contêm *links* para reportagens [35] e podem conter conteúdo não relacionado ao campo pesquisado, mesmo que possua palavras relacionadas a esse domínio.

Objetivando obter um maior aproveitamento de sistemas analisadores de sentimentos, cada trabalho adota uma sequência de atividades para preparar os dados coletados para a análise. Técnicas para remoção de *stop-words*³, pontuação e mensagens com "www" e "http:" em seu conteúdo (para eliminação de spam e informativos) podem ser adotadas, bem como agrupamento de tweets por data e seleção dos que continham palavras-chave os quais foram a escolha de [10]. Já [14], preferiu converter maiúsculas em minúsculas, remover stop-words e retweets e detectar qual o idioma foi utilizado. Em seu trabalho [4] concentrou em separar mensagens que continham palavras-chave do domínio por ele pesquisado.

Neste trabalho as transformações escolhidas para pré-processamento dos dados para a primeira fase de colera de dados foram:

- Filtragem de retweets;
- Filtragem de relevância, remoção de tweets que abrigaram palavras ou expressões selecionadas;
- Filtragem de *links*, remoção de tweets com "http:" e "www";
- Filtragem de pontuação.

As transformações acima foram selecionadas após a realização de análise visual de amostra com quatrocentos mil dados coletados. Com isso, teve-se a oportunidade de vivenciar a realidade dos dados apontados por tantos artigos publicados na área. As mensagens postadas por usuários de redes sociais são descompromissadas de quaisquer formalismos, isto é, possuem em seu conteúdo erros ortográficos em abundância, ausência de estrutura gramatical, palavras em idiomas diversos, caracteres que representam o estado atual de espírito do emissor (chamados emoticons), xingamentos, palavras chulas, entre outras características já citadas no capítulo anterior. Além disso, comprova-se uma enorme quantidade de retweets, *spam* e mensagens informativas. Para reduzir a quantidade de conteúdo desnecessário, o banco passou pelas transformações acima pontuadas.

Após a remoção de retweets, permaneceu no banco apenas a primeira mensagem postada e não as compartilhadas pelos amigos do primeiro usuário. Durante a análise visual dos dados, várias

³Palavras não significativas, podem ser pronomes, artigos, preposições, conjunções ou outras.

palavras e expressões muito utilizadas e sem relação com o domínio estudado foram selecionadas e removidas na filtragem "com limpeza", tais como xingamentos, avisos de eventos musicais, esportivos, textos em outros idiomas, avisos de incêndio, roubo, localização de usuários como, por exemplo, "estou ao lado do Itaú da Av. Raposo Tavares", etc.. A filtragem de pontuação não removeu mensagens do banco e sim caracteres de pontuação como ". , ; ! ?". Por último foi realizada a remoção de mensagens com *links*.

Para executar a filtragem de relevância removendo do banco as mensagens que continham as mais de trezentas palavras e expressões selecionadas, foi desenvolvido um *script* SQL. Uma amostra dessas expressões pode ser observada na Tabela 3.2.

As Figuras 3.5 (a) e (b) apresentam o volume de dados antes e depois do pré-processamento, com remoção de retweets, com remoção de retweets e filtragem e relevância, "http" e "www", separados por empresas e por ações.

Tabela 3.2: Palavras e expressões utilizadas para limpeza do banco de mensagens.

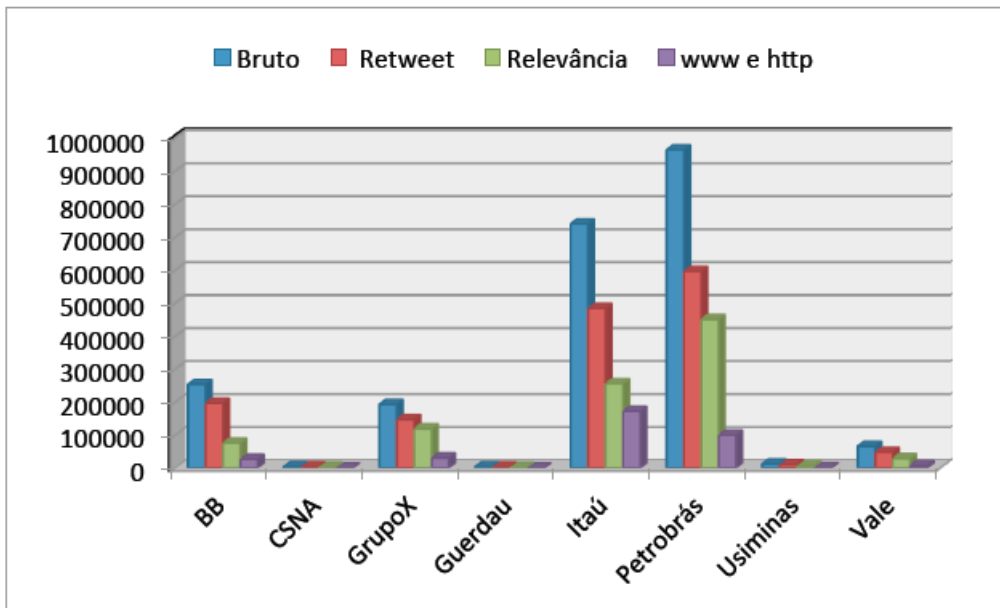
Expressões	bom pra todos, histórias que inspiram, cheque especial, cartão de crédito, São Caetano, I'am at, feito pra você, feito para você, Associação Atlética, em frente o, festival curtas, inscrição de estágio, meu tio, greve do banco, fundação banco, jogo do Brasil, código de barra, sua vida
Palavras	estágio, apostila, música, maracanã, campeonato, circuito, vasco, corinthias, FIFA, ingresso, bicicleta, brasileiro, amor, fotografia, natura, coca-cola, cinema, neymar, pulseira, ciclovia, bandido, cofrinho, ônibus, avenida, lotérica, gugu, escola, vereador, senha, lento, município, whatsapp

Ao verificar a diferença entre volume de dados com ou sem limpeza, algumas particularidades já citadas podem ser claramente observadas no domínio pesquisado:

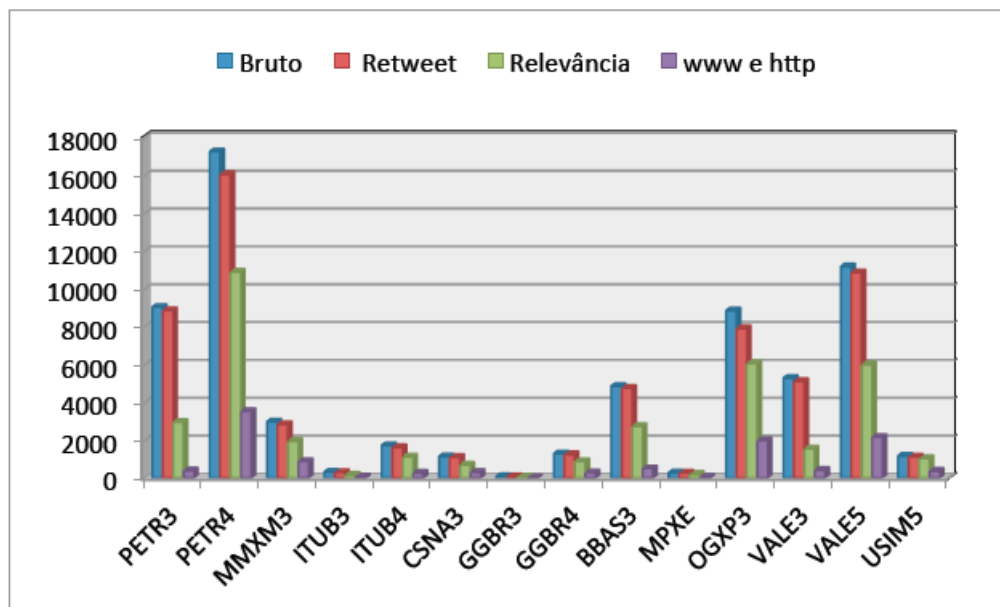
- grande parte das mensagens possuem *links*;
- grande quantidade de mensagens de retweet presente no banco;
- Banco do Brasil, Itaú e Petrobrás contém uma quantidade significativa de dados irrelevantes para a pesquisa;
- CSNA, Guerdau, Usiminas e Vale permaneceram praticamente as mesmas.

3.6 Classificação - Análise de tendência e sentimento

Neste estágio, procura-se extrair dos dados algum conhecimento que possa ser utilizado para estimação ou verificação de relacionamento com o mercado de ações. Essa informação será utilizada na próxima etapa, a de análise estatística do sistema - Figura 3.1.



(a)



(b)

Figura 3.5: Gráfico de volume de tweets coletados durante o período de 8 meses para cada empresa.

Para obter tendências a partir dos dados coletados, serão adotados dois processos independentes. Um levará em conta os dados brutos do banco, ou seja, dados "sem limpeza" e o outro os dados pré-processados descritos na Seção 3.5.

Sobre os dados "sem limpeza", serão explorados o volume de mensagens e os resultados de dois contadores ingênuos de palavras que serão descritos nas subseções seguintes. Nos dados pré-processados, isto é, o banco "com limpeza", será aplicado um analisador de sentimentos que polarizará os tweets em positivos e negativos e também será descrito abaixo.

Importante salientar que não faz parte deste projeto de doutorado a tarefa de desenvolver uma ferramenta para análise de sentimentos de documentos, para tal, foram estudadas várias ferramentas que realizam tal tarefa e dentre elas foi selecionada a que mais se adequava às necessidades da pesquisa. Tal ferramenta será comentada nesse texto mais adiante.

Os experimentos e resultados obtidos a serem apresentados neste relatório levaram em conta quatro empresas das selecionadas inicialmente. As escolhidas são aquelas cujas ações apresentaram maior quantidade de tweets após limpeza dos dados - Seção 3.5. Essas empresas e suas respectivas ações estão marcadas na Tabela 3.3, na qual as linhas apresentam coloração cinza.

3.6.1 Janela de análise

Apesar de os dados serem coletados no instante em que são publicados na plataforma Twitter, nesta pesquisa a análise é feita por dia de coleta e não por sequências de horas ou segundos coletados. Ou seja, as análises de tendências e sentimento da multidão realizada sobre os tweets obtêm o sentimento da multidão do dia t . Todos os tweets coletados no dia t contribuirão para afirmar um valor que representará a tendência ou o sentimento do dia t .

3.6.2 Volume

O volume de dados é bastante utilizado em estimação por pesquisadores da área de relacionamento de dados de mídias sociais *online* e mercado de ações. Baseando-se na ideia de que características textuais não estão limitadas apenas a sentimentos, [6] utilizou o fluxo não estruturado do tráfego web produzido pela comunidade *online* para investigar seu poder preditivo com relação ao mercado de ações. Através de correlações entre dados do fluxo *web* diário e preços de ações, ele desenvolveu modelos de predição e realizou experimentos que apresentaram evidências encorajadoras que justificam e incentivam investigações em definição de novos indicadores complementadores dos já existentes baseados em textos, especialmente os que usam análise de sentimentos.

Baseando seus estudos no volume de busca por nomes de ações e termos econômicos e financeiros no Google - dado amplamente utilizado por pesquisadores como representante do humor do público e investidor [106] - e no volume de tweets coletados diariamente para compará-los ao volume de negociação do mercado de ações, [10] sugere que o volume de buscas do Google é menos eficiente do que o volume obtido através do Twitter. Para [4], o volume de negociação é geralmente

correlacionado com a atenção do investidor. Em sua pesquisa, afirmou que o volume de mensagens coletadas do Twitter é a variável mais intimamente relacionada com a atenção do investidor.

Analisando a quantidade de tweets publicados com conteúdo relacionado aos filmes em determinados intervalos de tempo, [5] mostrou que existe forte correlação entre a atenção dada a um tópico, no caso, o burburinho sobre filmes nas redes sociais e sua classificação no futuro, ou seja, o resultado de bilheteria.

Os volumes das empresas escolhidas para o experimento deste trabalho a ser utilizado na próxima etapa do sistema - análise estatística - são os apresentados nas Tabelas 3.3 (a) e (b).

Tabela 3.3: Volume total de tweets coletados separados por (a) empresas e (b) por ações.

Empresas	Total Tweets	Ações	Total Tweets
Banco do Brasil	254.079	Petr3	9.057
C. Siderúrgica Nacional	3.966	Petr4	17.274
Grupo X	193.363	MMXM3	2.996
Guerdau	3.723	ITUB3	338
Itaú – Unibanco	740.268	ITUB4	1.750
Petrobrás	964.628	CSNA3	1.158
Usiminas	12.680	GGBR3	117
Vale S.A.	66.419	GGBR4	1.307
		BBAS3	4.876
		MPXE3	294
		OGXP3	8.865
		VALE3	5.303
		VALE5	11.186
		USIM5	1.190

(a)

(b)

Total Tweets :
2.071.975

Empresas e ações selecionadas para avaliação

3.6.3 Contador de palavras

O mercado de ações possui uma série de termos que qualificam o estado do mercado. Expressões como "em baixa", "descendo", "caindo", "padrão três corvos pretos", "padrão em golfo de baixa" e outras podem indicar uma tendência de queda nos valores das ações. Outras como "em alta", "subindo", "padrão três métodos de subida", "se levanta" podem indicar tendência de subida no preço. Muitas dessas expressões permanecem no vocabulário dos investidores e outras são geradas, assim como gírias, diariamente.

Esse trabalho explorará uma pequena parte do vocabulário dos investidores através de contadores ingênuos de palavras. Serão adotadas para experimentação duas abordagens de contagem diária:

1. Contador de expressões para alta e baixa: Nesse, para cada dia útil da bolsa, serão obtidos

dois contadores para cada ação. Um armazenará a quantidade de expressões de alta do dia e o outro a quantidade de expressões de baixa do dia. Esses contadores indicarão ingenuamente, a tendência de baixa ou de alta do dia.

2. Contador de compra, venda ou aluga: No mercado de ações, essas três palavras têm um significado bem definido e podem apontar tendências do mercado:

Comprar - De acordo com a Bovespa [138], um investidor opta por adquirir ações quando objetiva um ganho através dos direitos e proventos distribuídos aos acionistas pela companhia e também pela valorização do preço das ações.

Vender - Segundo [138], um investidor decide vender sua ação quanto, ao analisar suas perspectivas, verifica que essas estão menos favoráveis em relação a outras ações do mercado, ou quando necessita do dinheiro investido.

Alugar ou venda a descoberto - O aluguel de ações acontece quando um investidor loca a outro, por prazo determinado, certa quantidade de ações. O locador ganha um rendimento adicional com a operação e o locatário utiliza a ação para fazer a chamada venda a descoberto [139]. Essa prática permite a quem aluga a obtenção de lucros quando há queda no preço das ações, ele a vende para comprá-la com preço mais baixo ou para compor estratégias do investidor.

Dado o nome da ação, o contador de compra venda ou aluga percorre toda a tabela e conta para cada dia a quantidade de palavras que começam com "compra", "vend" e "alug".

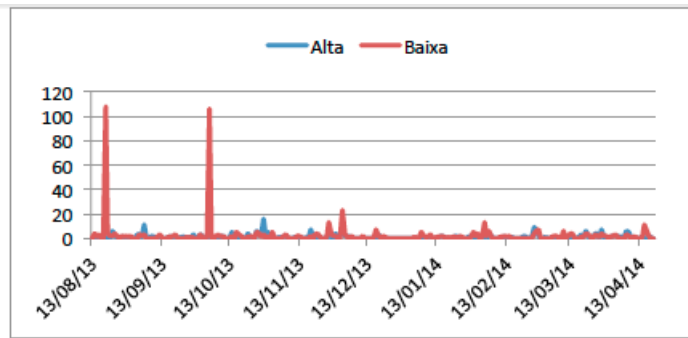
As Figuras 3.6 e 3.7 apresentam os valores obtidos para os contadores. A primeira mostra os valores de Alta e Baixa ao longo do tempo de coleta (de agosto de 2013 a abril de 2014). A segunda apresenta o valor dos contadores obtidos comprar, vender e alugar em duas janelas de tempo distintas para as ações selecionadas: PETR4, OGXP3, BBAS3, e VALE5. Essas janelas de tempo compreendem o período de duas semanas iniciando com a segunda-feira e terminando com a sexta-feira da segunda semana. Estão inclusos na figura o sábado e o domingo. As datas foram selecionadas por apresentarem uma grande quantidade de palavras captadas entre agosto de 2013 e abril de 2014.

3.6.4 Polarização de tweets

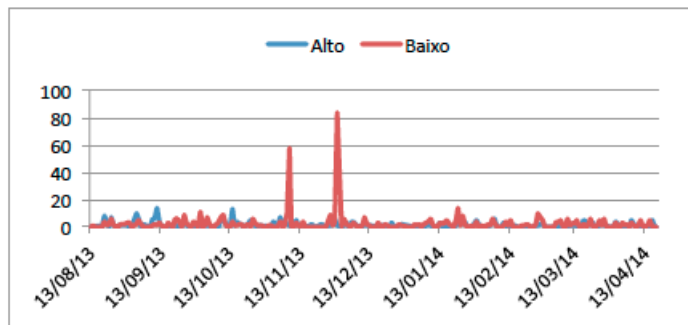
3.6.4.1 Análise de Sentimento

Segundo [44]: "Análise de sentimento, também chamada de mineração de opinião, é um campo de estudo que analisa a opinião das pessoas, sentimentos, avaliações, apreciações, atitudes e emoções para entidades como produtos, serviços, organizações, indivíduos, questões, eventos, tópicos e seus atributos. (...). Análise de sentimentos e mineração de opinião concentra-se principalmente em opiniões que expressam ou implicam em sentimentos positivos ou negativos."

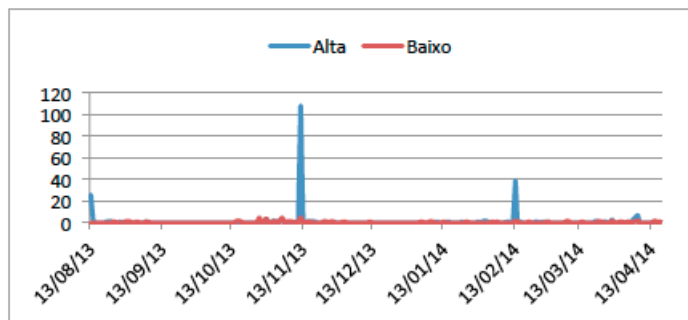
Para [49], a análise de sentimentos consiste em classificar a polaridade da opinião contida em um texto em positiva, negativa ou neutra. No campo da análise de sentimentos, três níveis principais têm sido investigados:



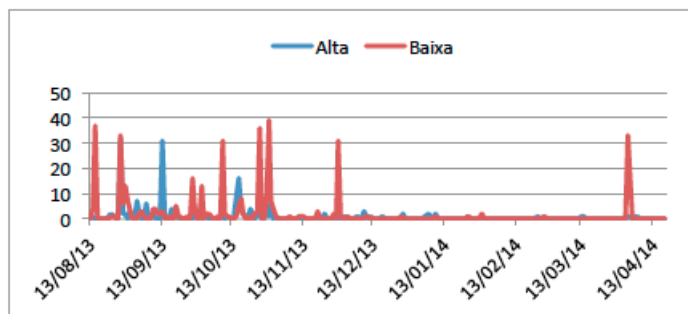
(a)



(b)



(c)



(d)

Figura 3.6: Quantidade de palavras relacionadas à alta e baixa durante os oito meses de captação de tweets sendo (a)PETR4, (b)VALE5, (c)BBAS3 e (d)OGXP3.

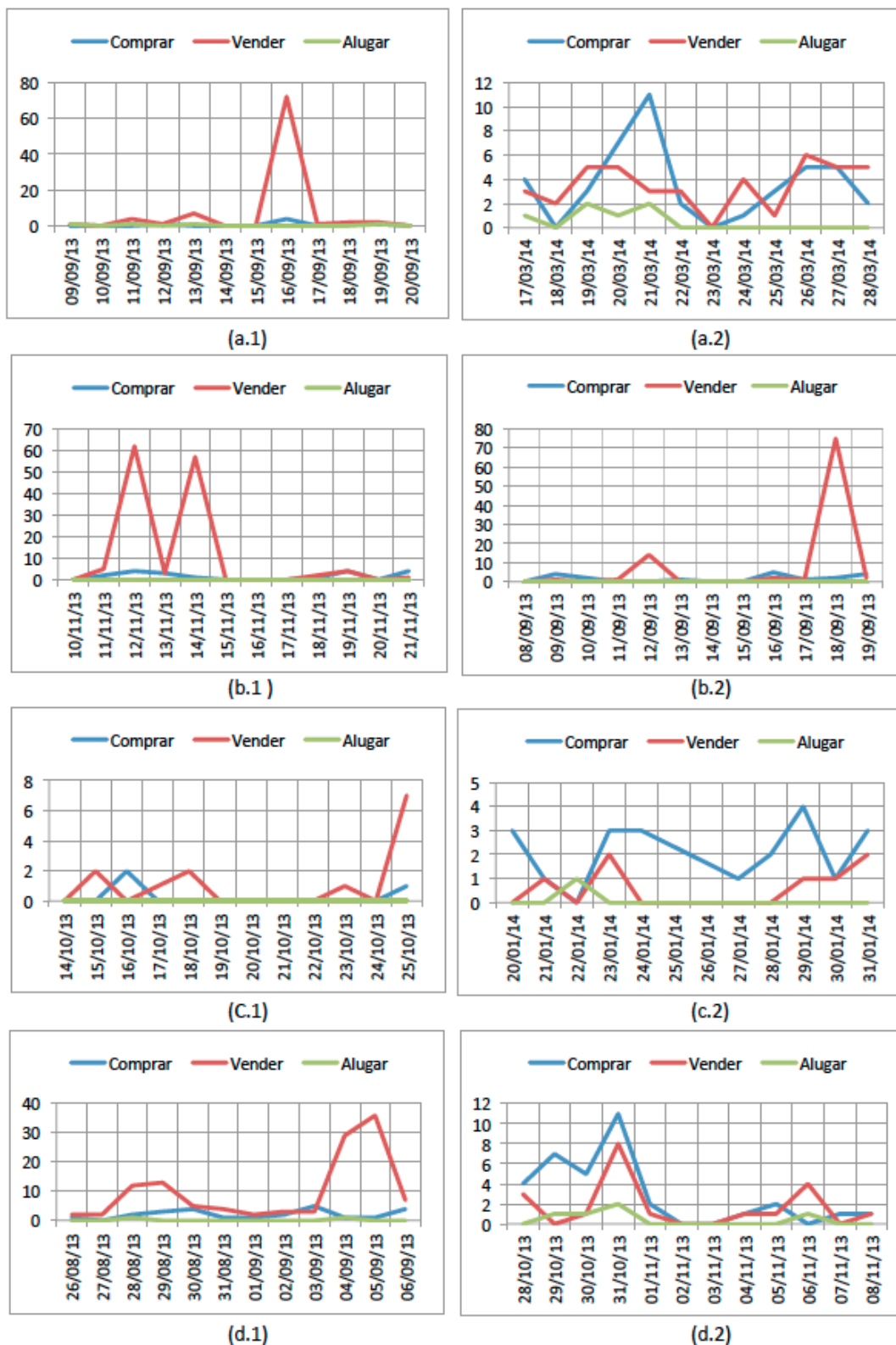


Figura 3.7: Quantidade de palavras com iniciais "compra", "vend" e "alug" por dia para duas janelas de tempo de duas semanas para (a)PETR4, (b)VALE5, (c)BBAS3 e (d)OGXP3.

- Análise em nível de documento: Envolve a classificação de um documento como um todo em positivo ou negativo;
- Análise em nível de sentença: Capta as sentenças do documento e determina a polaridade da opinião, positivo ou negativo, para cada uma;
- Análise em nível de entidade e aspectos: É uma análise de granulação mais fina, não preocupada com o documento ou sentença, mas busca alvos e a opinião sobre esses. Exemplo retirado de [44] na frase: "A qualidade da ligação do iPhone é boa, mas a vida útil da bateria é curta" existe uma entidade, o iPhone, e dois aspectos, cada um com um sentimento, qualidade de ligação que é positivo e vida útil da bateria, que é negativo. Neste caso, a qualidade da ligação e a vida útil da bateria são os alvos de opinião.

Um sistema de análise consiste em receber um corpus de documentos do tipo texto, aplicar pré-processamento usando uma variedade de ferramentas, como feito na Seção 3.5, e, no componente principal, adotar recursos linguísticos para anotar os documentos com rótulos de sentimento. As anotações podem ser feitas para um documento inteiro, quando a análise acontecer em nível de documento, para uma sentença, quando em nível de sentença ou para aspectos específicos de entidades, quando a análise for baseada em aspectos [31].

Um fator importante para o componente principal de um analisador é o conjunto de palavras que expressam o sentimento positivo ou negativo. Palavras como bom, maravilhoso e surpreendente demonstram positividade, enquanto que ruim, pobre e terrível expõem negatividade. Não só palavras, mas também expressões idiomáticas, podem carregar sentimentos e o conjunto dessas são instrumentais para a análise de sentimentos. A esse conjunto de palavras e expressões dá-se o nome de lexicon de sentimentos.

Muitas pesquisas têm sido realizadas em algoritmos para a formação de lexicons. Embora sejam de grande valor para a análise, são insuficientes [44]. Algumas questões podem ser pontuadas:

- Uma palavra polarizada como positiva pode ter uma orientação oposta em um domínio diferente. Exemplo: "Esse refrigerante está gelado!", é positivo, mas "Essa pizza está gelada!" é negativo;
- Uma sentença pode conter palavras polarizadas com sentimentos, entretanto, pode expressar sentimento algum. Exemplo: "Aquele restaurante é bom?".
- Sentenças com conteúdo sarcástico são problemáticas. Exemplo: "Que beleza de TV, pifou na primeira vez que foi utilizada!";
- Sentenças sem palavras sentimentais podem relatar opiniões. Exemplo: "Essa lavadora gasta muita água!".

Essas são apenas algumas questões relacionadas com a análise de sentimentos em textos, outras relacionadas à desambiguação e manipulação de negação remetem a um problema de processamento natural de linguagem e demonstram quão grandes são os desafios dessa área.

Textos podem ser objetivos ou subjetivos; os primeiros quando tratam de informações factuais, os seguintes quando tratam de opiniões, crenças e pontos de vista sobre entidades específicas. Existem duas abordagens principais para análise de sentimentos de documentos:

1. Supervisionada: Assume-se que existe um conjunto finito de classes no qual o documento deve ser classificado. Dados de treinamento estão disponíveis para cada classe. O caso mais simples consiste em determinar se o texto é positivo ou negativo. Extensões desse modelo adicionam a classe neutra, outros constituem uma escala numérica discreta na qual o documento pode ser classificado. Fornecidos os dados de treinamento, o sistema aprende um modelo de classificação usando um dos algoritmos Support Vector Machine (SMV, do inglês), Naïve Bayes, Logistic regression ou K-nearest neighbors (KNN, do inglês) [31]. Essa classificação é usada para rotular novos documentos em uma das classes de sentimentos.
2. Não-supervisionada: Baseada em determinar a orientação semântica de frases específicas do documento. Se a média dessa orientação estiver acima de um limiar pré-definido, então o documento é classificado como positivo, caso contrário, negativo. Maiores informações a respeito dessa abordagem podem ser obtidas em [44].

Nesse trabalho, será adotada a abordagem supervisionada para análise de sentimentos. A descrição da ferramenta adotada será realizada na próxima subseção.

3.6.4.2 Ferramenta para análise de sentimento

O objetivo deste trabalho, ao utilizar uma ferramenta para análise de sentimentos, é tentar extrair o sentimento da multidão advindo das postagens na rede social Twitter em português, mesmo sendo essas cheias de ruídos. Por isso, optou-se por utilizar uma ferramenta de análise de sentimentos baseada em treinamento e teste de N-grama.

O N-grama é um modelo de linguagem que permite captar palavras em sequência através de junção, ou seja, o N representa a quantidade de palavras unidas para a formação de um atributo.

Algoritmos que trabalham com treinamento e teste, ao utilizarem N-grama, admitem a utilização de um corpus de treinamento que ensinará ao classificador quais sequências de palavras estão associadas a uma determinada categoria de classificação [140].

Existem várias abordagens para a implementação de algoritmos para análise de sentimento de textos. Em [49], há uma ampla descrição e comparação de diversas dessas abordagens. Há também ferramentas prontas e disponíveis para a realização de análise de sentimentos, em [48], os autores comparam algumas delas.

Apesar de existirem alguns métodos de análise de sentimentos e de popularidade, na literatura não fica claro qual é o melhor de todos [48], e para tentar amenizar essa questão vários autores estudam e fazem comparativos sobre ferramentas buscando entender suas limitações, vantagens e desvantagens em análise de conteúdo de mensagens.

Tabela 3.4: Ferramentas para análise de sentimentos em textos.

FERRAMENTAS	DESCRIÇÃO	IDIOMA
SentiWordNet	Baseada no dicionário léxico de inglês WordNet que agrupa várias classes gramaticais, classifica o sentimento do texto em positivo, negativo e neutro. Adota como método o aprendizado de máquina semi-supervisionado. Disponível em http://sentiwordnet.isti.cnr.it	Inglês e línguas indianas(Bengali, Hindi e Telugu)[141].
SenticNet	Ferramenta disponível para análise de sentimento a nível conceitual. Usa processamento natural de linguagem para inferir a polaridade de conceitos de senso comum, explorando a semântica dos textos analisados. Disponível em http://sentic.net/about/ .	Inglês
LIWC	Linguistic Inquiry and Word Count - ferramenta comercial que analisa componentes estruturais, cognitivos e emocionais de um texto baseando-se em dicionário. Classifica o texto em positivo, negativo e em outras categorias. Disponível em http://www.liwc.net/ .	Inglês, alemão, espanhol, italiano e holandês.
SentiStrength	Analisador de sentimentos que estima o quão positivo ou negativo são os pequenos textos analisados. Segundo [48], implementa o estado da arte em aprendizado de máquina no contexto de redes sociais <i>online</i> . Sua classificação se apoia em palavras do dicionário LIWC. Disponível em http://sentistrength.wlv.ac.uk/ .	Inglês
SASA	SailAil Sentiment Analyser é uma ferramenta open source baseada em aprendizado de máquina para análise de sentimentos. Testado classifica mensagens do Twitter em positivo, negativo e neutro nas eleições presidenciais dos EUA, de 2012, está disponível em http://code.google.com/p/sasa-tool/ .	Inglês
Python NLTK	Baseado em um classificador ingênuo de, a linguagem de programação Python oferece um conjunto de ferramentas para processamento de linguagem natural. O classificador de sentimentos foi treinado com tweets e comentários sobre filmes. Disponível em http://text-processing.com/demo/sentiment/ .	Inglês, holandês e Francês.
Lingpipe	Conjunto de ferramentas para processamento de texto. Disponível em http://alias-i.com/lingpipe/index.html onde há também descrições sobre seu uso. Essa foi a escolhida para uso nesta pesquisa.	Multilingual
R -Text Minig Package	Trata-se de um pacote de análise de sentimento para a linguagem R de mineração de texto Disponível em https://r-forge.r-project.org/R/?group_id=1048 .	Inglês
Emoticons	Detecção de sentimentos através de símbolos que representam como o autor de um texto está se sentindo no momento de confecção do mesmo, por exemplo, " :-) " que significa feliz e consequentemente expressa um sentimento positivo.	Qualquer
Gate	Ferramenta que utiliza aprendizado de máquina supervisionado treinado com dados anotados por humanos, estatísticas de coocorrência e léxicos com palavras negativas e positivas para identificar problemas com produtos e serviços de empresas relatados em blogs. Disponível em https://gate.ac.uk/sentiment/ .	Inglês, francês, alemão, espanhol e símbolos.
Outras	Sentiment140(http://www.sentiment140.com/), Twends (http://twendz.waggenereidstrom.com), Twitratr (http://twitratr.com), SocialMention (http://socialmention.com), TipTop (http://feeltiptop.com/) e TweetFeel (www.tweetfeel.com).	Maioria em inglês.

Não é objeto de estudo deste trabalho de doutorado o desenvolvimento de ferramenta para análise de sentimento de tweets. Dessa forma, fez-se um levantamento de várias ferramentas existentes para tal atividade no intuito de encontrar uma que se adequasse ao problema proposto e que, principalmente, permitisse a análise de sentimento de textos na língua portuguesa. A Tabela 3.4 apresenta algumas das ferramentas pesquisadas e consultadas. Em [48, 49, 31], há comparativos de métodos e ferramentas para análise de sentimentos em textos.

Com a finalidade de obter a polarização dos tweets coletados, após estudo e análise de abordagens e ferramentas prontas, optou-se por adotar a ferramenta LingPipe. Tal escolha se deu por essa atender as necessidades da pesquisa com relação ao treinamento de sistema para um domínio específico, suporte à língua portuguesa e ser livre para uso.

A biblioteca LingPipe⁴ oferece um conjunto de ferramentas para atividade de processamento de texto e é o instrumento selecionado para a análise de sentimentos realizada neste trabalho. O analisador do LingPipe atua em nível de sentença e possui um modelo de classificador probabilístico de aprendizado de máquina baseado em regressão logística [142]. Essa ferramenta foi escolhida por ser estável, permitir o uso de N-grama, ser adotada em diversas pesquisas⁵, por, dentre tantas ferramentas disponíveis, se adequar bem ao problema em questão e permitir o treinamento para dados em qualquer idioma. Para os trabalhos com análise de sentimentos, elegeu-se as tabelas de tweets - a PETR4 e VALE5 (armazenam os tweets sobre a ação da Petrobrás e Vale) por possuir a maior quantidade diária de dados após a limpeza descrita na Seção 3.5.

O LingPipe trabalha com um conjunto de documentos de treinamento e teste. Esse conjunto necessita ser inicialmente polarizado de forma a mostrar ao classificador quais dos documentos caracterizam uma opinião positiva e negativa. Desse conjunto, seleciona-se uma parte para teste que, ao ser submetida ao modelo treinado, é calculada a probabilidade de cada um se encaixar em uma das classificações.

Nesse trabalho, cada tweet é adotado como um documento para o LingPipe. Com o intuito de polarizar um conjunto de treinamento, foi desenvolvida uma ferramenta em JAVA que carrega os tweets postados por dia para que o pesquisador os classifique. Dessa forma, são gerados dois arquivos; um com documentos polarizados como positivos, e outro os negativos. A Figura 3.8 apresenta uma janela do *software* desenvolvida para ler os tweets do banco de dados e disponibilizá-los para a polarização manual.

Após uma visualização rápida dos documentos pelo pesquisador, percebeu-se certa dificuldade em realizar a polarização, pelo fato de os dados possuírem um conteúdo atrelado ao domínio do mercado de ações. Optou-se, então, por utilizar a ajuda de um especialista para essa tarefa.

Durante a atividade de polarização dos dados de treinamento junto ao especialista, muitos questionamentos e características do perfil de usuários do twitter para esse domínio foram levantados. Alguns pontos julgados importantes:

⁴<http://alias-i.com/lingpipe/>

⁵<http://alias-i.com/lingpipe/web/citations.html>

ID	Tweet	POS	NEG	GERADO ...	AVALIAR...
438670108869677056	Venda coberta PETR4 (últ. fech.): PETRC14 (Tx. 3,47%, prot. 4,58%); PETRC15 (Tx. 7,68%, prot. 1,76%); PETRC16... http://t.co/ebFdyNSUrf	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
438670589893804033	Resultado 'maquiado' da Petrobrás #Petr4?! Hummm o mercado já entendeu assim. Estão batendo no ativo.	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438671093017350144	depois do balanço de ontem, #PETR3 e #PETR4 começando a manhã levando fumo! Vamos girar pelo mercado. Bom dia.	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438673614121238528	#PETR4 dívida da Petrobrás subiu só 30% no último ano. Se continuar nesta toada, o Brasil vai parar dentro do... http://t.co/kj9CZYLWIS	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
43867498656203008	#PETR4 dívida da Petrobrás subiu só 30% em 2013. Deste jeito, o Brasil vai parar dentro do Pré-cal aha... uhu... a dívida da Petro é nossa!	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438675824146120704	E o mercado recebeu mal o lucro merreca, dívida fabulosa - água no óleo da bolivariana Petrobras. #PETR4 #ForaDilma #PTnuncaMais	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438676859224862720	Não é Petrobrax, é #PETRODVIDAS! #Petrobrás #PETR4 #Bovespa	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
43867818641302912	#PETR4 ... e já chegamos no suporte dos R\$13,88! O se furar, a próxima parada é nos R\$13,56 ... Os PTralhas acabando com o Brasil..	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438684763189415936	@PedroCerize BTG Pactual maior comprador do dia, com saldo de 675 mil PETR4, correspondendo a 16,10% do saldo comprador total	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438695068854419456	@Infomoney parece que o mercado não curtiu a graça da Graça. #PETR4	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438698218428563456	#PETR4 RT @radaronline: Aumenta (e muito) o endividamento da Petrobras http://t.co/BUdbyVX1Vd	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438701161877889024	Será que PETR4 atingiu o "ponto Tiiririca", pior que tá não fica?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438701592121200640	"@topBovespa: #PETR4 RT @radaronline: Aumenta (e muito) o endividamento da Petrobras http://t.co/4IYMXAmx6"	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438701663298543616	A minha opinião sobre o resultado da Petrobrás: #PETR4: https://t.co/Jg6hrjEebG	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438708418250285057	Petr4 aos 14.00 com fundo em 13,56. e topo em 15,10.	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
43872967752955808	Se petr4 perder R\$13,56 min 2008, pode buscar os R\$12,30. PTralhas quebrando tudo...	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438732391700787201	Pessoal tá quieto na #PETR4 não acham? Antes era festa qdo caiu, festa qdo subiu, agora virou assunto do passado? :)	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438738324736315392	Hoje a Petr4 vai a R\$13,75, representando 48% do valor de liquidação da Petrobras. Em outras palavras, se... http://t.co/TWAs5fuZXR	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438738488192942081	Ativo c/ vol Financeiro Superior à sua MM21 -15h: ABCB4 ALL3 ALUP11 BPHA3 BRIN3 CCR03 CPLE5 CYR83 EMBR3 EURFU7 HGTX3 KEPL3 LAME3 ODP...	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438743205954486272	Fi só a CAVEIRA SINTETICA Graça Foster sair da tumba e abrir a mandíbula pras ações PETR4 desabarem! Saiba por que: http://t.co/IRBgRfCeG	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438744048598540289	VALLUREC : Vallourec-2014 será stable ou en "croissance modérée" http://t.co/NqWZHI8nr \$PETR4	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438752595658964992	Ativo c/ vol Financeiro Superior à sua MM21-16h: ITSA4 KEPL3 LAME3 LEV3 LUIS3 ODPV3 PETR3 PETR4 PFRM3 PSSA3 RADL3 RDN3 SMT03 TAE11 T...	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438755045371547648	#PETR3 #PETR4 - Lucro da Petrobras sobe 11% em 2013 e soma R\$ 23,57 bilhões - http://t.co/hQI6h0IPbu	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438766707189248001	BOVESPA -0,30% -141,60 pontos Graças à #PETR4 - 3,24% R\$ 13,72 - No mínimo deveria ser R\$ 71,00...	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438769458304544768	Petrobrás (PETR4) fechou hj em sua menor cotação desde outubro/2005. A diferença é que, naquela época, sua dívida não era de R\$ 221 bi.	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438770151711068161	26/02/14 - 17:19: Maiores Baixas: PETR4 -3,24% R\$13,72; PETR3 -2,48% R\$12,95; RSID3 -2,38% R\$1,64; EVEN3 -2,34% R\$7,08; ELP4 -1,98% ...	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438778654987145216	Muita gente comprou #PETR4 com aquele martelo semanal uma semana antes dele ser confirmado. Agora na ALL os mesmos não compraram né...	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438779137423978496	Correção: Carvalho defende patrocínio ao MST. A Petrobras patrocinou, com R\$ 650 mil - http://t.co/b7QpKWhKEM \ \ #PETR4 R\$13,72 - 3,31%? \ \	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438781231912980480	petr4 -3,39%, apesar do propalado lucro recorde, que veio acompanhado de uma dívida recorde.	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438784012178706432	Acabei de atualizar as análises de BVMF3, ITUB4, PDGR3, PETR4, USIMS e VALES no INVISTA EM AÇÕES http://t.co/LNFxOOww7g	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438782218341994496	Análise - 26.02.14 - IBOV, PETR4, VALES, EMBR3, DIRR3, RAPT4, TIMP3, GRND3, BBAS3, GGBR4, BRIN3 e JSLG3. Assista! http://t.co/LIq9tuztj	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438784015731286017	Análise de Vale5 e Petr4 para 27/02/2014.	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438784426399772672	Análise de Vale5 e Petr4 para 27/02/2014. http://t.co/EoZwZkdqkq	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438784820177825792	@petrobras @germano_mergel Análise por este caminho quem quer banana?hj, petr4 fechou 13,68 mínima desde 2005 http://t.co/wXr25pi8LT	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438785197979750400	#PETR4 #Petrobrás #GovernoCorruTo #PTralhas parabéns pela boa gestão do governo destaque no quesito politica fiscal e controle orçamentário	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438786814405804033	Argentina honres Repsol deal will dispel investor doubts http://t.co/Gctuo6xLFCVX \$FPX \$SXP \$PETR4 \$YPFD	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438789186540552192	Petr4 no menor valor desde 2005. Isso pq a Desgraça Foster falou que o resultado de 2013 foi EXCELENTE! Que piranha.	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438791328441593856	PETR4 descendo a ladeira.	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438791931519565825	26/02/14 Obvespa fechou em queda de -0,25%, aos 46.599 pts. Destaques: PETR4 -3,39%, OIBR4 -2,98% e ELP4 -2,72%. http://t.co/j9yUOf2ipY	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438793100816039936	Maiores Baixas: PETR4 -3,53% PETR3 -2,86% ELP4 -2,60% EVEN3 -2,34% BRPR3 -1,81%.	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
43879738096603137	Gráfico diário #PETR4 http://t.co/UvIVR4Nm1l	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438816440914018304	não foi e derreteu...RT @Live_Trade: 14,64 divisor de águas intra do King Kong #petr4	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438654553115078656	PETR4 pelo Credit Suisse: sem pressa para comprar...	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438708418250285057	Petr4 aos 14.00 com fundo em 13,56. e topo em 15,10.	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438781231912980480	petr4 -3,39%, apesar do propalado lucro recorde, que veio acompanhado de uma dívida recorde.	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438789186540552192	PETR4 no menor valor desde 2005. Isso pq a Desgraça Foster falou que o resultado de 2013 foi EXCELENTE! Que piranha.	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
438791328441593856	PETR4 descendo a ladeira.	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Figura 3.8: Amostra de dados coletados do Twitter sendo polarizados manualmente como positivos, negativos e selecionados para avaliação.

- Muitas pessoas utilizam-se de palavras para expressarem suas opiniões sobre os acontecimentos no mercado, e isso fez com que as limpezas realizadas na Seção 3.5 fossem revistas.
- O linguajar e interlocuções adotadas traçam um perfil de usuário que, aparentemente, trata-se de pessoa física e que adota a postura de "grafista" ou "traydeiro", isto é, são pessoas que realizam operações em curto prazo chamadas de *daytrade* e, por isso, são especialistas em gráficos emitidos a cada segundo pela bolsa de valores. Percebe-se essa característica pelo uso constante de expressões que indicam um formato do gráfico e por estratégias de negociação expostas nas postagens.
- Os tweets são dependentes do contexto. Muitos fazem referência a outros comentários e a características visualizadas nos gráficos da bolsa de valores;
- Mesmo com a limpeza realizada nos dados, existem milhares de tweets de notícias publicadas, sendo estes fatos objetivos que não expressam o sentimento da multidão;
- Comprova-se a quantidade de ruído nas mensagens, essas são carregadas de erros ortográficos, truncagem de palavras, mistura de idiomas, uso de sarcasmo e ironia e caracteres especiais.

Seguindo orientações de bons resultados obtidos no uso de N-grama com $N=8$ [140], essa também foi adotada nesse trabalho.

Para o experimento, foram polarizados 504 tweets pelo especialista (utilizando dados com

limpeza de agosto a dezembro de 2013), desses 38% foram utilizados para treinamento e 62% para avaliação. Desses, a média de acerto foi de 62%.

Capítulo 4

Análise Estatística Inicial

4.1 Introdução

Neste capítulo, será apresentada uma análise estatística inicial que verifica o relacionamento entre variáveis obtidas através dos dados coletados do Twitter e os da Bolsa de Valores de São Paulo - Bovespa para as ações PETR4 da Petrobrás, VALE5 da Vale S.A., BBAS3 do Banco do Brasil e OGXP3 da Óleo e Gás Participações S.A.. Esta é a descrição da realização da etapa III da arquitetura mostrada na Figura 3.1.

A análise realizada permitirá estudar possíveis relações entre dados de microblog e o mercado de ações brasileiro com a finalidade de verificar se o burburinho produzido na rede social pode, de alguma forma, contribuir para uma modelagem mais eficaz da dinâmica do mercado de ações no Brasil.

Para essa análise, foram utilizados dados coletados do dia 13 de agosto de 2013 ao dia 19 de abril de 2014, sendo um total de oito meses, aproximadamente.

4.2 Amostras

Para efetuar os experimentos iniciais, selecionou-se para cada ação uma janela de tempo de 9 semanas. No período escolhido como amostra estão concentrados volumes expressivos de tweets entre agosto de 2013 e abril de 2014. Como a quantidade de tweets postados em finais de semana é pequena em relação aos dias úteis, optou-se por remover das amostras os sábados, domingos e feriados. As Figuras 4.1 (a), (b), (c) e (d) apresentam graficamente e respectivamente, dentro do montante de dados coletados do Twitter, o período selecionado para o ensaio para as ações PETR4, VALE5, BBAS3 e OGXP3.

4.2.1 Dados do Twitter

Da coleção de dados coletados a partir do twitter, é possível extrair uma variedade de indicadores, entretanto, para esse experimento foram adotados os seguintes:

- B_t - Está relacionado ao burburinho, ou seja, representa a quantidade de tweets postados no dia t comentando sobre determinada ação;
- H_t - Representa o humor do dia t , otimista para compra ou pessimista para a venda de ações. Como a quantidade de tweets relacionados com a palavra "alugar" foi inexpressivo, optou-se apenas para o uso dos contadores "Compra" e "Venda". O valor de H_t é obtido através do contador ingênuo descrito na Seção 3.6.3, sendo computado como $H_t = Compra_t - Venda_t$, ou seja, a quantidade de tweets otimistas subtraída dos pessimistas. Se o valor é positivo, então, o mercado está otimista, caso contrário, pessimista.
- E_t - Relacionado ao sentimento de tendência do mercado no dia t . É baseado no contador de expressões alto e baixo definido na Seção 3.6.3, sendo $E_t = Alta_t - Baixa_t$.
- S_t - Representa o sentimento obtido através do analisador de sentimentos descrito na Seção 3.6.4.2. Se o valor é positivo, então, a maioria dos tweets coletados no dia possui um sentimento positivo em relação à ação, senão, o sentimento é negativo para o dia t . Foi definido $S_t = Positivos_t - Negativos_t$. Como comentado na Subseção 3.6.4.2, esse valor será computado apenas para os dados coletados para as ações da Petrobrás - PETR4 e VALE5 da Vale S.A.

4.2.2 Dados da Bolsa de valores

As variáveis da bolsa de valores a serem consideradas para os experimentos são:

- v_t - Volume de negociação: Trata-se da quantidade de ações negociadas em um dia de negociação. Esse volume é geralmente correlacionado com a atenção do investidor, ou seja, é o dado que mais se aproxima com o interesse do investidor. Neste trabalho, assim como em [4], testou-se esse relacionamento para cada ação através da medição da regressão entre o número total de tweets B_{t-1} , e as tendências $H_{t-1}, E_{t-1}, S_{t-1}$ do dia anterior, e o volume de negociações de hoje v_t . A série temporal utilizada foi transformada para a escala logarítmica para análise;
- r_t - Retornos diários: Trata-se da porcentagem de mudança no valor do ativo. Há uma escassa evidência da predicabilidade de um retorno [4], entretanto, o retorno provê informação útil sobre a distribuição de probabilidade do preço do ativo. É calculado utilizando-se o valor do preço de fechamento ajustado para dividendos p_t , juros sobre o capital, desdobramentos e agrupamentos para o dia t subtraindo deste o preço do dia anterior p_{t-1} . A fórmula abaixo adotada é similar à utilizada por [10]. Outros autores também adotaram o logaritmo dos retornos como [14, 8, 4]. Assumindo que os retornos vêm de uma distribuição log-normal, então, seu logaritmo é normalmente distribuído. Dessa forma, ao utilizar retornos

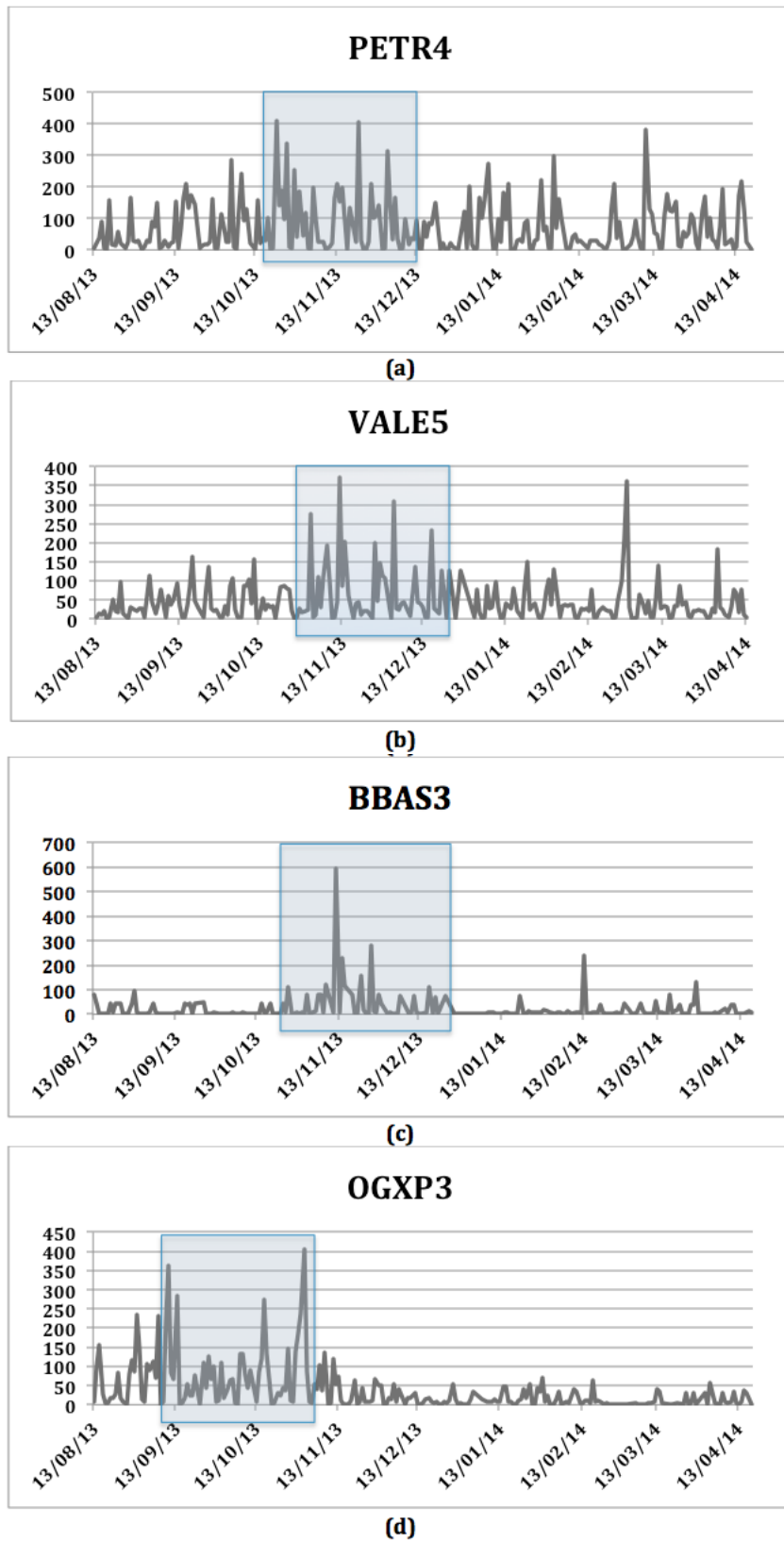


Figura 4.1: Janelas de dados escolhidas para experimento baseadas na quantidade de tweets postados para cada ação.

logarítmicos pode ser mais conveniente no trabalho com análises estatísticas que assumem a normalidade [4],

$$r_t = \ln(p_t) - \ln(p_{t-1}). \quad (4.1)$$

Ambos os conjuntos de valores foram obtidos do *site* da Bolsa de Valores de São Paulo - Bovespa¹.

A Figura 4.2 apresenta, para cada uma das ações selecionadas e para a janela de 9 semanas de experimento, o relacionamento entre as percentagens de volume de tweets coletados em relação ao preço e em relação ao volume de negociação na bolsa.

4.3 Modelos Adotados

Para iniciar os experimentos com os dados, optou-se por realizar um modelo de regressão linear simples com a finalidade de investigar a relação entre as variáveis. Medidas de qualidade também foram produzidas para avaliação e serão descritas abaixo.

4.3.1 Modelo de Regressão Simples

Supondo que a relação entre duas variáveis seja aproximadamente linear, os dados podem ser ajustados por uma reta passando pelos pontos. Um modelo de regressão linear simples pode descrever como o valor esperado y_i para um dado observado y está relacionado com um valor também observado x e com uma parcela de erro. A equação abaixo define o modelo de regressão linear simples, em que y_i é o valor estimado para a variável dependente y e x_i a variável independente para a i -ésima observação. β_0 e β_1 são os parâmetros a serem ajustados e ϵ o erro responsável pela variabilidade em y que não pode ser explicada pela relação linear entre x e y :

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i. \quad (4.2)$$

Objetivando encontrar valores para os parâmetros da equação de regressão de forma que se tenha uma reta ajustada de maneira eficiente, deseja-se que as diferenças entre os valores observados e os estimados sejam mínimas. Por isso, aplica-se o algoritmo dos mínimos quadrados que utilizará os dados amostrais para conhecer os valores β_0 e β_1 , ou seja, o argumento que minimize a soma dos quadrados dos desvios. O critério dos mínimos quadrados é dado por:

$$\arg_{\beta_0 \beta_1} \min \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2. \quad (4.3)$$

Os parâmetros a serem estimados, definidos abaixo, com \bar{y} sendo valor médio da variável dependente, \bar{x} o valor médio da variável independente e n o número de observações. Quando um

¹[http://http://www.bmfbovespa.com.br](http://www.bmfbovespa.com.br)

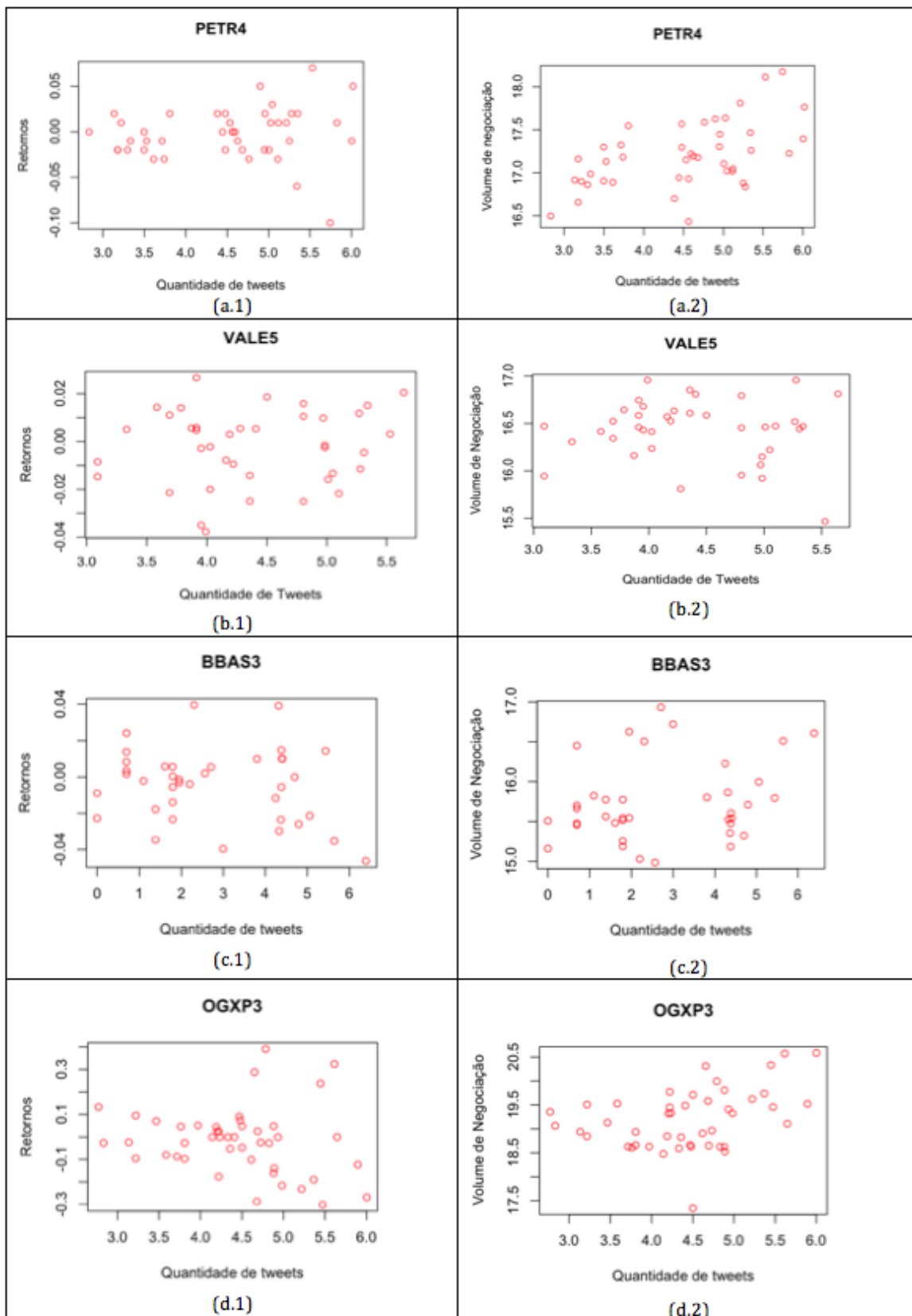


Figura 4.2: Dados coletados para a janela de tempo de 9 semanas definido na Seção 4.2 para as ações (a) PETR4, (b) VALE5, (c) BBAS3 e (d) OGXP3.

valor positivo é obtido para β_1 é indicado que à medida em que x aumenta, aumenta também y :

$$\beta_1 = \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i - \frac{\left(\sum_{i=1}^n x_i \sum_{i=1}^n y_i\right)}{n}}{n \sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}}, \quad (4.4)$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}. \quad (4.5)$$

4.3.2 Medidas de Qualidade

Para medir a qualidade do ajuste do modelo de regressão, serão adotados o coeficiente de correlação, o de determinação R^2 e o erro relativo absoluto - *Relative Absolute Error* (RAE, do inglês), assim como em [4].

Em probabilidade e estatística, a correlação ou coeficiente de correlação se refere a uma medida descritiva da intensidade de associação linear entre duas variáveis x e y , embora não implique em causalidade. Existem vários coeficientes que podem ser utilizados em situações diversas. Um bastante utilizado é o coeficiente de correlação de Pearson, o qual é obtido dividindo-se a covariância de duas variáveis pelo produto de seus desvios padrão. Assim como em [10, 106], que realizaram a correlação entre variáveis de dados de redes sociais e mercado de ações, será extraído o coeficiente de correlação entre dados do mercado de ações e os indicadores descritos na Seção 4.2.1.

Para obter uma medida do grau de associação entre variáveis, usa-se o coeficiente de correlação definido como :

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y} \quad (4.6)$$

onde $\rho_{x,y}$ é o coeficiente de correlação, $cov(x,y)$ a covariância, que mede o grau de interdependência numérica entre as variáveis e $\sigma_x \sigma_y$ e os desvios padrão das variáveis aleatórias x, y . A expressão analítica é dada por:

$$\rho_{x,y} = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{\left[n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 \right] \left[n \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i \right)^2 \right]}}. \quad (4.7)$$

Se o valor desse coeficiente for próximo a 1, então há a indicação de que as variáveis são positivamente e linearmente relacionadas, ou seja, os pontos de dados estão próximos à reta que

tem inclinação positiva, caso contrário, acontecerá quando o valor estiver próximo a -1. Sendo o valor próximo de zero é provável um relacionamento fraco entre as mesmas.

As equações para o cálculo do coeficiente de determinação R^2 e erro relativo absoluto (RAE) são definidas por:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} \quad (4.8)$$

$$RAE = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|^2}{\sum_{i=1}^n |y_i - \bar{y}_i|^2} \quad (4.9)$$

sendo y_i , \hat{y}_i e \bar{y}_i os valores alvo medido, ajustado e média dos valores para a i -ésima das n amostras consideradas. As métricas R^2 e RAE são independentes de escala.

O coeficiente de determinação resume o poder explicativo do modelo de regressão, ou seja, o quão adequadamente a equação de regressão estimada se ajusta aos dados. Ele é calculado a partir das somas dos quadrados dos resíduos, como apresentado na Equação 4.8. O valor de R^2 descreve a proporção da variância da variável dependente explicada pelo modelo de regressão. Sendo 1.0 o modelo de regressão se mostra ideal, se 0.0, nenhuma variação é explicada por meio da regressão. Esse valor expressa a habilidade do modelo estatístico na estimação correta dos valores da variável y .

Quanto ao erro relativo absoluto, quanto menor o seu valor melhor o modelo. Ele expressa quão boa é uma medida em relação ao tamanho do que se está medindo. Comparando o coeficiente de determinação e o erro relativo absoluto, o primeiro é mais sensível a erros individuais mais altos [4].

4.3.3 Teste de Significância

Apenas as medidas de qualidade não são suficientes para determinar a significância da relação entre as variáveis. No caso da equação de regressão linear simples, a média ou valor esperado de y é uma função linear de x eq. 4.2 e sendo $\beta_1 = 0$, o valor de y não dependerá de x , apontando que esses não são linearmente relacionados. Caso contrário se $\beta_1 \neq 0$, conclui-se que x e y estão linearmente relacionados. Desse modo, com a finalidade de verificar a significância de uma relação de regressão, realiza-se um teste de hipóteses para determinar se o valor de $\beta_1 = 0$ [143]. Nesse trabalho, será utilizado o teste t de distribuição t -student.

Assumindo que as variáveis de estudo tenham distribuição normal e variância desconhecida $N(\beta_0 + \beta_1 x_i, \sigma^2)$, e que os erros são independentes e identicamente distribuídos $N(0, \sigma^2)$, para realizar o teste é necessário obter uma estimativa da variância σ^2 de ϵ , que também representa

a variância dos valores de y nas proximidades da reta de regressão. A soma dos quadrados dos resíduos (desvios de y nas proximidades da reta de regressão) fornece uma medida de variabilidade das observações reais em torno da reta estimada, e essa dividida por seus graus de liberdade oferece uma estimativa de σ^2 . Para essa regressão com estimação de dois parâmetros β_0 e β_1 , a soma dos quadrados dos resíduos tem 2 graus de liberdade, produzindo assim uma estimativa não viesada da variância [143]. A estimativa de σ^2 , aqui simbolizada por s , e o desvio padrão podem ser estimados por:

$$s^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2}, \quad (4.10)$$

$$\sigma = s = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2}}. \quad (4.11)$$

Dois hipóteses são testadas para a realização do teste t : $H_0 : \beta_1 = 0$ e $H_a : \beta_1 \neq 0$. Se a primeira for aceita, conclui-se que não há relação estatisticamente significativa entre x e y , o contrário se esta for rejeitada e a segunda aceita. Dessa forma, tem-se uma estatística de teste dada por [143]:

$$T_0 = \frac{\beta_1}{\frac{s}{\sqrt{\sum_{i=1}^n (x_i - \bar{x}_i)^2}}}. \quad (4.12)$$

Logo, H_0 é rejeitado se $|T_0|$ possuir valor inferior à um nível de confiança α considerado. Geralmente adota-se $\alpha = 0,05$.

4.4 Ambiente Computacional

Para a realização dos experimentos para análise estatística e apresentação de resultados foram utilizadas funções disponibilizadas pela ferramenta R² rodando em um MacOS 10.9.3.

O R é um ambiente de *software* livre para computação estatística e gráficos bastante utilizado por grupos de pesquisa espalhados pelo mundo³. Possui um conjunto de funcionalidades para manipulação, cálculo e exibição gráfica de dados. Permite facilidades para armazenamento, operações com vetores, matrizes e listas, ferramentas para análise entre outros. Criada no departamento de estatística da Universidade de Auckland na Nova Zelândia, seu desenvolvimento se deu através da colaboração de desenvolvedores de várias partes do mundo. Seu código fonte está disponível sobre licença GNU GPL. Todos os comandos do R, utilizados nesta pesquisa de doutorado para análise

²<http://www.r-project.org/>

³<http://www.nature.com/news/programming-tools-adventures-with-r-1.16609>

de dados, foram pesquisados em materiais disponíveis em [144, 145].

As tabelas de dados dos resultados foram organizadas no aplicativo Excel, padrão americano, para MacOS da Microsoft. Por esse motivo as casas decimais dos valores apresentados encontram-se separadas por "." e não por ",", que é o padrão adotado em português.

4.5 Resultados da análise estatística inicial

Nas Tabelas 4.1, 4.2 e 4.3 estão disponibilizados os resultados de avaliação e teste de significância para o modelo de regressão linear entre variáveis da bolsa de valores brasileira e variáveis obtidas a partir dos tweets.

Para cada ação selecionada, PETR4, VALE5, BBAS3 e OGXP3, foram avaliados os relacionamentos entre volume de negociação e retornos obtidos da bolsa de valores e os indicadores de tendência B_t , H_t , E_t e de sentimento S_t , esse último apenas para PETR4 e VALE5. Foram avaliados conjuntos de dados em janelas de nove, seis, três e uma semana e também de três dias.

De todos os tweets coletados para as ações, os da PETR4 e VALE5 se destacaram pela quantidade total e diária de tweets suficiente para a realização de testes.

A análise de sentimentos com o polarizador produziu o indicador S descrito na Seção 4.2.1. Os resultados dos relacionamentos estatísticos entre volume de negociação e retornos com o indicador S estão disponíveis nas últimas linhas das Tabelas 4.1 e 4.2.

Nas tabelas de resultados a serem apresentados, estão disponíveis as medições de coeficiente de correlação (Cor), coeficiente de determinação (R^2), nível de significância (p-value) e erro relativo absoluto (RAE) para os relacionamentos entre os indicadores do Twitter e volumes de negociação e retornos das ações da bolsa de valores. O valor de p-value deve ser inferior ao valor de um nível de confiança adotado 0,05, conforme definido na Subseção 4.3.3.

O relacionamento entre os dados coletados do Twitter e os da bolsa de valores, para cada ação dentro da janela de tempo selecionada, foi testado pelo modelo de regressão para a evolução do retorno (r_t), relacionado aos preços dos ativos, sendo $y = r_t$, e volume de negociação (v_t) com $y = v_t$. A variável x está associada a um dos indicadores relacionados aos tweets B_t , H_t , E_t e S_t descritos na Seção 4.2.1, e neste caso de estudo, $t = t - 1$, ou seja y_t , do dia t será correlacionado com o x_{t-1} do dia anterior.

Como informado nas seções anteriores, diferentes indicadores obtidos do montante de dados coletados do Twitter foram utilizados para inferir o relacionamento de regressão linear com volume de negociação e evolução de preço de ações.

Para as ações BBAS3 e OGXP3, foram realizados trinta conjuntos de testes, cada um obtendo um valor para correlação, coeficiente de determinação, teste de significância e RAE. Para as ações PETR4 e VALE5, foram realizados 40 testes em razão do indicador S obtido da polarização dos tweets coletados e pré-processados.

Tabela 4.1: Medições para PETR4, sendo B ; H ; E ; S indicadores obtidos do Twitter, Cor o Coeficiente de Correlação e p-value o valor do teste t de significância.

PETR4											
Tweets	Medição	9 semanas		6 semanas		3 semanas		1 semana		3 dias	
		v_t	r_t	v_t	r_t	v_t	r_t	v_t	r_t	v_t	r_t
B	Cor	0.3562	0.2589	0.2529	0.3197	0.2576	0.3001	0.5444	0.2050	0.9851	0.7594
	R ²	0.1269	0.0670	0.0639	0.1022	0.0663	0.0900	0.2965	0.0420	0.9705	0.5768
	p-value	0.02058	0.0977	0.1774	0.085	0.3539	0.2771	0.3427	0.7408	0.1098	0.4509
	RAE	0.8730	0.9329	0.9360	0.8977	0.9336	0.9099	0.7035	0.9579	0.0294	0.4231
H	Cor	0.1173	0.1609	0.0978	-0.2071	0.1905	0.4993	0.6369	0.8636	0.9705	0.9912
	R ²	0.0137	0.0259	0.0095	0.04291	0.0362	0.2494	0.4057	0.7459	0.942	0.9826
	p-value	0.4591	0.3085	0.607	0.2721	0.4964	0.0580	0.2478	0.0591	0.1549	0.0843
	RAE	0.9862	0.9740	0.9904	0.9570	0.9637	0.7506	0.5942	0.2540	0.0580	0.0174
E	Cor	0.1984	-0.0525	0.7696	0.5385	0.0198	0.5174	-0.2333	-0.5517	-0.6938	-0.6099
	R ²	0.03939	0.0027	0.5923	0.29	0.0003	0.2678	0.0544	0.3044	0.4814	0.3721
	p-value	0.2077	0.7409	6.66e-07	0.0021	0.9441	0.0481	0.7057	0.335	0.5118	0.5823
	RAE	0.9606	0.9972	0.4076	0.7099	0.9996	0.7322	0.9455	0.6955	0.5185	0.6279
S	Cor	0.1871	-0.0870	0.7376	-0.0461	0.2098	0.0150	0.6369	0.8636	0.9705	0.9705
	R ²	0.0350	0.0075	0.5441	0.0021	0.0440	0.0002	0.4057	0.7459	0.942	0.942
	p-value	0.2352	0.5836	3.308e-06	0.8087	0.453	0.9575	0.2478	0.0591	0.1549	0.1549
	RAE	0.96496	0.9924	0.4559	0.9978	0.9559	0.9997	0.5942	0.2540	0.0580	0.0580

Tabela 4.2: Medições para VALE5, sendo B ; H , E e S indicadores obtidos do Twitter, Cor o Coeficiente de Correlação e p-value o valor do teste t de significância.

Vale5											
Tweets	Medição	9 semanas		6 semanas		3 semanas		1 semana		3 dias	
		v_t	r_t	v_t	r_t	v_t	r_t	v_t	r_t	v_t	r_t
B	Cor	-0.0407	0.2071	0.0175	0.1868	-0.0124	0.0403	-0.8631	-0.1487	-0.9355	0.1468
	R ²	-0.0260	0.0163	-0.0341	0.0016	-0.0767	-0.075	0.66	-0.3038	0.7505	-0.9568
	p-value	0.8082	0.2121	0.9253	0.3143	0.9649	0.8866	0.0595	0.8113	0.2298	0.9061
	RAE	0.9983	0.9570	0.9996	0.9651	0.9998	0.9983	0.2550	0.9778	0.1247	0.9784
H	Cor	-0.1182	0.0627	-0.1608	0.0569	0.2767	-0.014	-0.6885	-0.0538	-0.8868	0.8353
	R ²	-0.0134	-0.0237	-0.0077	-0.0311	0.0055	-0.076	0.2987	-0.3295	0.573	0.3956
	p-value	0.4795	0.7083	0.3873	0.7608	0.3181	0.9596	0.1987	0.9314	0.3058	0.3705
	RAE	0.9860	0.9960	0.9741	0.9967	0.9234	0.9997	0.5259	0.9970	0.2134	0.3021
E	Cor	0.0780	-0.1329	0.1023	0.0539	-0.1400	-0.319	0.1979	-0.0538	0.0081	-0.8774
	R ²	-0.0215	-0.0096	-0.0236	-0.0314	-0.0557	0.0329	-0.2811	-0.3295	-0.9999	0.5397
	p-value	0.6415	0.4263	0.5837	0.7733	0.6186	0.2459	0.7496	0.9314	0.9948	0.3185
	RAE	0.9939	0.9823	0.9895	0.9970	0.9803	0.8979	0.9608	0.9970	0.9999	0.2301
S	Cor	-0.0652	0.0270	-0.0918	0.1987	0.0470	0.1493	-0.6885	-0.0538	0.7395	0.2279
	R ²	-0.0234	-0.0270	-0.0257	0.00637	-0.0745	-0.052	0.2987	-0.3295	0.0937	-0.8961
	p-value	0.6971	0.872	0.6233	0.2838	0.8677	0.5952	0.1987	0.9314	0.4701	0.8536
	RAE	0.9957	0.9992	0.9915	0.9605	0.9977	0.9776	0.5259	0.9970	0.4531	0.9480

Tabela 4.3: Medições para BBAS3, sendo B ; H e E indicadores obtidos do Twitter, Cor o Coeficiente de Correlação e p -value o valor do teste t de significância..

BBAS3											
Tweets	Medição	9 semanas		6 semanas		3 semanas		1 semana		3 dias	
		v_t	r_t	v_t	r_t	v_t	r_t	v_t	r_t	v_t	r_t
B	Cor	0.2479	-0.1067	0.2580	-0.0933	0.0648	-0.1060	-0.2642	-0.6518	-0.0479	-0.2321
	R ²	0.0615	0.0113	0.0666	0.0087	0.0042	0.0112	0.0698	0.4249	0.0022	0.0539
	p-value	0.1389	0.5295	0.1685	0.6237	0.8183	0.7069	0.6675	0.2333	0.9695	0.8508
	RAE	0.9384	0.9886	0.9333	0.9912	0.9957	0.9887	0.9301	0.5751	0.9977	0.9460
H	Cor	0.2976	0.0918	0.4330	0.1011	-	-	-	-	-	-
	R ²	0.0885	0.0084	0.1875	0.0102	-	-	-	-	-	-
	p-value	0.0736	0.5888	0.0168	0.5947	-	-	-	-	-	-
	RAE	0.9114	0.9915	0.8124	0.9897	-	-	-	-	-	-
E	Cor	0.2463	0.0093	0.2626	0.0250	0.5259	0.0422	-0.3056	-0.1103	0.0479	0.2321
	R ²	0.0607	8.798e-05	0.0689	0.0006	0.2766	0.0017	0.0934	0.0121	0.0022	0.0539
	p-value	0.1416	0.9561	0.1609	0.8954	0.0440	0.8811	0.617	0.8598	0.9695	0.8508
	RAE	0.9393	0.9999	0.9310	0.9993	0.7234	0.9982	0.9065	0.9878	0.9977	0.9460

Tabela 4.4: Medições para OGXP3, sendo B ; H e E indicadores obtidos do Twitter, Cor o Coeficiente de Correlação e p -value o valor do teste t de significância.

OGXP3											
Tweets	Medição	9 semanas		6 semanas		3 semanas		1 semana		3 dias	
		v_t	r_t	v_t	r_t	v_t	r_t	v_t	r_t	v_t	r_t
B	Cor	0.4093	-0.0746	0.4329	-0.0916	0.4589	-0.2625	-0.1337	-0.9031	0.6543	0.5076
	R ²	0.1675	0.0055	0.1874	0.0083	0.2106	0.0689	0.0179	0.8157	0.4282	0.2577
	p-value	0.0058	0.6303	0.0168	0.6302	0.0852	0.3445	0.8302	0.0356	0.5459	0.661
	RAE	0.8324	0.9944	0.8125	0.9916	0.7893	0.9310	0.9820	0.1842	0.5717	0.7423
H	Cor	0.2548	0.0177	0.3022	0.2141	0.6499	0.3656	0.5581	0.7079	0.4637	0.2967
	R ²	0.0649	0.0003	0.0913	0.0458	0.4225	0.1337	0.3116	0.5011	0.215	0.0880
	p-value	0.0949	0.909	0.1045	0.2557	0.0087	0.1801	0.3282	0.181	0.693	0.8082
	RAE	0.9350	0.9996	0.9086	0.9541	0.5775	0.8662	0.6884	0.4988	0.7849	0.9119
E	Cor	0.2079	-0.0536	0.2168	0.0250	0.2862	-0.1825	0.0500	-0.5465	0.9962	0.9644
	R ²	0.0432	0.0028	0.047	0.0006	0.0819	0.1007	0.0025	0.2987	0.9925	0.9301
	p-value	0.1756	0.7292	0.2498	0.8955	0.301	0.5149	0.9362	0.0649	0.0551	0.1703
	RAE	0.9567	0.9971	0.9529	0.9993	0.9180	0.9666	0.9974	0.7012	0.0074	0.0699

Observando o valor de R^2 nas Tabelas 4.1, 4.2, 4.3 e 4.4, são verificados valores inferiores a 0.1, bem próximos de zero, que indicam um relacionamento fraco entre as variáveis analisadas ou nenhum relacionamento entre as variáveis investigadas. De igual modo, valores próximos à zero também podem ser verificados para o coeficiente de correlação - Cor e medições de nível de significância (p -value) superiores a 0.05 indicando uma medição ruim.

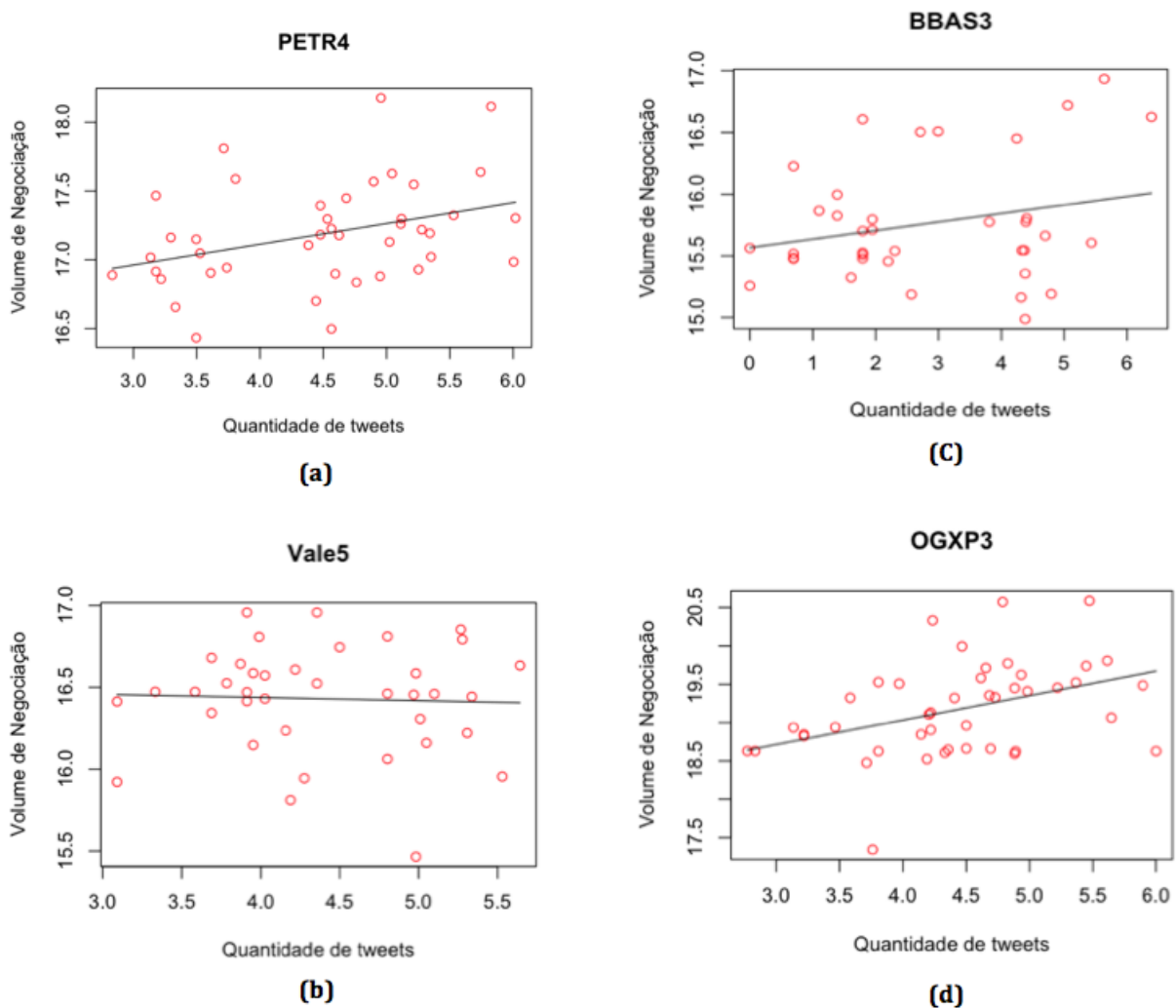


Figura 4.3: Retas Ajustadas (a) PETR4 e (b) VALE5, ambas para a janela de amostra de 9 semanas com variáveis Volume de Negociação e Burburinho.

No entanto, podem ser verificados vários valores aceitáveis. Dos quarenta testes realizados com os dados da PETR4, 6 obtiveram nível de significância inferior 0,05, esses estão marcados com borda dupla na célula da Tabela 4.1, 22 testes obtiveram valor de coeficiente de correlação superior a 0,3, desses 17 foram superiores a 0,5, e 19 com valor de coeficiente de determinação superior a 0,25, dez deles com valor superior a 0,5. Tais valores demonstram que há correlação entre os valores apresentados, embora alguns sejam fracos, existem os que apresentam forte correlação quando os valores ficam próximos a 1,0 e $-1,0$, no caso do coeficiente de determinação os melhores valores são os próximos a 1,0 e para RAE, os menores valores possíveis.

Na Tabela 4.1, últimas cinco linhas em cinza, também estão listados os resultados de análise estatística para o relacionamento entre dados obtidos do analisador de sentimentos supervisionado e os dados do mercado de ações. Nesse contexto, os resultados para seis e uma semana e para três dias foram bons em termos de valores dos coeficientes de determinação e correlação. Porém, como informado na Subseção 3.6.4.2 do Capítulo 3, o nível de acerto do polarizador ficou em 62%, o que pode ser melhorado com ajustes em treinamento, teste e aumento da base de dados de treinamento

com polarização supervisionada.

Na Tabela 4.2, podem ser observados os valores obtidos para os relacionamentos das variáveis pertencentes a VALE5. Para essa ação, dos quarenta testes realizados apenas dez obtiveram valores de correlação superiores a 0,3, desses, oito obtiveram valor de correlação maior que 0,5. Quinze valores de coeficiente de determinação foram superiores a 0,25, desses, sete foram maiores que 0,5. Valores inferiores a 0,05 para testes de significância não foram obtidos. Nesta tabela as quatro últimas linhas na cor cinza apresentam os resultados das análises para o indicador S . Os melhores resultados obtidos foram encontrados nos intervalos de uma semana e de três dias.

A Tabela 4.3 apresenta os valores obtidos para os testes com os dados análogos à BBAS3. Apenas três dos testes apresentaram valores satisfatórios com dois valores de coeficiente de correlação superiores a 0,5 e um com coeficiente de correlação 0,4249 e apenas dois testes de significância menores que 0,05. Essa tabela apresenta em cinza células que não foram preenchidas por falta de dados para o cálculo. Isso significa que não havia quantidade significativa de tweets para a realização dos cálculos de relacionamento estatístico.

Dos testes realizados com a OGXP3, 15 apresentaram valores de coeficiente de correlação superiores a 0,3, nove valores de coeficiente de determinação superiores a 0,25 e apenas três testes de significância menores que 0,05.

Todas as ações apresentaram uma combinação de valores aceitáveis para Cor , R^2 e p -value para a janela de uma semana. Porém, como comentado nos parágrafos anteriores, valores interessantes e isolados de R^2 , Cor e RAE podem ser observados em todas as tabelas e em especial a da PETR4 detentora da maior quantidade de tweets.

Em relação aos indicadores de tweets B ; H e E , todos mostraram resultados mais interessantes para R^2 e teste de correlação para três e uma semana e também para três dias, com exceção de H para BBAS3 por falta de valores. O B certamente por expressar o que está sendo comentado sobre a ação na rede, independente do sentimento, ou seja, sempre possui valores.

Quanto ao RAE, que representa a quão boa a medida é em relação ao tamanho do que está sendo medido, apresentou menores valores para as janelas de uma semana e de três dias para PETR4 e VALE5. Para BBAS3, os valores RAE ficaram próximos de 0,9, uma medida ruim. Para a OGXP3, os melhores valores de RAE foram obtidos para três e uma semana.

Apesar de muitos dos valores apresentados nas tabelas e gráficos não estarem dentro de um intervalo de valores que reforcem o relacionamento linear entre as variáveis consultadas, ao analisar visualmente os textos de tweets postados ao longo dos dias, juntamente com gráficos de preços diários, percebe-se uma forte relação do burburinho com o que está acontecendo no dia. Ao ler as postagens é humanamente possível compreender tendências de queda e subida no preço das ações, volume de negociação e volatilidade do mercado. No caso das ações da PETR4 e VALE5, isso é ainda mais visível, pois há muito comentário sobre ambas relacionado ao que está acontecendo no mercado. As ações da OGXP3 também são bastante comentadas diariamente e, em especial, muito se especula sobre as atitudes de seu proprietário.

Em relação à BBAS3, existem dias em que não há comentários relacionados à ação. Tais

acontecimentos provocam a reunião de dados esparsos que inviabilizam os testes e proporcionam valores de medição ruim.

As Figuras 4.3 (a), (b), (c) e (d) exibem os gráficos de plotagem dos dados em relação à reta ajustada obtida para 9 semanas de medição, explorando assim o relacionamento entre volume de negociação e volume de tweets postados. Observando os dados plotados e as retas adquiridas percebe-se que o conjunto de amostras permanece bem espalhado no gráfico e que a regressão estimou uma reta que habita no meio do espaço ocupado pelas medidas, mostrando que o ajuste levou em conta todos os pontos medidos.

É importante reforçar que para esse experimento foi utilizado um modelo de regressão linear simples sem o uso de pesos ou outros artifícios que podem ser adotados para melhorar a modelagem. Os valores ajustados foram alcançados na tentativa de obter valores para o dia t baseados unicamente em dados do dia $t - 1$. A intenção desse capítulo foi mostrar os passos iniciais e investigativos sobre as variáveis analisadas obtendo, assim, um estudo de verificação de relacionamento entre variáveis vislumbrando possibilidades que foram implementadas e relatadas no próximo capítulo.

4.6 Comentários

Diante dos resultados apresentados, abre-se um amplo espaço de possibilidades. Com relação à análise realizada, é possível perceber o quão desafiador é trabalhar com dados obtidos de rede social e o quanto o caminho é promissor. A cada dia que se passa, ao visualizar o banco de tweets, é possível acompanhar o crescimento do interesse pelo cidadão brasileiro para comentar negociações e o ambiente da bolsa de valores no país. Isso só confirma que o estudo na área é de grande interesse, tendo em vista que o interesse pelo mercado de ações também é crescente pela pessoa física no Brasil.

O modelo de regressão linear, como comentado anteriormente, é o simples. Esse fato demonstra que os resultados apresentados aqui são iniciais e que necessitam ser melhorados. Entretanto, muitas alternativas podem ser aplicadas para análise e averiguação de relacionamento entre retornos calculados dos preços das ações, volume de negociação e os comentários do público na rede social Twitter, algumas dessas outras possibilidades são:

- Aplicação de pesos que favoreçam um conjunto de pontos em relação a outro;
- Adoção de modelo de ajuste de uma curva ampliando a quantidade de parâmetros a serem ajustados com a finalidade de captar um intervalo de confiança maior;
- Adoção de modelos não-lineares e modelos para análise de séries temporais muito utilizados em economia como ARMA e ARIMA;
- Aplicação de indicadores de Twitter juntamente com vários indicadores de mercado para o enriquecimento de modelos de predição e obtenção de estimadores mais robustos;

- Adoção de modelos de predição baseados no burburinho e indicadores do mercado do dia $t-1$, estimar o comportamento do mercado no dia t ;
- Simulação de compra e venda de ações realizadas à partir da análise dos indicadores de sentimento obtidas.

Esse último item será o adotado na próxima etapa desta pesquisa e será relatada no próximo capítulo.

Quanto à análise de sentimentos e polarização ingênua, várias questões levantadas durante esse estudo serão averiguadas e apresentadas no próximo capítulo como forma de melhorar o desempenho desses polarizadores como:

- Melhorar o dicionário de palavras e expressões que indicam tendências no mercado de ações com a finalidade de melhorar a qualidade dos dados obtidos a partir dos contadores ingênuos;
- Verificar a limpeza realizada no banco de dados para a polarização supervisionada, pois a que foi realizada removeu conteúdo que poderia ter valor como, por exemplo, frases com palavrões, tweets com indicação de sites e outros;
- Ajustar melhor os dados de treinamento e avaliação no caso do analisador de sentimentos para obter uma polarização mais acurada.

Capítulo 5

Tomada de decisão para compra e venda de ações

5.1 Introdução

Neste capítulo, será apresentado o desenvolvimento do último estágio (IV) da arquitetura apresentada na Figura 3.1 do Capítulo 3. Este estágio envolve o desenvolvimento do simulador de compra e venda de ações para auxílio na tomada de decisão baseado em indicadores obtidos a partir do Twitter e de preços da bolsa de valores brasileira - Bovespa. Os resultados finais obtidos através do uso desse simulador serão apresentados no capítulo seguinte.

5.2 Arquitetura do Simulador

A Figura 5.1 apresenta a arquitetura planejada para o simulador que recebe dados de indicadores obtidos através do processamento de tweets e de preços de ações e os transforma em decisões de compra e venda dependendo da estratégia e objetivo de lucro escolhidos pelo usuário. Nas seções seguintes serão detalhados cada item dessa arquitetura.

5.3 Dados e Indicadores

Dados recolhidos do Twitter e processados para transformação em indicadores de sentimento e tendências foram tratados no Capítulo 5. Esses indicadores foram analisados a partir de relacionamentos estatísticos com dados de preços de ações e volume de negociação na bolsa de valores - Bovespa no Capítulo 4. Nesse também foram pontuadas algumas questões que poderiam provocar uma melhora na qualidade dos indicadores de Twitter obtidos a partir dos dados, assim, após avaliações optou-se por realizar um ajuste nas técnicas aplicadas que serão comentadas.

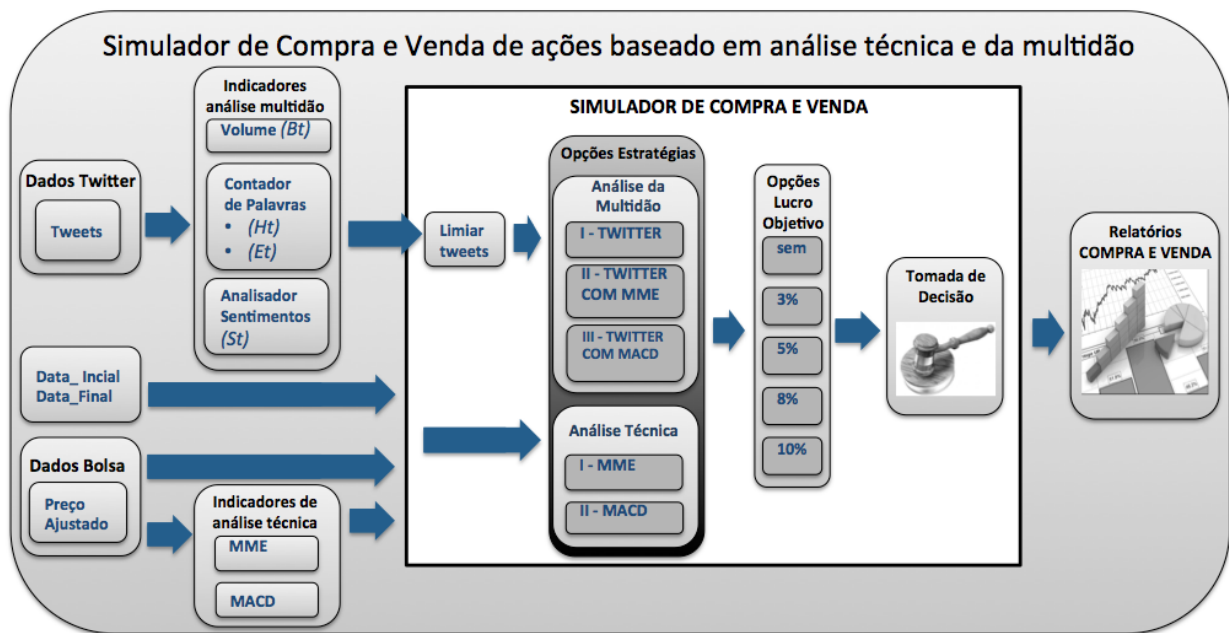


Figura 5.1: Arquitetura do simulador de compra e venda de ações.

5.3.1 Dados do Twitter - Janela de Dados, Pré-Processamento e Classificação

Para a realização de testes com o simulador de compra e venda a ser apresentado neste capítulo, optou-se por trabalhar apenas com as ações PETR4 da Petrobrás e VALE5 da Vale S.A.. Essas foram selecionadas por representarem ações de duas empresas diferentes e de grande importância no mercado de ações brasileiro e por possuírem, dentre as demais ações investigadas, um volume de mensagens (tweets) captadas diariamente suficientes para a realização das simulações e testes.

O intervalo de tempo de coleta de dados utilizado para os testes inicia em 13 de agosto de 2013 e finaliza em 04 de maio de 2015. Os tweets selecionados para PETR4 e VALE5 foram todos os que possuíam em seu conteúdo o nome dessas ações. Dos dados selecionados, foram removidos os tweets coletados nos dias de sábado, domingo, feriados e dias em que ocorreram problemas de indisponibilidade de coleta por motivos técnicos. Para a simulação, são utilizados apenas dos tweets coletados em dias da semana nos quais ocorreu pregão, ou seja, dias comerciais no Brasil. Após essa seleção, foram levantados 426 dias de dados de tweets para a simulação. Para cada um desses dias foram coletados também, a partir do histórico de preços da Bovespa, os preços de abertura, mínimo, máximo e fechamento ajustado para as ações em questão.

Os indicadores obtidos através do processamento de tweets a serem utilizados no simulador, B_t burburinho, H_t humor para compra e venda, E_t expressões de tendência de alta ou baixa nos preços e S_t sentimento positivo ou negativo em relação à ação no mercado para o dia t , são os mesmos definidos na Subseção 4.2.1 do Capítulo 5.

As Figuras 5.2, 5.3 e 5.4 apresentam respectivamente o volume total de tweets coletados, o volume diário sem limpeza e com limpeza para PETR4 e VALE5, sendo que as duas últimas possuem um intervalo vazio localizado entre os dias 13/10/14 e 13/11/14. Estes dias sem coleta

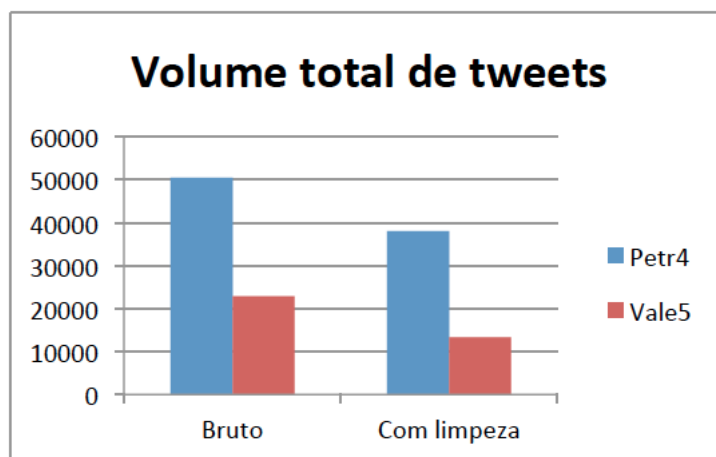


Figura 5.2: Volume de tweets coletados para PETR4 e VALE5 entre 13 de agosto de 2013 e 04 de maio de 2015.

de dados se devem a problemas técnicos (indisponibilidade de hardware e software para realização de coleta).

Da mesma forma que no Capítulo 3, os tweets são coletados na medida em que são publicados na plataforma Twitter, entretanto, o sistema implementado para simulação adota indicadores que refletem o sentimento do dia e não da hora, minuto ou segundo de captação. Todos os tweets coletados no dia t contribuirão para afirmar um valor que representará a tendência ou o sentimento do dia t .

Após análises e reflexões realizadas no Capítulo 4 acerca dos indicadores obtidos dos tweets, optou-se por uma alteração nas atividades de limpeza - etapa de pré-processamento da Figura 3.1 descrita na Seção 3.5 do Capítulo 3. A única filtragem pela qual os tweets coletados foram expostos para essa etapa de simulação foi a de relevância, ou seja, foram removidos da base de dados apenas os tweets que continham palavras ou expressões selecionadas. A filtragem de retweets, bem como a de *links* e pontuação foram removidas da etapa de pré-processamento.

Retweets são cópias de mensagens publicadas por outras pessoas que os usuários postam, ou seja, é replicar algo que foi escrito, sem que o autor perca os créditos por sua autoria. A escolha por manter os retweets na base de dados se deve ao fato de esses expressam o pensamento de outrem reafirmado pelo último usuário que postou. Ou seja, o usuário que retweeta concorda com a postagem do primeiro, sendo que esta passa a ser também sua opinião. Dessa forma, não se conta apenas o pensamento de apenas uma pessoa, mas também o pensamento de todos os que concordaram com o primeiro.

A opção por manter tweets que possuíam *links* (endereços eletrônicos) em seu conteúdo, foi escolhida porque muitas dessas mensagens possuíam texto do usuário acompanhadas de endereços de *sites*. Verificou-se que quando eram removidas, muito do pensamento do povo também era removido dos dados prejudicando o valor e a qualidade dos indicadores de tendência e sentimento.

Após a realização de uma leitura de amostras de toda a base de dados, o script com expressões

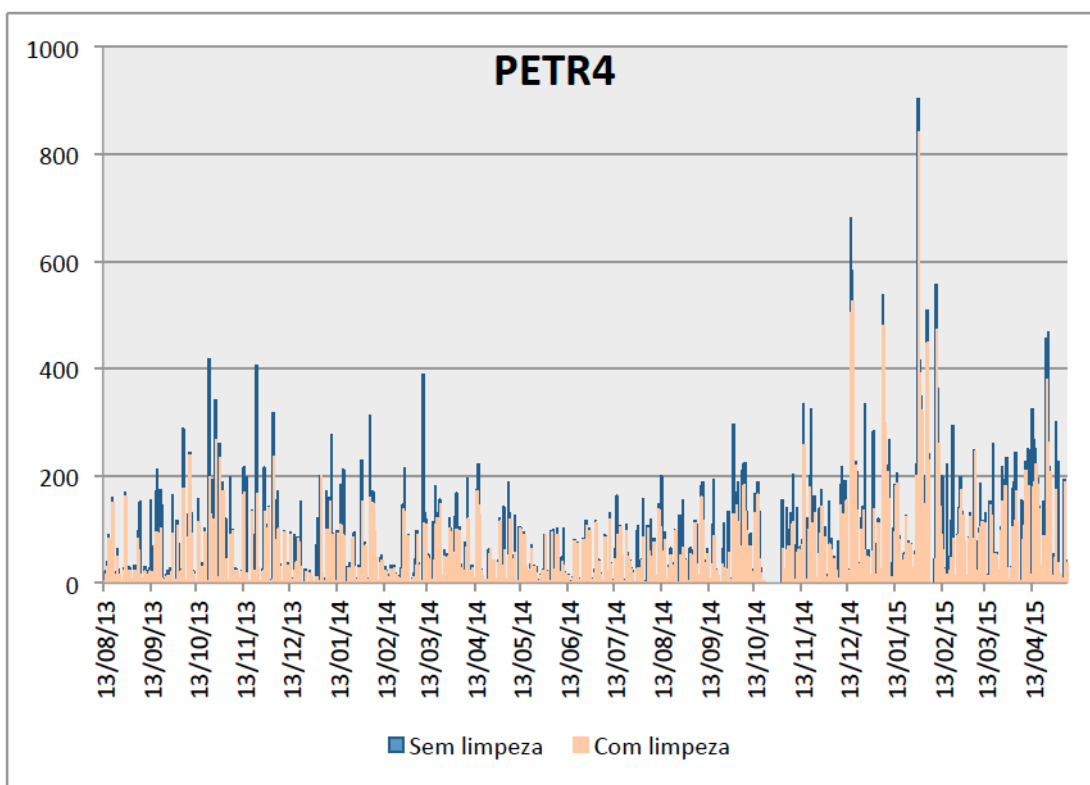


Figura 5.3: Volume de tweets para PETR4 com e sem limpeza.

e palavras selecionadas para remoção foi atualizado e saltou de 300 (quantidade de palavras e expressões utilizadas na limpeza dos dados para análise estatística do Capítulo 4) para 412 expressões. Esse arquivo ao ser aplicado nos dados realiza a limpeza removendo todos os tweets que contenham em seu conteúdo as palavras e expressões selecionadas.

Após a limpeza dos tweets, seguem as atividades da etapa de classificação. Essas foram realizadas conforme definidas na Seção 3.6 do Capítulo 3 e com diferenças em relação à:

- janela de dados que no Capítulo 3 era de oito meses (período de coleta de agosto de 2013 a abril de 2014) e passou para vinte meses (de agosto de 2013 a maio de 2015);
- adoção do elemento neutro na classificação de tweets pelo analisador LingPipe que antes era apenas positivo e negativo;
- formação de um novo conjunto de dados de treinamento para o analisador de sentimentos LingPipe para as ações PETR4 e VALE5.

Tweets selecionados aleatoriamente foram polarizados manualmente em positivos, negativos e neutros com a finalidade de formar um conjunto de dados de treinamento e avaliação para o *software* de análise de sentimentos Lingpipe. Sobre o treinamento e avaliação de tweets:

- PETR4: Dos dados polarizados manualmente, 60% foram utilizados para treinamento de positivos e negativos e 40% para avaliação. A média de acertos foi de aproximadamente 68%;

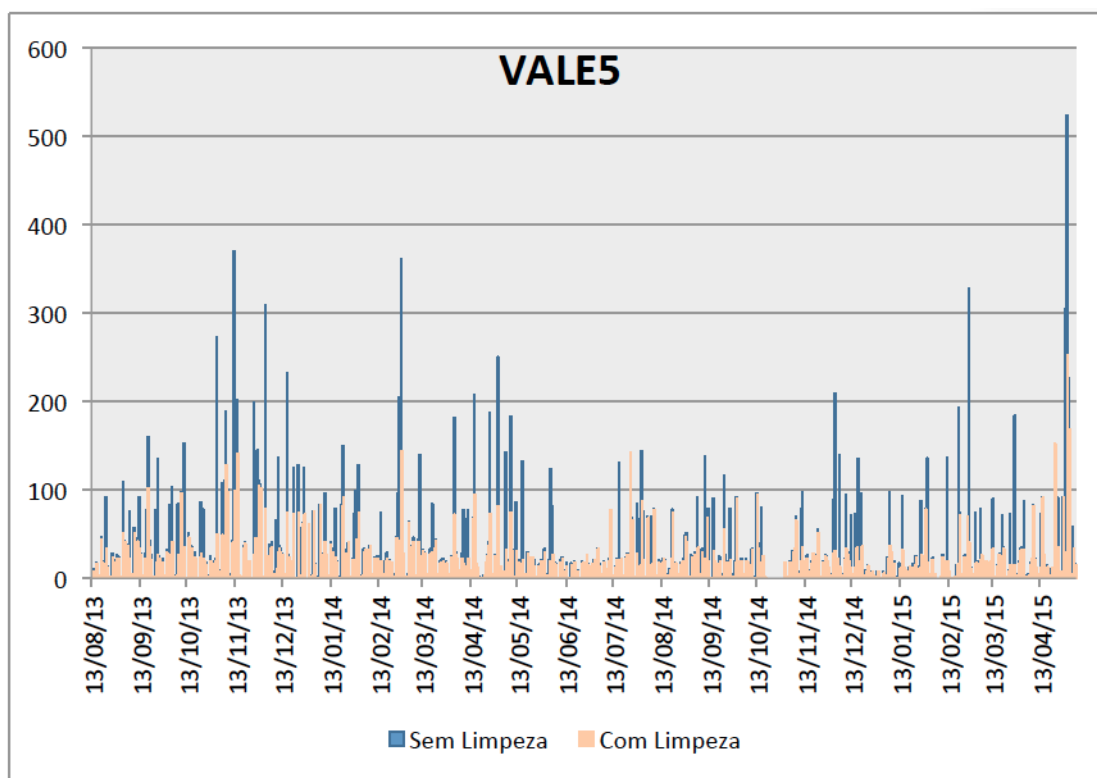


Figura 5.4: Volume de tweets para VALE5 com e sem limpeza

- VALE5: Dos dados polarizados manualmente, 80% foram utilizados para treinamento de positivos e negativos e 20% para avaliação. A média de acertos foi de aproximadamente 54%.

5.3.2 Dados da bolsa e Indicadores de Análise Técnica

Para a simulação de compra e venda, foram utilizados apenas os preços de ações ajustados de abertura, mínimo, máximo e fechamento para PETR4 e VALE5 nos dias em que ocorreu pregão no Brasil, ou seja, os 426 dias de coleta de tweets mencionados acima. Esses valores foram obtidos do site da Bolsa de Valores de São Paulo - Bovespa ¹. As indicações de compra e venda de ações para a simulação foram obtidas através da aplicação de metodologias de análise técnica. No mercado de ações existem duas escolas que se conjugam para a tomada de decisão em compra e venda de ações, segundo [146]:

- Escola gráfica ou técnica – Baseia-se em análise gráfica de volumes e preços pelos quais as ações foram comercializadas em pregões anteriores. Nesta, para comprar e vender ações não se faz necessário pesquisar os fundamentos da empresa, pois o gráfico representa a soma de todos os conhecimentos e expectativas sobre determinada ação e transmite o que o mercado está disposto a pagar por ela. As técnicas pertencentes a essa escola indicam a tendência

¹Bovespa: <http://www.bmfbovespa.com.br/pt-br/mercados/acoes.aspx?idioma=pt-br>

futura dos preços. A análise técnica é essencial para o chamado *market timing* que reflete o momento pertinente para a aquisição ou venda de ação de uma determinada empresa;

- Escola fundamentalista - Baseia-se em resultados setoriais da empresa que se deseja adquirir a ação, levando em consideração o contexto da economia nacional e internacional. A análise fundamentalista é essencial para o chamado *stock picking*, que reflete a seleção da empresa cuja ação deve ser adquirida ou de qual deverá ser vendida em determinado intervalo de tempo.

Dentro da análise técnica existem vários indicadores rastreadores de tendência e osciladores. Estes indicadores ajudam a detectar uma tendência no mercado, entretanto em mercados que não apresentam tendência definida os preços variam dentro de uma faixa de negociação, nesse caso indicadores osciladores ajudam a discernir níveis de suporte e resistência [147].

Alguns indicadores de tendência e osciladores, dentre vários muito utilizados, podem ser citados:

- Médias móveis - São as médias do preço das ações que se deslocam no tempo em decorrência da entrada de novos preços e saída dos mais antigos. Elas promovem a suavização dos ruídos do gráfico de preços, proporcionando uma melhor observação da tendência. Os tipos mais comuns são: simples (média aritmética dos valores); ponderada (média aritmética ponderada); e exponencial (atribui peso crescente exponencialmente do preço mais antigo ao recente). São obtidas por períodos de cinco, nove, vinte, cem ou quantos dias o investidor desejar. Quanto menor o período, menor o atraso no acompanhamento dos preços. As de menor período, geralmente, de cinco, nove ou vinte dias são chamadas de rápidas e as de maior, cinquenta, cem, duzentos dias são chamadas de médias móveis lentas, pois acompanham o preço lentamente. As médias móveis são utilizadas na compra e venda de ações em cruzamentos entre essas e os preços e serão melhor explicadas na próxima subseção, pois serão adotadas pelo simulador, assim como a convergência/divergência de médias móveis – MACD;
- Bandas de Bollinger - São formadas por três curvas que oferecem informações sobre a volatilidade do mercado, pois ofertam uma maneira de detectar a relação risco x retorno do investimento através da análise de limites de volatilidade esperados para tal ativo;
- Volume - Representa a quantidade de ações negociadas em um período de tempo. Podem indicar uma tendência. Usualmente, quando há um alto volume é provável que haja um aumento no preço da ação, quando baixo, pode indicar uma tendência de queda nos preços;
- Regressão linear - Técnica estatística para prever valores futuros baseado em valores do passado (conforme apresentado no capítulo anterior);
- Índice de força relativa - Mede evolução da relação de forças entre compradores e vendedores ao longo do tempo, permitindo a visualização da atenuação de uma tendência ou a identificação de possíveis pontos de reversão;
- Volatilidade - Descreve oscilações diárias nos preços das ações sendo importante para a visualização de potenciais reversões do mercado.

Para o desenvolvimento do simulador, foram selecionados apenas dois indicadores de tendência; o cruzamento de médias móveis exponenciais e a convergência/divergência de médias móveis - MACD, por serem bastante populares e utilizados por investidores que adotam análise técnica para compra e venda de ações.

5.3.2.1 Cruzamento de médias móveis exponenciais

A média móvel exponencial é uma média ponderada das observações passadas e promove a suavização da quantidade de sinais de compra e venda presentes no caso do gráfico de média de preços de ações. Ela permite identificar a tendência do preço do ativo, se de alta ou de baixa e também serve de suporte quando o preço se encontra acima da média, e resistência quando abaixo desta. É representada por uma linha que se movimenta a cada novo dado adicionado.

Para o seu cálculo em relação ao preço do ativo, é necessário possuir o preço, no caso desse trabalho foi utilizado o preço de fechamento ajustado, e o período que representa a quantidade de dias a serem considerados. Dessa forma, sendo n o período (quantidade de dias considerados), $k = 2/(n + 1)$, t o dia em questão, e P_1 o preço atual do ativo, MME pode ser definida como:

$$MME_t = (1 - k) * MME_{t-1} + k * P_1. \quad (5.1)$$

A MME enfatiza os valores mais recentes, pois esses recebem maior peso que os mais antigos cujos pesos caem exponencialmente. Dessa forma, essa média reduz o atraso ao aplicar pesos maiores aos dados mais recentes e, conseqüentemente, reagirá mais rapidamente a uma alteração nos preços atuais sendo sensível a esses. Os valores mais antigos, à medida que o tempo passa, vão sendo esquecidos. Os valores mais comumente utilizados por investidores para o período são [148, 147, 149, 150]:

- 5 a 13 para curto prazo;
- 20 a 50 para médio prazo;
- Acima de 70, longo prazo. As mais utilizadas, nesse caso, são as médias de 100 e 200 dias.

Médias móveis de curto prazo são denominadas rápidas, as demais são lentas por moverem mais lentamente no intervalo de tempo e sofrerem menos influência de valores adicionados recentemente.

Após várias leituras de textos de investidores disponibilizados em sites da Internet [148, 150, 149] e tendo como base o conteúdo dos tweets capturados, nos quais muitos investidores comentam valores de períodos para cálculo de médias, optou-se por adotar os valores MME para os períodos de 5 e 20 e 20 e 200 dias sobre o preço de fechamento ajustado para ações PETR4 e VALE5. Para isso, foram utilizados dados de históricos de preços captados do site da Bovespa de 18 de novembro de 2012 a 03 de maio de 2015 com a finalidade de computar as médias para os preços de 13 de agosto de 2013 a 04 de maio de 2015 compatíveis com dias de dados capturados do Twitter. A Figura 5.5 demonstra as linhas de média móvel exponencial para a ação VALE5.

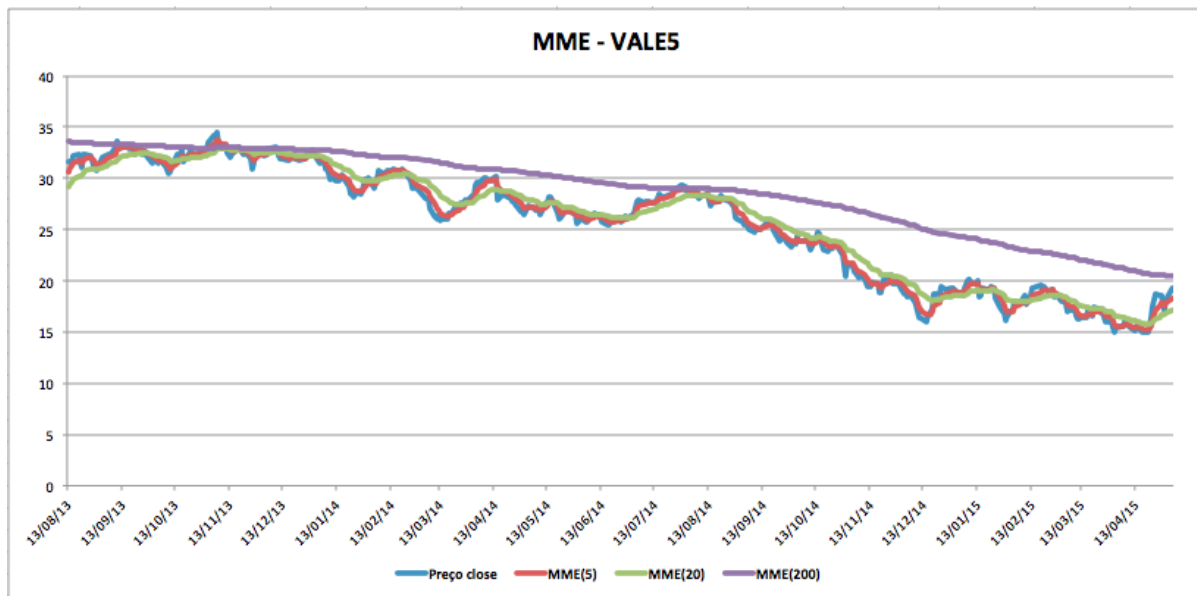


Figura 5.5: Linhas de Média Móvel Exponencial de 5, 20 e 200 dias para o preço de fechamento ajustado da ação VALE5 de 13/08/2013 a 04/05/2015.

5.3.2.2 Convergência/Divergência de médias móveis

O *Moving Average Convergence Divergence* (MACD, do inglês), desenvolvido por Gerald Appel, na década de 60, é um indicador bastante utilizado por investidores para observar pontos de reversão de tendência, auxiliando a tomada de decisão sobre o momento certo para entrar ou sair do mercado. Apesar de não refletir condições de volatilidade frequentes no mercado, continua muito popular.

Esse indicador é composto por duas linhas, uma denominada MACD e outra sinal. A linha MACD é formada pela diferença entre duas médias exponenciais, uma de curto e outra de maior prazo. Geralmente, são utilizadas MME [12] para a mais curta e MME[26] para a mais longa [149, 147], e esses são os valores adotados para a linha MACD utilizada pelo simulador. Dessa forma:

$$MACD = MME[12] - MME[26]. \quad (5.2)$$

Os valores de MME são calculados conforme explicado na subseção anterior.

A linha MACD oscila sem limite inferior ou superior e responde de forma relativamente rápida às mudanças nos preços. Quando a linha se encontra acima de zero, é provável que as entradas recentes reflitam expectativas de alta em relação ao passado, caso contrário, quando abaixo de zero as expectativas recentes são de baixa. Se a linha se mantiver em zero, é provável que operações de compra e venda estejam em equilíbrio. Estando as linhas muito acima de zero, pode-se inferir que o mercado encontra-se comprado, o inverso, vendido.

A linha sinal é computada pela média móvel exponencial de nove períodos dos valores obtidos para a linha MACD:

$$Sinal = MME[9]_{MACD}, \quad (5.3)$$

ela responde às mudanças do mercado mais lentamente que a MACD.

Do cruzamento das linhas Sinal (mais lenta) e MACD (mais rápida) reflete alterações no equilíbrio de forças entre compradores e vendedores. E nesses momentos são identificadas oportunidades para compra ou venda de ações no mercado. Quando a linha MACD cruza a linha sinal para cima o mercado está propício para compra, o oposto acontece se a MACD cruza a sinal para baixo.

Nos anos 80, Thomas Aspray introduziu o conceito de histograma de MACD no qual um gráfico de barras sob a linha de referência zero é arranjado a partir da diferença entre a linha MACD e a linha sinal, sendo:

$$Histograma = MACD - sinal. \quad (5.4)$$

O histograma permite ao investidor a percepção de movimento de queda ou alta mesmo que a linha de preços ofereça tendência conflitante. O início de uma tendência de alta pode ser identificado quando a linha MACD está acima de zero e inicia um movimento de queda enquanto a linha de preço se mantém em tendência de alta, esse movimento é chamado de divergência de baixa.

O contrário acontece quando a linha MACD está bem abaixo de zero e inicia um movimento de subida no momento em que os preços se mantêm em tendência de baixa, nesse caso, é provável o início de uma tendência de alta [149].

No caso do simulador, serão utilizados apenas os cruzamentos das linhas MACD e sinal. Na Figura 5.6, são apresentadas a linha MACD e Sinal obtidas para o período de realização da simulação, bem como uma linha representando os preços de fechamento ajustado para a ação PETR4.

5.4 Simulador de Compra e Venda

O simulador foi desenvolvido em linguagem JAVA por esta ser uma ferramenta que não exige pagamento para a licença uso, também por possuir uma vasta *Application Programming Interface* (API, do inglês) de programação que reduz a necessidade do uso de bibliotecas de terceiros para desenvolvimento, e pelo domínio da linguagem pelo programador.

Conforme o fluxograma da Figura 5.1, o simulador recebe como entrada o valor os indicadores obtidos através do processamento de tweets B_t , H_t , E_t e S_t (burburinho, humor para compra e venda, expressões de tendência de alta ou baixa nos preços e sentimento positivo ou negativo), os preços ajustados da ação $(abertura, mínimo, máximo, fechamento)_t$, as médias móveis exponenciais $(MME(5), MME(20), MME(200))_t$ e o indicador convergência/divergência de média móvel $(MACD, Sinal)_t$ para cada dia t da janela de dados de 13 de agosto 2013 a 04 de maio de 2015.

São disponibilizadas cinco estratégias de compra e venda de ações divididas em duas categorias, uma é baseada em tweets e é denominada análise da multidão e a outra análise técnica. A partir

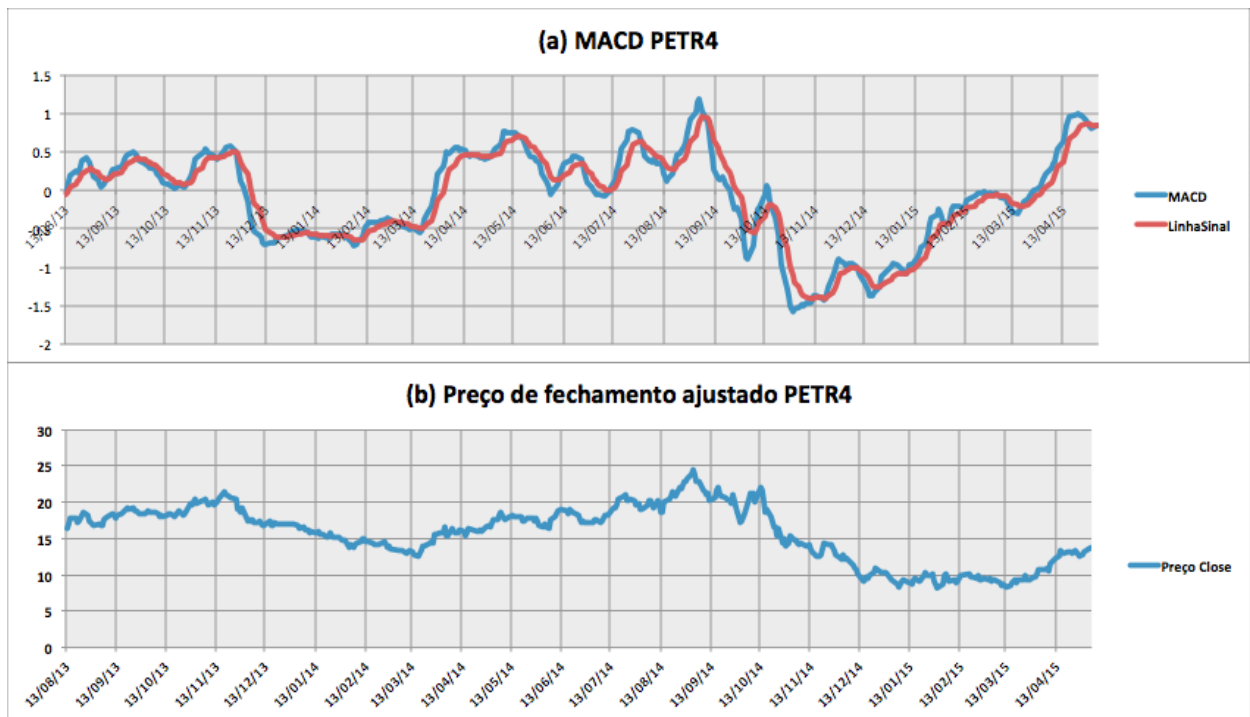


Figura 5.6: (a) Linha MACD e Linha Sinal gerada a partir do preço de fechamento, (b) diário da ação PETR4 de 13/08/2103 a 04/05/2015.

do momento em que uma estratégia é escolhida pelo usuário, o simulador inicia sua atividade.

No Algoritmo 5.1, é apresentado o pseudocódigo de rotinas do simulador, algumas características a serem pontuadas:

- Valores de entrada: : obtidos das bases de dados e histórico de preços: Dados_Bolsa (preços ajustados para a ação) e Dados_Twitter (indicadores $B; H; E; S$); indicados pelo usuário: Limiar (valor que será utilizado quando houver análise da multidão) - é a quantidade mínima de tweets a ser considerada por dia para realização de uma operação, Objetivo_lucro - é o objetivo de lucro desejado na operação e datas data_inicial e data_final - período de tempo a ser considerado; variável Status - reflete o estado do simulador, que inicialmente é NULO e informa ao sistema que não há operação de compra em andamento;
- Processamento: A estrutura de repetição é executada enquanto a Data, que recebe inicialmente a Data_Inicial, ou seja, a partir de quando o usuário quer realizar a simulação, não alcança a Data_Final selecionada pelo usuário. Dentro dela é verificada a situação da variável Status, se uma ação não foi adquirida então é verificada a possibilidade de aquisição através do módulo de Decisão, se esse retornar "COMPRAR" então o status passa a ser "COMPRADO" e os dados da compra (data, preço de fechamento ajustado da ação) são armazenados. Se o Status estiver "COMPRADO" então é verificada a possibilidade de vender através do módulo Decisão, e se esta for verdadeira Status é alterado para "VENDIDO" e os dados da venda são armazenados. No final da estrutura de repetição a variável Data é atualizada.

Algoritmo 5.1 Algoritmo do simulador de compra e venda de ações.

```
1  Dados: Dados_Bolsa, Dados_Twitter, Limiar, Objetivo_lucro, Data_Final,
2      Data = Data_Inicial,
3      Status=NULO, /* Indica o estado do simulador: Comprado ou Vendido*/
4      Opção_Estratégia = ESCOLHA_DA_ESTRATÉGIA (Twitter, MME, MACD, Twitter_com_MME, Twitter_com_MME )
5  Enquanto (Data ≠ Data_Final) repetir
6      Se (status≠Comprado)
7          Então se (Decisão(Data, Opção_Estratégia, Dados_Bolsa e/ou Dados_Twitter e/ou Limiar, Objetivo_lucro)=Comprar)
8              então status = COMPRADO
9                  Armazena(DADOS_Compra)
10             fim_se
11         fim_se
11         Se (status = Comprado)
13             então se (Decisão(Data, Opção_Estratégia, Dados_Bolsa e/ou Dados_Twitter e/ou Limiar, Objetivo_lucro)=Vender)
14                 então status = VENDIDO
15                     Armazena(DADOS_Venda)
16             fim_se
17         fim_se
18         Atualiza(Data)
19 FimEnquanto
20     Mostrar_Relatório(DADOS_Compra, DADOS_Venda)
```

- Saída: Dados de todas as operações de compra e venda da ação efetuadas durante o intervalo de tempo solicitado.

A estrutura do simulador permite à compra de uma única ação, quando houver uma possibilidade de compra indicada pela estratégia, a variável Status é checada, se seu conteúdo for "NULO" ou "VENDIDO" a compra é realizada. A cada operação de venda fica liberada a oportunidade de uma nova compra, sinalizada a partir da variável Status. Dessa forma, durante o período selecionado pelo usuário de Data_Inicial e Data_Final é possível realizar várias operações de compra e venda de uma mesma ação.

5.4.1 Estratégias de análise técnica

A estratégia de análise técnica no simulador é composta de duas opções:

1. MME - Para a simulação foram adotadas as médias móveis exponenciais de 5 e 20 períodos, sendo a primeira a rápida e a segunda a mais lenta. No Algoritmo 5.2 (a) são apresentados os passos para a aplicação desta técnica. O módulo Estratégia_MME recebe a data e os dados da bolsa (preços da ação) como parâmetros de entrada, no processamento os módulos MME_5 e MME_20 retornam os respectivos valores das médias. Se houver um cruzamento de baixo para cima da média mais rápida em relação a mais lenta, então, há a indicação de compra, caso contrário, venda. Se ambas apresentarem o mesmo valor o módulo retornará "NULO".
2. MACD - O Algoritmo 5.2 (b) mostra os passos para a aplicação desta técnica. O módulo Estratégia_MACD recebe a data e os dados da bolsa (preços da ação) como parâmetros

Algoritmo 5.2 Estratégias (a)MME e (b)MACD.

(a) MME

```
1 Estratégia_MME (Data, Dados_bolsa)
2 Se (MME_5(Data)>MME_20(Data)
3     então retornar COMPRAR
4     senão Se (MME_5(Data)<MME_20(Data)
5         então retornar VENDER
6         senão retornar NULO
7     fim_se
8 fim_se
9 retornar NULO
```

(b) MACD

```
1 Estratégia_MACD (Data, Dados_bolsa)
2 Se (MACD(Data)>SINAL(Data))
3     então retornar COMPRAR
4     senão Se (MACD(Data)<SINAL(Data))
5         então retornar VENDER
6         senão retornar NULO
7     fim_se
8 fim_se
9 retornar NULO
```

de entrada, no processamento os módulos MACD e Sinal retornam os respectivos valores, definidos na Seção 5.3.2.2, calculados para a data. Se há o cruzamento de baixo para cima da linha MACD em relação a Sinal, então, há a indicação de compra, caso contrário, venda. Caso ambas apresentem o mesmo valor o módulo retornará "NULO".

5.4.2 Estratégias de análise da multidão

Para a realização de simulação de compra e venda de ações por análise da multidão, foram estabelecidas e implementadas três possibilidades de estratégia:

1. Twitter - Esta se baseia unicamente no sentimento da multidão obtido através dos indicadores E_t ; H_t ; S_t do Twitter. Tais valores são utilizados em uma verificação de combinação de indicadores para a tomada de decisão de compra e venda, ou seja, quando dois deles concordarem em relação a uma tendência, logo esta tendência é levada em consideração para a tomada de decisão. Dessa forma, a estratégia sinaliza a compra ou venda da ação. As regras desenvolvidas para combinação dos indicadores e seus resultados podem ser observadas nas colunas "Regra" e "Estratégia" da Figura 5.7 (b). Uma tendência de alta no preço da ação é observada quando o indicador de expressões E_t retornar um valor positivo. Da mesma forma, quando H_t - indicador de compra e venda - retornar um valor positivo significa que a tendência é de compra para o dia t , provavelmente o preço da ação vai subir. Semelhantemente o indicador S_t , quando positivo, indica que a maior parte dos tweets tuitados no dia t possuem um sentimento positivo, então, é provável que o mercado está passível de alta. Caso contrário acontece quando os indicadores E_t ; H ; S_t possuem valores negativos. Há também a possibilidade de os indicadores possuírem valor zero em seu conteúdo. Nesse caso, não há indicação de tendência e ocorre quando não há tweets para o dia em questão ou a quantidade de sentimento positivo ou de alta ou de compra é igual a quantidade de negativos, baixa e venda respectivamente. Na Figura 5.7 (c), algoritmo de implementação das regras;
2. Twitter com MME - Este método é uma combinação de análise da multidão - Twitter com cruzamento de médias móveis exponenciais. Ela levará em consideração as regras de ambas as técnicas expostas na Figura 5.7 (b) - e no Algoritmo 5.2 (a). O Algoritmo 5.3(a) implementa a junção dos métodos. A rotina indicará uma compra quando as estratégias MME e Twitter

Algoritmo 5.3 Algoritmos das estratégias (a)Twitter com MME e (b)Twitter com MACD.

(a) MME com Twitter

```
Estratégia (Data, Dados_bolsa, Dados_twitter)
1 Se ( Estratégia_MME( Data, Dados_bolsa)=COMPRAR e
2   ( Estratégia_Twitter( Data, Dados_Twitter)= COMPRAR) ou
3   Estratégia_Twitter( Data - 1, Dados_Twitter)= COMPRAR ou
4   Estratégia_Twitter( Data - 2, Dados_Twitter)= COMPRAR ) )
5   então retornar COMPRAR
6   senão Se ( Estratégia_MME(Data, Dados_bolsa) = VENDER e
7     ( Estratégia_Twitter( Data, Dados_Twitter) = VENDER) ou
8     Estratégia_Twitter( Data - 1, Dados_Twitter) = VENDER ou
9     Estratégia_Twitter( Data - 2, Dados_Twitter) = VENDER ) )
10    então retornar VENDER
11    senão retornar NULO
12  fim_se
13 fim_se
```

(b) MACD com Twitter

```
Estratégia (Data, Dados_twitter)
1 Se ( Estratégia_MACD( Data, Dados_bolsa) = COMPRAR e
2   ( Estratégia_Twitter( Data, Dados_Twitter) = COMPRAR) ou
3   Estratégia_Twitter( Data - 1, Dados_Twitter) = COMPRAR ou
4   Estratégia_Twitter( Data - 2, Dados_Twitter) = COMPRAR ) )
5   então retornar COMPRAR
6   senão Se ( Estratégia_MACD ( Data, Dados_bolsa) = VENDER e
7     ( Estratégia_Twitter( Data, Dados_Twitter) = VENDER) ou
8     Estratégia_Twitter( Data - 1, Dados_Twitter) = VENDER ou
9     Estratégia_Twitter( Data - 2, Dados_Twitter) = VENDER ) )
10    então retornar VENDER
11    senão retornar NULO
12  fim_se
13 fim_se
```

retornarem "COMPRA", uma venda quando os métodos retornarem "VENDER" e "NULO" caso sua combinação seja incompatível;

3. Twitter com MACD - É uma combinação de análise da multidão - Twitter com a técnica convergência/divergência de médias móveis MACD. As regras do Twitter dispostas na Figura 5.7 e as MACD no Algoritmo 5.2 (b), juntas serão responsáveis pela compra e venda de ações (b). Haverá uma compra ou venda quando ambas as estratégias retornarem "COMPRAR" ou "VENDER" e nada será feito quando o retorno for "NULO", Algoritmo 5.3(b).

5.4.3 Módulo de Decisão

O módulo Decisão está exposto no Algoritmo 5.4. Ele recebe como parâmetros de entrada a data, a opção selecionada para estratégia, os dados da bolsa e do twitter, o limiar e o objetivo de lucro como entrada. Os valores de opção de estratégia, limiar e objetivo de lucro são escolhidos pelo usuário.

Se a opção de estratégia selecionada pelo usuário for relacionada ao Twitter então inicialmente é realizada a verificação do limiar de tweets. Esse valor é passado pelo usuário ao sistema através da

ESTRATÉGIA TWITTER	
QUANDO sinaliza Positividade é provável que ocorra uma alta no preço do papel = HORA DE COMPRAR	
QUANDO sinaliza Negatividade é provável que ocorra uma baixa no preço do papel = HORA DE VENDER	
OBJETIVO	
Verificar se dois indicadores concordam, isto é, se eles refletem juntos o mesmo sentimento em conjunto.	
REGRAS	
Para comprar	
Se (A e Pos) OU (A e C) OU (C e Pos) então COMPRAR	
Para vender	
Se (B e Neg) OU (B e V) OU (V e Neg) então VENDER	

(a)

Formação das REGRAS para a Estratégia				
Indicadores de Sentimento			Regra	Estratégia
E	H	S		
A	C	POS	Se (A e Pos) então	COMPRAR
A	V	POS	Se (A e Pos) então	COMPRAR
A	N	POS	Se (A e Pos) então	COMPRAR
A	C	NEG	Se (A e C) então	COMPRAR
A	V	NEG	Se (V e Neg) então	VENDER
A	N	NEG		NADA
A	C	N	Se (A e C) então	COMPRAR
A	V	N		NADA
A	N	N		NADA
B	C	POS	Se (C e Pos) então	COMPRAR
B	V	POS	Se (B e V) então	VENDER
B	N	POS		NADA
B	C	NEG	Se (B e Neg) então	VENDER
B	V	NEG	Se (B e Neg) então	VENDER
B	N	NEG	Se (B e Neg) então	VENDER
B	C	N		NADA
B	V	N	Se (B e V) então	VENDER
B	N	N		NADA
N	C	POS	Se (C e Pos) então	COMPRAR
N	V	POS		NADA
N	N	POS		NADA
N	C	NEG		NADA
N	V	NEG	Se (B e Neg) então	VENDER
N	N	NEG		NADA
N	C	N		NADA
N	V	N		NADA
N	N	N		NADA

(b)

Significados	
E - Expressões de Alta ou Baixa	
A - Alta	
B - Baixa	
N - Neutro	
H - Humor para Compra e Venda	
C - Compra	
V - Venda	
S - Sentimento	
POS - Positivo	
NEG - Negativo	
N - Neutro	

(c)

ALGORITMO – ESTRATÉGIA TWITTER

```

1 Estratégia (Data, Dados_twitter)
2 Se ( (E = A) e (S = POS) ) ou ( (E = A) e (H = C) ) ou ( (H = C) e (S = POS) )
3   então retornar Comprar
4   senão Se ( (E = B) e (S = NEG) ) ou ( (E = B) e (H = V) ) ou ( (H = V) e (S = NEG) )
5     então retornar Vender
6     senão retornar NULO
7   fim se
8 fim se

```

(d)

Figura 5.7: (a)Estratégia baseada no Twitter, (b) formação das regras para compra e venda de ações, (c) significado dos símbolos utilizados nas regras e (d) algoritmo da estratégia Twitter.

Algoritmo 5.4 Módulo de tomada de decisão do simulador.

```
1 Decisão (Data, Opcao_Estrategia,Dados_Bolsa e/ou Dados_Twitter e/Limiar, Objetivo_lucro)
  /*Para estratégias com twitter verificar VolumeTwitter > Limiar.
  As estruturas abaixo são realizadas conforme a estratégia escolhida e recebida por Opção_Estratégia */
2 se ( Estratégia(Data, Dados_Bolsa e/ou Dados_Twitter ) = COMPRAR )
3   então retornar COMPRAR
4 fim_se
5 se ( Objetivo_Lucro > 0 )
6   então se ( (Estratégia(Data, Dados_Bolsa e/ou Dados_Twitter ) = VENDER) ou
7     (LucroProvável(Data,Dados_Bolsa) >= Objetivo_Lucro ) )
8     então retornar VENDER
9   fim_se
10  senão se (Estratégia(Data, Dados_Bolsa e/ou Dados_Twitter ) = VENDER )
11    então retornar VENDER
12  fim_se
13 fim_se
14 retornar NULO
```

variável Limiar que está presente na Figura 5.1 e no Algoritmo 5.1. Se o valor do Limiar for maior que zero, uma operação de compra ou venda só poderá ser realizada se houver, no dia avaliado, no mínimo a quantidade de tweets apontada por Limiar para realizar a operação. Caso o usuário opte pelo valor zero, qualquer quantidade de tweets poderá ser utilizada para a realização de decisão de compra e venda de uma ação.

Seguidamente, o módulo Estratégia é invocado, este retornará "COMPRAR", "VENDER" ou "NULO" dependendo da situação dos indicadores e da estratégia escolhida. Seja "COMPRAR" o retorno do módulo Estratégia, o módulo Decisão retornará "COMPRAR" ao que o chamou - Algoritmo 5.1. Se o retorno do módulo Estratégia for diferente de "COMPRAR", é checado se há especificação de objetivo de lucro para venda, se houver uma venda, será retornada quando o módulo Estratégia retornar "VENDER" ou quando o objetivo de lucro para venda for alcançado. Se não houver objetivo de lucro para venda, apenas o método Estratégia é testado. Se o retorno de Estratégia for diferente de "COMPRAR" e "VENDER", o retorno de Decisão será "NULO".

5.5 Saídas do Simulador

O Algoritmo 5.1 apresentou os passos de processamento das entradas do simulador seguidos para a tomada de decisão de compra e venda de ação pelo sistema. Entretanto, na última linha do algoritmo há uma chamada para o módulo Mostrar_Relatório, esse recebe como entrada os dados de compra e venda de cada ação negociada entre as datas inicial e final e emite na tela os resultados da simulação.

Antes de detalhar as saídas, é importante explicar as formas possíveis de atuação de um investidor e como o simulador os representará. Um investidor pode optar por realizar operações no mercado de ações conduzido pela tendência dos indicadores, ou não, seguindo então contratendência. Quando segue a tendência, ele age de acordo com os indicadores obtidos pelas técnicas adotadas. No caso do simulador, ele segue a tendência dos indicadores de análise da multidão e análise técnica, ou seja, compra quando estes apontarem "COMPRAR" e vende quando sinalizarem "VENDER". A atuação contratendência funciona exatamente o contrário, se a indicação é

"COMPRAR" então, se opta por vender, se "VENDER" então é o momento para comprar. Pode ainda atuar na condição de comprado, vendido ou comprado/vendido:

1. Comprado - Assume-se uma postura otimista em relação ao ativo confiando na alta dos preços, ou seja, adquire-se um ativo na baixa com a expectativa de vender na esperança de uma alta dos preços. Quando uma ação é adquirida pelo simulador o status do sistema passa a ser "COMPRADO", ou seja, significa que o sistema através de suas regras determinou a compra de uma ação. O status do sistema só mudará quando o sistema acusar venda através de suas análises.
2. Vendido - Espera-se lucrar com a baixa dos preços. Vende-se um ativo por um preço mais caro crendo que poderá adquiri-lo novamente por um preço mais baixo. O simulador passa para o estado de "VENDIDO" quando de posse de uma ação, e ainda no estado "COMPRADO", o sistema identifica um momento de venda. Assim, a ação é vendida e o estado do sistema passa a ser "VENDIDO", podendo a partir desse instante adquirir outra ação.
3. Comprado/Vendido - Se ganha com o empréstimo ou aluguel de ações. Funciona da seguinte forma: se o investidor não tem a ação para vender, ele a toma emprestado pagando uma taxa de aluguel ao dono, em seguida, o investidor a vende. Estando certo de que o preço da ação cairá, o investidor a compra de volta por um preço inferior e a devolve a quem o emprestou. Dessa forma, o investidor ganhou com a queda de uma ação que a princípio ele não tinha, assumindo primeiramente a posição de vendido e posteriormente de comprado ao adquirir a ação de volta.

O sistema emitirá para cada simulação três janelas de resultados, essas janelas serão apresentadas visualmente no próximo capítulo, que descreverá os resultados obtidos com a simulação. A primeira é constituída por quatro gráficos:

- o primeiro deles apresenta as linhas de preço de fechamento ajustado e as MME de 5, 20 e 200 períodos, bem como as marcações com os caracteres "C" e "V" que indicam respectivamente o momento para compra e venda segundo o Twitter;
- o segundo MACD apresenta as linhas MACD e Sinal;
- o terceiro, gráfico de barras, apresenta o volume de Tweets capturados para o período.

A segunda janela apresentará os dados de saída da simulação dispostos em uma tabela. Nas linhas os resultados de cada operação e nas colunas as saídas:

- Data Compra - Mostra os dias em que a compra de uma ação foi realizada;
- Data Venda - Mostra os dias em que a ação adquirida na Data Compra foi vendida;
- Lucro - lucro médio por operação para cada compra e venda efetuada, conforme a forma de atuação no mercado que pode ser comprado, vendido ou comprado/vendido seguindo ou não a tendência, conforme os termos explicados acima;

- Preço Compra - Preço de fechamento ajustado para a ação na data da compra;
- Preço Venda - Preço de fechamento ajustado para a ação na data da venda;
- Quantidade dias Comprado - Quantidade de dias em que o simulador permaneceu com status COMPRADO.

A terceira janela emitirá um relatório final de toda a simulação com os dados:

- Quantidade de dias em que o sistema permaneceu como COMPRADO;
- Quantidade de operações realizadas;
- Lucro acumulado com todas as operações;
- Lucro médio por operação;
- *Sharpe ratio*;
- Porcentagem de operações lucrativas.

O simulador, ao realizar suas atividades de compra e venda de ações, atua como comprado e seguindo a tendência e, a partir dos dados de saída obtidos (data_compra, data_venda, preço compra, preço venda, quantidade de dias comprado), ele calcula também as saídas para uma atuação na forma de vendido e comprado/vendido seguindo ou não a tendência. Para isso, foram realizados os seguintes cálculos, sejam $r_1, r_2, r_3, \dots, r_n$ os retornos obtidos por cada uma das n operações realizadas em um período de N dias de simulação, assim:

- o retorno para cada operação seguindo a tendência - d - atuando como comprado - c - , vendido - v - ou compra/vendido - cv :

$$r_{dc} = \frac{P_{venda_i}}{P_{Compra_i}} - 1, \quad (5.5)$$

$$r_{dv_i} = \frac{P_{venda_i}}{P_{Compra_{i+1}}} - 1, \quad (5.6)$$

$$r_{dcv} = \left(\frac{P_{venda_i}}{P_{Compra_i}} \right) * \left(\frac{P_{venda_i}}{P_{Compra_{i+1}}} \right) - 1. \quad (5.7)$$

- O retorno para cada operação seguindo contra a tendência - cd - atuando como comprado - c - , vendido - v - ou compra/vendido - cv :

$$r_{cdc_i} = - \left(\frac{P_{venda_i}}{P_{compra_i}} - 1 \right), \quad (5.8)$$

$$r_{cdv_i} = - \left(\frac{P_{venda_i}}{P_{compra_{i+1}}} - 1 \right), \quad (5.9)$$

$$r_{cdcv_i} = - \left(\left(\frac{P_{venda_i}}{P_{compra_i}} \right) * \left(\frac{P_{venda_i}}{P_{compra_{i+1}}} \right) - 1 \right). \quad (5.10)$$

- O lucro médio por operação para operações que seguem ou não a tendência, optando por operar comprado, vendido ou comprado/vendido é calculado pela média aritmética dos retornos:

$$\bar{r} = \frac{1}{n} \sum_{i=1}^n r_i. \quad (5.11)$$

- O lucro médio acumulado - a -, obtido para comprado, vendido ou comprado/vendido, mas seguindo a tendência é computado pelo produtório:

$$r_{ad} = \left(\prod_{i=1}^n (1 + r_i) \right) - 1. \quad (5.12)$$

- O lucro médio acumulado seguindo contra a tendência para comprado, vendido ou comprado/vendido é dado pelo produtório:

$$r_{acd} = \left(\prod_{i=1}^n (1 - r_i) \right) - 1. \quad (5.13)$$

Além dos retornos e lucro médio acumulado, o simulador também calcula o índice de *sharpe ratio* definido matematicamente como:

$$SR = \frac{E[r_i - r_f]}{\sigma_i}. \quad (5.14)$$

Sendo $E[.]$ o valor esperado para o retorno i do ativo, r_f o retorno sobre um ativo de referência, tal como uma taxa livre de risco ou um índice como CDI² no caso do Brasil ou S&P500³ nos EUA, e σ_i seu desvio padrão.

O índice de *sharpe ratio* é um importante indicador financeiro desenvolvido por William F. Sharpe para avaliação de rentabilidade e risco do investimento, ou seja, ele permite julgar o quanto o retorno compensa o risco assumido pelo investidor. Sua definição mede a relação entre o retorno excedente ao ativo livre de risco e a volatilidade. Se uma comparação for feita entre dois ativos e um ponto de referência comum, aquele que possuir um *sharpe ratio* mais elevado proporcionará um melhor retorno para o mesmo risco, assim quanto maior o valor do índice, melhor.

Para a aplicação do *sharpe ratio* sobre os dados de saída do simulador foram adotados os seguintes esquemas, sejam x_j , x_j^* , i e l respectivamente para o mês j , a rentabilidade das operações encerradas, o CDI do mês (a Tabela 5.1 apresenta os valores do CDI para os meses avaliados), i iniciando da primeira operação finalizada até l a última e I sendo o total de meses nos quais

²Certificado de Depósito Interbancário, atualmente DI (Depósito Interfinanceiro) - São títulos de emissão das instituições financeiras monetárias e não-monetárias que lastreiam as operações do mercado interbancário, mais informações em [146].

³Standard & Poor's 500 - Índice que representa o desempenho do mercado de bolsa de valores norte-americano. Os 500 representam as quinhentas ações mais importantes do mercado, essas são de empresas escolhidas por um comitê levando em consideração o tamanho, liquidez e setor dessas empresas.

as operações foram completadas durante toda a simulação. Dessa forma, a partir definição da Equação 5.14, tem-se que:

$$x_j = \left(\prod_{i=1}^l (1 + r_i) \right) - 1, \quad (5.15)$$

$$SR = \frac{\frac{1}{I} \cdot \sum_{j=1}^I (x_j - x_j^*)}{\sqrt{\frac{1}{I} \cdot \sum_{j=1}^I (x_j - x_j^* - m)}}, \quad (5.16)$$

com $m = \frac{1}{I} \sum_{j=1}^I (x_j - x_j^*)$.

Os resultados obtidos pelo simulador para a tomada de decisão para a compra e venda de ações para PETR4 e VALE5 serão apresentados e discutidos no próximo capítulo.

Tabela 5.1: Índice CDI mensal (<http://www.cetip.com.br>).

Ano-Mês	% CDI
ago/13	0.6957%
set/13	0.6991%
out/13	0.8033%
nov/13	0.7105%
dez/13	0.7803%
jan/14	0.8398%
fev/14	0.7827%
mar/14	0.7600%
abr/14	0.8154%
mai/14	0.8583%
jun/14	0.8174%
jul/14	0.9404%
ago/14	0.8595%
set/14	0.9006%
out/14	0.9448%
nov/14	0.8379%
dez/14	0.9558%
jan/15	0.9293%
fev/15	0.8185%
mar/15	1.0361%
abr/15	0.9483%
mai/15	0.9838%

Capítulo 6

Resultados6

6.1 Introdução

Este capítulo apresenta os resultados alcançados com os dados de entrada processados pelo simulador. As saídas obtidas para as ações PETR4 e VALE5 serão analisadas e discutidas nas seções posteriores.

6.2 Dados e Ferramentas

Sobre os dados de entrada utilizados para a simulação:

- Quantidade total de dias: 426 dias foram apresentados ao sistema que compreendem os dados coletados entre 13 de agosto de 2013 e 04 de maio de 2015. Destes foram removidos os dias não comerciais (sábados, domingos e feriados) e dias com problemas técnicos (quando não houve captura por motivos técnicos de queda de sinal de energia, sinal Internet, problemas no software de captura ou banco de dados);
- Quantidade total de tweets: A base de dados total possui 8.144.657 tweets, a partir desses, uma rotina de seleção os separou em 50.500 tweets para PETR4 e 22.868 para VALE5. Após a aplicação das rotinas de limpeza dos dados, conforme os critérios estabelecidos na Seção 5.3.1 do Capítulo 5, PETR4 passou a conter 38.070 tweets e VALE5 13.218;
- Dados da bolsa de valores: Foram utilizados os valores de preço de fechamento ajustado para as ações PETR4 e VALE5 para os 426 dias de negociação. No caso da análise técnica que realiza os cruzamentos de médias móveis, o histórico de preços das ações foi utilizado para os cálculos das médias necessárias.
- Limiar de tweets: Foram utilizados inicialmente dois valores, $Limiar = 0$ e $Limiar = (\frac{1}{n} \sum_{t=1}^n B_t)/2$ com n amostras, o segundo foi calculado a partir do indicador Twitter B_t , que reflete a quantidade total de tweets trafegados no dia t . Esse valor foi escolhido de forma aleatória;

- Objetivo de lucro desejado: O usuário pode optar por realizar as simulações sem objetivo de lucro ou optando por lucro desejado. As opções testadas durante a simulação foram 3%, 5%, 8% e 10%, também escolhidas aleatoriamente.

As saídas do simulador são: data_compra; data-venda; preço compra; preço venda; quantidade de dias comprado; o retorno para as estratégias que seguem ou não a tendência atuando como comprado; vendido e comprado/vendido; e índice *sharpe ratio*. Para melhor visualização e organização das saídas do simulador, essas foram exportadas para uma planilha do Microsoft Excel para Mac (2011) e nela foram calculadas para cada simulação o lucro médio por operação e a porcentagem de operações lucrativas.

As informações sobre as simulações realizadas, bem como os dados de saída, serão exibidos em tabelas detalhando o objetivo de lucro desejado, a estratégia (comprado, vendido, comprado/vendido seguindo a tendência ou contra a tendência) e o limiar adotado. Para facilitar a visualização dos lucros acumulados das operações, cores diferentes foram adotadas:

- Para o lucro acumulado maior que 0,00% e menor que 30,00%, este será marcado na cor amarela;
- Sendo maior ou igual a 30,00% e menor que 100,00% será marcado na cor laranja;
- Se maior ou igual a 100,00% será marcado com a cor azul;
- Os três maiores lucros acumulados de cada tabela estão em negrito.

As mesmas condições de simulação elaboradas foram empregadas para as ações da PETR4 e VALE5. As saídas para cada simulação serão comentados nas seções seguintes.

As tabelas de dados dos resultados foram organizadas no aplicativo Excel, padrão americano, para MacOS da Microsoft. Por esse motivo as casas decimais dos valores apresentados encontram-se separadas por "." e não por "," que é o padrão adotado em português.

6.3 Resultados para a simulação de análise técnica

As Tabelas 6.1 e 6.2 apresentam os valores obtidos para a simulação de compra e venda das ações PETR4 e VALE5 por métodos de análise técnica MACD, MME de 5 e 20 períodos e MME de 20 e 200 períodos, todas sem interferência alguma de dados de redes sociais. Essas técnicas utilizam do preço das ações para indicar momentos propícios para a compra e venda, conforme relatado no capítulo anterior. As tabelas contêm células vazias porque para alguns dados de entrada não houve resultados.

Para a análise técnica, foram executadas 90 simulações para cada uma das ações. Sobre cada método de análise técnica simulada, algumas informações interessantes:

- MACD: foram 5 operações lucrativas para PETR4 e 15 para VALE5. O maior lucro acumulado e *sharpe ratio* para PETR4 foi de 56,95% e 0,23, adotando a condição de COM-

PRADO/VENDIDO, contratendência e com ou sem objetivo de lucro. Para a VALE5, adotando as mesmas características, o lucro acumulado foi de 65,19%. O menor lucro acumulado para VALE5 foi de 9,90% na condição de VENDIDO, seguindo a tendência e com ou sem objetivo de lucro;

Tabela 6.1: Resultados da simulação de compra e venda da PETR4 por análise técnica.

PETR4 -ANÁLISE TÉCNICA											
Objetivo	Estratégias	Quantidade operações	Dias Comprado	Tendência - COMPRADO				Contra - tendência - COMPRADO			
				Lucro Médio por operação	Lucro Acumulado	Sharpe Ratio	% Operações Lurativas	Lucro Médio por operação	Lucro Acumulado	Sharpe Ratio	% Operações Lurativas
0%	MACD	19	252	-0.63%	-24.51%	-0.1074	42.11%	0.63%	-9.27%	0.2348	57.89%
	MME 5 e 20	12	186	-1.37%	-17.99%	-0.2879	50.00%	1.37%	14.03%	-0.1893	50.00%
	MME 20 e 200	2	149	-16.68%	-30.69%	-5.0326	0.00%	16.68%	36.02%	0.8854	100.00%
3%	MACD	19	252	-0.63%	-24.51%	-0.1074	42.11%	3.11%	-9.27%	0.2348	57.89%
	MME 5 e 20	12	186	-1.37%	-17.99%	-0.2879	50.00%	1.37%	14.03%	-0.1893	50.00%
	MME 20 e 200	2	149	-16.68%	-30.69%	-5.0326	0.00%	16.68%	36.02%	0.8854	100.00%
5%	MACD	19	252	-0.63%	-24.51%	-0.1074	42.11%	3.11%	-9.27%	0.2348	61.11%
	MME 5 e 20	12	186	-1.37%	-17.99%	-0.2879	50.00%	1.37%	14.03%	-0.1893	50.00%
	MME 20 e 200	2	149	-16.68%	-30.69%	-1.5626	0.00%	16.68%	36.02%	0.8854	100.00%
8%	MACD	19	252	-0.63%	-24.51%	-0.1074	42.11%	3.11%	-9.27%	0.2348	57.89%
	MME 5 e 20	12	186	-1.37%	-17.99%	-0.2879	50.00%	1.37%	14.03%	-0.1893	50.00%
	MME 20 e 200	2	149	-16.68%	-30.69%	-5.0326	0.00%	16.68%	36.02%	0.8854	100.00%
10%	MACD	19	252	-0.63%	-47.59%	-0.1074	42.11%	0.63%	-9.27%	0.2348	57.89%
	MME 5 e 20	12	186	-1.37%	-17.99%	-0.2879	50.00%	1.37%	14.03%	-0.1893	50.00%
	MME 20 e 200	2	149	-16.68%	-30.69%	-5.0326	0.00%	16.68%	36.02%	0.8854	100.00%
Tendência - VENDIDO											
Objetivo	Estratégias	Quantidade operações	Dias Vendido	Tendência - VENDIDO				Contra - tendência - VENDIDO			
				Lucro Médio por operação	Lucro Acumulado	Sharpe Ratio	% Operações Lurativas	Lucro Médio por operação	Lucro Acumulado	Sharpe Ratio	% Operações Lurativas
0	MACD	18	251	0.28%	-3.60%	-0.0510	38.89%	-0.28%	-15.47%	-0.1073	61.11%
	MME 5 e 20	12	186	5.36%	43.58%	0.2024	25.00%	-5.36%	-80.13%	-0.2612	75.00%
	MME 20 e 200	1	148	-3.52%	-3.52%	-1.5626	0.00%	3.52%	3.52%	0.4876	100.00%
3%	MACD	18	251	0.28%	-3.60%	-0.0510	38.89%	-0.28%	-15.47%	-0.1073	61.11%
	MME 5 e 20	12	186	5.36%	43.58%	0.2024	25.00%	-5.36%	-80.13%	-0.2612	75.00%
	MME 20 e 200	1	148	-3.52%	-3.52%	-1.5626	0.00%	3.52%	3.52%	0.4876	100.00%
5%	MACD	18	251	0.28%	-3.60%	-0.0510	38.89%	-0.28%	-15.47%	-0.1073	61.11%
	MME 5 e 20	12	186	5.36%	43.58%	0.2024	25.00%	-5.36%	-80.13%	-0.2612	75.00%
	MME 20 e 200	1	148	-3.52%	-3.52%	-1.5626	0.00%	3.52%	3.52%	0.4876	100.00%
8%	MACD	18	251	0.28%	-3.60%	-0.0510	38.89%	-0.28%	-15.47%	-0.1073	61.11%
	MME 5 e 20	12	186	5.36%	43.58%	0.2024	25.00%	-5.36%	-80.13%	-0.2612	75.00%
	MME 20 e 200	1	148	-3.52%	-3.52%	-1.5626	0.00%	3.52%	3.52%	0.4876	100.00%
10%	MACD	18	251	0.28%	-3.60%	-0.0510	38.89%	-0.28%	-15.47%	-0.1073	61.11%
	MME 5 e 20	12	186	5.36%	43.58%	0.2024	25.00%	-5.36%	-80.13%	-0.2612	75.00%
	MME 20 e 200	1	148	-3.52%	-3.52%	-1.5626	0.00%	3.52%	3.52%	0.4876	100.00%
Tendência - COMPRADO/VENDIDO											
Objetivo	Estratégias	Quantidade operações	Dias Comprado/Vendido	Tendência - COMPRADO/VENDIDO				Contra - tendência - COMPRADO/VENDIDO			
				Lucro Médio por operação	Lucro Acumulado	Sharpe Ratio	% Operações Lurativas	Lucro Médio por operação	Lucro Acumulado	Sharpe Ratio	% Operações Lurativas
0	MACD	18	251	-3.11%	-49.48%	-0.3987	38.89%	3.11%	56.95%	0.2348	61.11%
	MME 5 e 20	12	186	3.35%	17.75%	0.1347	33.33%	-3.35%	-58.69%	-0.1893	66.67%
	MME 20 e 200	1	148	-16.33%	-16.33%	-1.1169	0.00%	16.33%	16.33%	0.8854	100.00%
3%	MACD	18	251	-3.11%	-49.48%	-0.3987	38.89%	3.11%	56.95%	0.2348	61.11%
	MME 5 e 20	12	186	3.35%	17.75%	0.1347	33.33%	-3.35%	-58.69%	-0.1893	66.67%
	MME 20 e 200	1	148	-16.33%	-16.33%	-1.1169	0.00%	16.33%	16.33%	0.8854	100.00%
5%	MACD	18	251	-3.11%	-49.48%	-0.3987	38.89%	3.11%	56.95%	0.2348	61.11%
	MME 5 e 20	12	186	3.35%	17.75%	0.1347	33.33%	-3.35%	-58.69%	-0.1893	66.67%
	MME 20 e 200	1	148	-16.33%	-16.33%	-1.1169	0.00%	16.33%	16.33%	0.8854	100.00%
8%	MACD	18	251	-3.11%	-49.48%	-0.3987	38.89%	3.11%	56.95%	0.2348	61.11%
	MME 5 e 20	12	186	3.35%	17.75%	0.1347	33.33%	-3.35%	-58.69%	-0.1893	66.67%
	MME 20 e 200	1	148	-16.33%	-16.33%	-1.1169	0.00%	16.33%	16.33%	0.8854	100.00%
10%	MACD	18	251	-3.11%	-49.48%	-0.3987	38.89%	3.11%	56.95%	0.2348	61.11%
	MME 5 e 20	12	186	3.35%	17.75%	0.1347	33.33%	-3.35%	-58.69%	-0.1893	66.67%
	MME 20 e 200	1	148	-16.33%	-16.33%	-1.1169	0.00%	16.33%	16.33%	0.8854	100.00%

- MME 5 e 20 períodos: 15 operações lucrativas para PETR4 e VALE5. O maior e menor lucro acumulado para PETR4 foi de 43,58% e 14,03% e de 56,68% e 19,58% para VALE5;
- MME 20 e 200 períodos: 15 operações lucrativas para PETR4, cujo maior lucro acumulado foi de 36,02% e o menor 3,52%. Nenhuma dessas simulações resultou em saídas para VALE5.

Tabela 6.2: Resultados da simulação de compra e venda da VALE5 por análise técnica.

Vale5 - ANÁLISE TÉCNICA											
Objetivo	Estratégias	Quantidade operações	Dias Comprado	COMPRADO				Contra - tendência - COMPRADO			
				Lucro Médio por operação	Lucro Acumulado	Sharpe Ratio	% Operações Lurativas	Lucro Médio por operação	Lucro Acumulado	Sharpe Ratio	% Operações Lurativas
0%	MACD	17	210	-3.11%	-42.68%	-0.7946	17.65%	3.11%	65.19%	0.1877	82.35%
	MME 5 e 20	11	159	-1.77%	-18.48%	-0.8128	18.18%	1.77%	20.41%	-0.2218	81.82%
	MME 20 e 200	-	-	-	-	-	-	-	-	-	-
3%	MACD	17	210	-3.11%	-42.68%	-0.7946	17.65%	3.11%	65.19%	0.1877	82.35%
	MME 5 e 20	11	159	-1.77%	-18.48%	-0.8128	18.18%	1.77%	20.41%	-0.2218	81.82%
	MME 20 e 200	-	-	-	-	-	-	-	-	-	-
5%	MACD	17	210	-3.11%	-42.68%	-0.7946	17.65%	3.11%	65.19%	0.1877	82.35%
	MME 5 e 20	11	159	-1.77%	-18.48%	-0.8128	18.18%	1.77%	20.41%	-0.2218	81.82%
	MME 20 e 200	-	-	-	-	-	-	-	-	-	-
8%	MACD	17	210	-3.11%	-42.68%	-0.7946	17.65%	3.11%	65.19%	0.1877	82.35%
	MME 5 e 20	11	159	-1.77%	-18.48%	-0.8128	18.18%	1.77%	20.41%	-0.2218	81.82%
	MME 20 e 200	-	-	-	-	-	-	-	-	-	-
10%	MACD	17	210	-3.11%	-42.68%	-0.7946	17.65%	3.11%	65.19%	0.1877	82.35%
	MME 5 e 20	11	159	-1.77%	-18.48%	-0.8128	18.18%	1.77%	20.41%	-0.2218	81.82%
	MME 20 e 200	-	-	-	-	-	-	-	-	-	-
VENDIDO											
Objetivo	Estratégias	Quantidade operações	Dias Vendido	VENDIDO				Contra - tendência - VENDIDO			
				Lucro Médio por operação	Lucro Acumulado	Sharpe Ratio	% Operações Lurativas	Lucro Médio por operação	Lucro Acumulado	Sharpe Ratio	% Operações Lurativas
0%	MACD	17	210	0.70%	9.90%	0.1249	47.06%	-0.70%	-13.59%	-0.2819	52.94%
	MME 5 e 20	11	159	4.16%	46.68%	0.2965	46.68%	-4.16%	-44.69%	-0.4266	45.45%
	MME 20 e 200	-	-	-	-	-	-	-	-	-	-
3%	MACD	17	210	0.70%	9.90%	0.1249	47.06%	-0.70%	-13.59%	-0.2819	52.94%
	MME 5 e 20	11	159	4.16%	46.68%	0.2965	54.55%	-4.16%	-44.69%	-0.4266	45.45%
	MME 20 e 200	-	-	-	-	-	-	-	-	-	-
5%	MACD	17	210	0.70%	9.90%	0.1249	47.06%	-0.70%	-13.59%	-0.2819	52.94%
	MME 5 e 20	11	159	4.16%	46.68%	0.2965	54.55%	-4.16%	-44.69%	-0.4266	45.45%
	MME 20 e 200	-	-	-	-	-	-	-	-	-	-
8%	MACD	17	210	0.70%	9.90%	0.1249	47.06%	-0.70%	-13.59%	-0.2819	52.94%
	MME 5 e 20	11	159	4.16%	46.68%	0.2965	54.55%	-4.16%	-44.69%	-0.4266	45.45%
	MME 20 e 200	-	-	-	-	-	-	-	-	-	-
10%	MACD	17	210	0.70%	9.90%	0.1249	47.06%	-0.70%	-13.59%	-0.2819	52.94%
	MME 5 e 20	11	159	4.16%	46.68%	0.2965	54.55%	-4.16%	-44.69%	-0.4266	45.45%
	MME 20 e 200	-	-	-	-	-	-	-	-	-	-
COMPRADO/VENDIDO											
Objetivo	Estratégias	Quantidade operações	Dias Comprado/Vendido	COMPRADO/VENDIDO				Contra - tendência - COMPRADO/VENDIDO			
				Lucro Médio por operação	Lucro Acumulado	Sharpe Ratio	% Operações Lurativas	Lucro Médio por operação	Lucro Acumulado	Sharpe Ratio	% Operações Lurativas
0%	MACD	17	210	-2.34%	-37.01%	-0.2676	29.41%	2.34%	38.97%	0.1877	70.59%
	MME 5 e 20	11	159	2.28%	19.58%	0.1241	36.36%	-2.28%	-30.85%	-0.2218	63.64%
	MME 20 e 200	-	-	-	-	-	-	-	-	-	-
3%	MACD	17	210	-2.34%	-37.01%	-0.2676	29.41%	2.34%	38.97%	0.1877	70.59%
	MME 5 e 20	11	159	2.28%	19.58%	0.1241	36.36%	-2.28%	-30.85%	-0.22181	63.64%
	MME 20 e 200	-	-	-	-	-	-	-	-	-	-
5%	MACD	17	210	-2.34%	-37.01%	-0.2676	29.41%	2.34%	38.97%	0.1877	70.59%
	MME 5 e 20	11	159	2.28%	19.58%	0.1241	36.36%	-2.28%	-30.85%	-0.2218	63.64%
	MME 20 e 200	-	-	-	-	-	-	-	-	-	-
8%	MACD	17	210	-2.34%	-37.01%	-0.2676	29.41%	2.34%	38.97%	0.1877	70.59%
	MME 5 e 20	11	159	2.28%	19.58%	0.1241	36.36%	-2.28%	-30.85%	-0.2218	63.64%
	MME 20 e 200	-	-	-	-	-	-	-	-	-	-
10%	MACD	17	210	-2.34%	-37.01%	-0.2676	29.41%	2.34%	38.97%	0.1877	70.59%
	MME 5 e 20	11	159	2.28%	19.58%	0.1241	36.36%	-2.28%	-30.85%	-0.2218	63.64%
	MME 20 e 200	-	-	-	-	-	-	-	-	-	-

Tanto na tabela referente aos resultados da ação da Petrobrás quanto da Vale é possível observar a repetição de dados de saída para todos os objetivos de lucros disponíveis. Isso se deve ao fato de que o algoritmo atua observando a condição de compra e venda disponibilizada pela técnica de cruzamento de médias móveis adotada ou a condição de objetivo de lucro. Qualquer das condições sendo satisfeita, o algoritmo efetua a compra ou venda sinalizada. Por isso, os valores de indicadores disponibilizados na entrada produziram os mesmos valores na saída para quaisquer objetivos de lucros da entrada.

Quanto ao valor do *sharpe ratio*, nas simulações de análise técnica para PETR4, o maior valor obtido foi de 0,8854 operando comprado e também comprado/vendido, contratendência com ou sem objetivo de lucro para MME de 20 e 200 períodos. Para VALE5, o maior valor de *sharpe ratio* foi de 0,3186 obtido na condição de vendido, seguindo a tendência e sem objetivo de lucro para MME de 5 e 20 períodos.

Das operações de análise técnica simuladas, 44,45% para PETR4 e 45,00% para VALE5 foram lucrativas, respeitando as condições impostas pelo sistema de período de atuação no mercado, tipo de técnica adotada e levando em consideração apenas o preço de fechamento ajustado para cada ação.

6.4 Resultados para a simulação de análise da multidão

As Figuras 6.1 e 6.2 exibem as janelas de saídas de gráficos possíveis do simulador; a primeira para PETR4 e a segunda para VALE5. Em ambas, são plotadas as estratégias adotadas para efeito de comparação. As janelas apresentam quatro gráficos cada:

- o primeiro de ambas as figuras representa as linhas plotadas para médias móveis de 5, 20 e 200 períodos, bem como a curva do preço de fechamento ajustado para PETR4 (Figura 6.1) e VALE5 (Figura 6.2). As marcações com os caracteres "C" e "V" indicam os momentos para compra e venda detectados pelos indicadores do Twitter sem aplicação de limiar;
- O segundo gráfico da primeira figura mostra a linha do preço ajustado da ação, as linhas MME 5 e 20 períodos e as marcações com os caracteres "C" e "V" que apontam quando o cruzamento das médias móveis indicam compra e venda. O segundo gráfico da segunda figura exibe a linha do preço ajustado da ação e as marcações com os caracteres "C" e "V" que apontam quando o MADC indica compra e venda;
- O terceiro da primeira figura mostra a linha do preço ajustado da ação, as linhas MME 20 e 200 períodos e as marcações com os caracteres "C" e "V" que apontam quando o cruzamento das médias móveis indica compra e venda. O terceiro da segunda figura expõe o cruzamento das linhas MACD e Sinal para os dados de preços da VALE5;
- No quarto e último gráfico de ambas as Figuras 6.1 e 6.2 é possível observar o volume de tweets da ação para o período de simulação.

Outra janela de saída do sistema está exposta na Figura 6.3. Em (a) é apresentada a janela de saída do sistema para os resultados numéricos de cada operação da simulação de compra e venda da PETR4 por análise da multidão com limiar igual a 40 e objetivo de lucro de 10%. Nas colunas da janela estão dispostos para cada operação a data de compra e venda, os lucros obtidos LTC, LTV e LTCV que estão, respectivamente, seguindo à tendência comprado, vendido e comprado/vendido e LCTC, LCTV e LCTCV contra a tendência comprado, vendido e comprado/vendido. Em (b), os valores do *sharpe ratio* para a simulação total, ou seja, leva em consideração todas as operações realizadas para o cálculo.

O resumo das saídas numéricas para as sessenta simulações executadas baseadas exclusivamente em análise da multidão podem ser visualizadas nas Tabelas 6.3 para PETR4 e 6.4 para VALE5.

Das operações simuladas, 40% das PETR4 e 50% das VALE5 foram lucrativas. Nessa última, todas as operações que seguiram a tendência foram positivas. Tanto para PETR4 quanto para VALE5 todas as operações contratendência não geraram lucro. Este é um ponto muito interessante a ser observado nessas tabelas, pelos resultados obtidos percebe-se que há uma verdade embutida nas mensagens trafegadas na rede social Twitter sobre as ações investigadas. Isto é, os resultados apontam que o burburinho das redes sociais é verdadeiro, pois, na simulação, ao operar contratendência não há obtenção de lucros.

A média aritmética da quantidade de operações realizadas e dias em que o sistema permaneceu na condição de comprado, sem limiar para tweets, é de 62 operações em 199 dias para PETR4 e 73 em 273 dias para VALE5. Quando o limiar de tweets é aplicada a média passa para 42 operações em 167 dias para PETR4 e 61 em 265 dias para VALE5.

Em operações que não levaram em conta a quantidade de tweets trafegados no dia, ou seja, sem limiar, o maior lucro obtido foi de 38,29% para PETR4 e de 224,28% para VALE5. Ao aplicar o limiar, que só permite a realização de uma operação se a quantidade de tweets trafegada no dia for igual ou superior a este, o maior lucro alcançado para operações da PETR4 foi de 277,86% e de 195,37% para VALE5. Os maiores lucros obtidos em todas as simulações realizadas estão nessas tabelas de resultados para análise da multidão.

Comparando a obtenção de lucros de PETR4 e VALE5, a primeira alcançou maiores valores quando o limiar foi utilizado. Entretanto, a quantidade de tweets disponíveis para a PETR4 é maior do que os da VALE5, portanto seu limiar é maior e este, por sua vez, ajuda a reforçar o pensamento da maioria no dia avaliado sobre a ação em questão. Quando não há limiar, se apenas uma pessoa opina negativamente ou positivamente na rede social, esse sentimento único será o responsável pela tomada de decisão. Se há o limiar, só serão avaliados para compra ou venda os dias em que várias pessoas comentaram, dessa forma, a decisão será tomada levando-se em consideração o pensamento de várias pessoas.

No caso da VALE5, os maiores lucros foram realizados sem a adoção de limiar, no entanto, a diferença entre o valor do limiar utilizado em PETR4 para o da VALE5 é grande, e a diferença entre o uso e o não uso de limiar para a VALE5 é pequena.

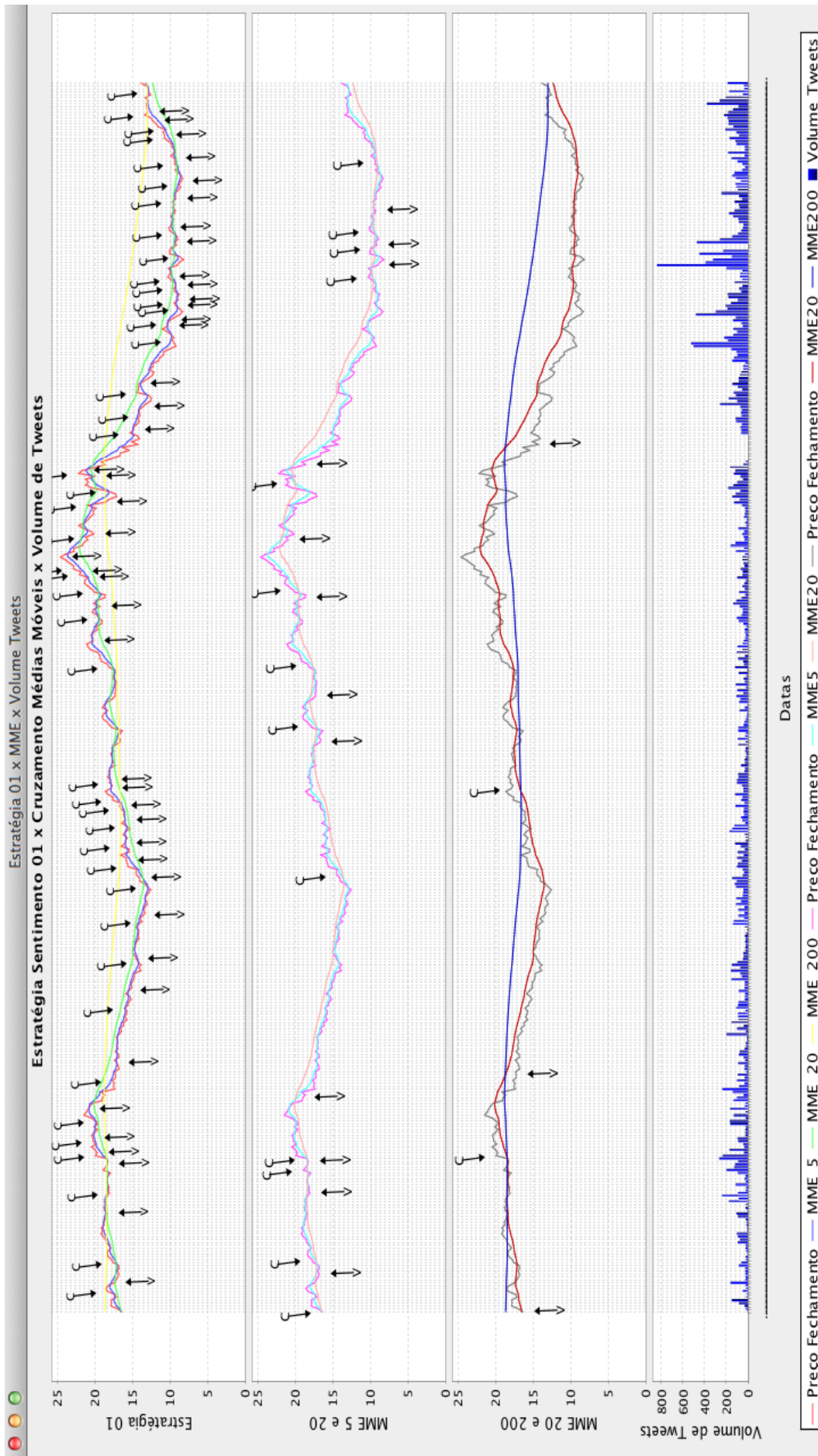


Figura 6.1: Janela de saídas gráficas do simulador para PETR4.

Os maiores valores obtidos para sharpe ratio, calculado em relação ao índice CDI da Tabela 5.1 e conforme explicado na Seção 5.5 do Capítulo 5, foram de 0,43 e 0,65. O primeiro para PETR4 em uma operação com limiar para tweets, atuando como COMPRADO/VENDIDO, seguindo a tendência e objetivando 10% de lucro, o segundo para VALE5 obtido na simulação de atuação COMPRADO/VENDIDO, seguindo a tendência e objetivando 3% de lucro.

Um valor também interessante em relação às simulações realizadas é a porcentagem de operações lucrativas, próxima dos 50% para a análise da multidão.

6.5 Resultados para a simulação de análise técnica com análise da multidão

As Tabelas 6.5, 6.6, 6.7 e 6.8 apresentam as simulações de análise técnica com análise da multidão, definidas na Seção 5.4.2 do capítulo anterior, para compra e venda de ações da PETR4 e VALE5. Nas subseções seguintes serão comentados os resultados alcançados.

6.5.1 Twitter com Convergência/Divergência de médias móveis - MACD

Os valores apresentados nas Tabelas 6.5 e 6.6 apresentam as simulações executadas para PETR4 e VALE5 baseadas nos indicadores obtidos do Twitter e no modelo MACD da análise técnica. De acordo com as definições para essa modalidade de simulação apresentadas no Capítulo 5, uma compra ou venda só é efetivada quando há uma indicação do MACD e esta é confirmada pelos indicadores do Twitter em um dos três dias t (o dia em questão), $t-1$ ou $t-2$.

Das sessenta simulações realizadas para PETR4, apenas 20% foram lucrativas. Um resultado positivo foi obtido ao seguir a tendência, os demais foram alcançados na contratendência. O maior valor de *sharpe ratio* foi de 0,5 para duas simulações com maior lucro acumulado: 100,86% e 162,31%. O primeiro desses foi obtido quando o simulador atuou como COMPRADO e o segundo como COMPRADO/VENDIDO, ambos contratendência, sem objetivo de lucro e com aplicação de limiar para tweets. A maior parte das operações lucrativas para PETR4 foram realizadas com aplicação de limiar para tweets.

Os resultados obtidos para VALE5 foram bem diferentes dos da PETR4 tanto as operações realizadas com limiar como as sem limiar alcançaram lucros. Das sessenta simulações realizadas para MACD com Twitter, 46% delas obtiveram lucro. O maior valor de *sharpe ratio* foi de 0,31 e foi obtido quando lucros acumulados de 77,37% e 62,79% foram alcançados. Esses são dois dos três maiores lucros acumulados obtidos para a VALE5. A média aritmética da quantidade de operações realizadas e dias em que o simulador permaneceu comprado para operações sem limiar de tweets foi de 18 operações em 202 dias, com limiar passa para 16 operações realizadas em 201 dias.

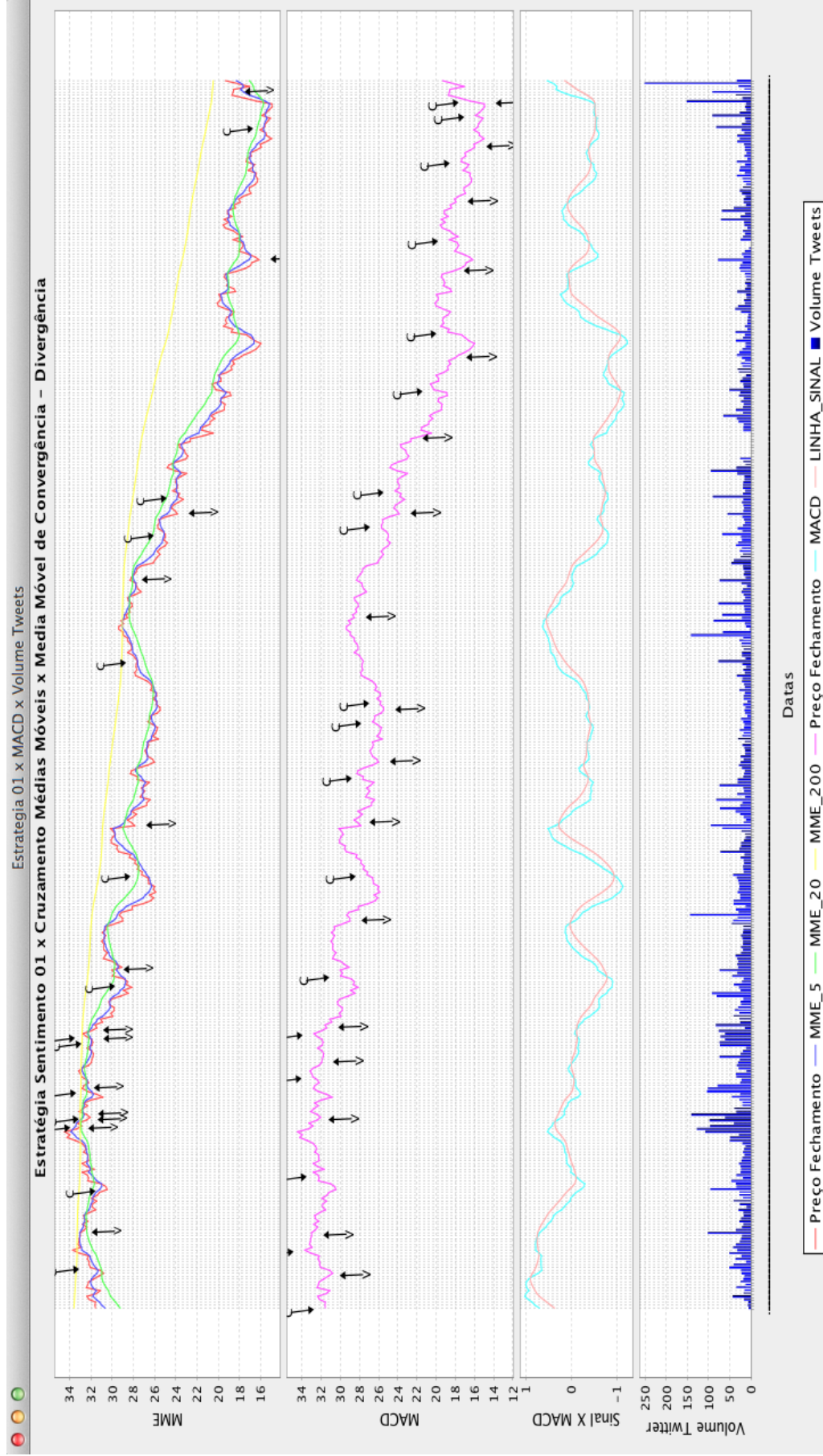


Figura 6.2: Janela de saídas gráficas do simulador para VALE5.

ESTRATÉGIA OBJETIVO TWITTER										
DATA COMPRA	DATA VENDA	LTC	LTV	LTCV	LCTC	LCTV	LCTCV	PREÇO COMPRA	PREÇO VENDA	QDE DIAS COMPRADO
2013-08-22	2013-08-27	-0.044...	-0.005...	-0.049...	0.0443...	0.0051...	0.0492...	18.26	17.45	4
2013-09-05	2013-09-30	0.0467...	0.0160...	0.0635...	-0.046...	-0.016...	-0.063...	17.54	18.36	18
2013-10-09	2013-10-23	0.0143...	-0.078...	-0.065...	-0.014...	0.0784...	0.0651...	18.07	18.33	11
2013-10-28	2013-11-01	-0.001...	-0.017...	-0.019...	0.0015...	0.0178...	0.0192...	19.89	19.86	5
2013-11-04	2013-11-05	0.0014...	0.0155...	0.0170...	-0.001...	-0.015...	-0.017...	20.22	20.25	2
2013-11-13	2013-11-21	0.0481...	0.1854...	0.2425...	-0.048...	-0.185...	-0.242...	19.94	20.9	5
2013-12-05	2013-12-13	-0.034...	0.0617...	0.0255...	0.0340...	-0.061...	-0.025...	17.63	17.03	7
2014-01-15	2014-01-23	-0.035...	0.0886...	0.0499...	0.0355...	-0.088...	-0.049...	16.04	15.47	7
2014-02-06	2014-02-07	0.0168...	0.0562...	0.0741...	-0.016...	-0.056...	-0.074...	14.21	14.45	2
2014-02-26	2014-02-28	-0.006...	0.0478...	0.0409...	0.0065...	-0.047...	-0.040...	13.68	13.59	3
2014-03-18	2014-03-21	0.0809...	-0.099...	-0.026...	-0.080...	0.0995...	0.0266...	12.97	14.02	4
2014-03-27	2014-03-31	0.0134...	-0.041...	-0.028...	-0.013...	0.0413...	0.0283...	15.57	15.78	3
2014-04-07	2014-04-08	-0.028...	0.0133...	-0.015...	0.0285...	-0.013...	0.0156...	16.46	15.99	2
2014-04-16	2014-04-22	-0.0114...	-0.093...	-0.082...	-0.011...	0.0931...	0.0828...	15.78	15.96	3
2014-05-02	2014-05-08	0.0159...	-0.008...	0.0074...	-0.015...	0.0083...	-0.007...	17.6	17.88	5
2014-05-12	2014-05-13	-0.003...	-0.008...	-0.012...	0.0038...	0.0082...	0.0121...	18.03	17.96	2
2014-07-10	2014-07-23	0.1187...	0.0416...	0.1653...	-0.118...	-0.041...	-0.165...	18.11	20.26	10
2014-08-04	2014-08-08	-0.007...	-0.037...	-0.044...	0.0071...	0.0373...	0.0443...	19.45	19.31	5
2014-08-15	2014-08-22	0.0428...	-0.050...	-0.010...	-0.042...	0.0508...	0.0101...	20.06	20.92	6
2014-08-25	2014-08-26	-0.009...	-0.043...	-0.052...	0.0090...	0.0437...	0.0524...	22.04	21.84	2
2014-08-27	2014-09-02	0.0753...	0.1728...	0.2611...	-0.075...	-0.172...	-0.261...	22.84	24.56	5
2014-09-26	2014-09-29	-0.111...	0.0136...	-0.099...	0.1117...	-0.013...	0.0996...	20.94	18.6	2
2014-10-03	2014-10-10	0.0910...	-0.095...	-0.013...	-0.091...	0.0953...	0.0130...	18.35	20.02	6
2014-10-13	2014-10-14	-0.021...	0.4168...	0.3861...	0.0216...	-0.416...	-0.386...	22.13	21.65	2
2014-10-31	2014-11-03	-0.028...	0.0622...	0.0323...	0.0281...	-0.062...	-0.032...	15.28	14.85	2
2014-11-10	2014-11-13	-0.027...	-0.048...	-0.074...	0.0271...	0.0489...	0.0748...	13.98	13.6	4
2014-11-21	2014-11-26	-0.013...	0.4596...	0.4392...	0.0139...	-0.459...	-0.439...	14.3	14.1	4
2014-12-17	2014-12-26	0.0662...	0.0019...	0.0683...	-0.066...	-0.001...	-0.068...	9.66	10.3	6
2014-12-29	2014-12-30	-0.025...	0.1557...	0.1264...	0.0252...	-0.155...	-0.126...	10.28	10.02	2
2015-01-07	2015-01-08	0.0588...	-0.023...	0.0340...	-0.058...	0.0234...	-0.034...	8.67	9.18	2
2015-01-09	2015-01-12	-0.052...	-0.056...	-0.105...	0.0521...	0.0561...	0.1053...	9.4	8.91	2
2015-01-16	2015-01-19	-0.026...	-0.064...	-0.088...	0.0264...	0.0641...	0.0889...	9.44	9.19	2
2015-01-21	2015-01-22	0.0437...	0.1836...	0.2354...	-0.043...	-0.183...	-0.235...	9.82	10.25	2
2015-02-02	2015-02-09	0.0715...	-0.023...	0.0467...	-0.071...	0.0231...	-0.046...	8.66	9.28	6
2015-02-12	2015-02-18	0.0652...	0.0541...	0.1229...	-0.065...	-0.054...	-0.122...	9.5	10.12	3
2015-03-03	2015-03-04	-0.040...	0.0477...	0.0052...	0.0406...	-0.047...	-0.005...	9.6	9.21	2
2015-03-11	2015-03-12	-0.032...	-0.090...	-0.120...	0.0329...	0.0909...	0.1209...	8.79	8.5	2
2015-03-20	2015-03-24	0.0042...	-0.124...	-0.120...	-0.004...	0.1240...	0.1203...	9.35	9.39	3
2015-04-02	2015-04-06	-0.002...	0.0084...	0.0056...	0.0027...	-0.008...	-0.005...	10.72	10.69	2
2015-04-08	2015-04-13	0.1575...	-0.079...	0.0654...	-0.157...	0.0795...	-0.065...	10.6	12.27	4
2015-04-15	2015-04-16	-0.030...	0.0117...	-0.018...	0.0300...	-0.011...	0.0186...	13.33	12.93	2
2015-04-28	?	0.0	0	0	-0.0	-0.0	-0.0	12.78	0.0	5

(a)

Sharp Ratio					
SRTC	SRTV	SRTC	SRTC	SRTC	SRTC
0.270649631149149	0.291001428944013...	0.433965848572906...	-0.5429453959935372	-0.4067147367611742	-0.5429453959935372

(b)

Figura 6.3: Janelas de saídas numéricas do sistema para análise da multidão PETR4 com limiar 40 e objetivo de lucro de 10%, (a) saídas por operação e (b) *sharpe ratio* de toda a simulação.

Tabela 6.3: Resultado da simulação por análise da multidão
- Twitter para PETR4 com e sem limiar de tweets.

PETR4 - ANÁLISE DA MULTIDÃO - TWITTER													
Ob. Lucro	Medições	Limiar=0						Limiar=40					
		Tendência			Contra-tendência			Tendência			Contra-tendência		
		Comprado	Vendido	C/Vendido	Comprado	Vendido	C/Vendido	Comprado	Vendido	C/Vendido	Comprado	Vendido	C/Vendido
0%	Quantidade Operações	62	62	62	62	62	62	41	41	41	41	41	41
	Dias Comprado	205	205	205	205	205	205	171	171	171	171	171	171
	Lucro Médio/Op	-0.12%	0.61%	0.46%	0.12%	-0.61%	-0.46%	1.12%	2.50%	3.58%	-1.12%	-2.50%	-3.58%
	Lucro Acumulado	-15.46%	17.94%	-0.29%	-1.12%	-49.33%	-48.18%	50.11%	114.47%	221.94%	-40.58%	-76.01%	-85.13%
	Sharpe Ratio	-0.1225	0.0604	0.0054	-0.1770	-0.1883	-0.1770	0.2260	0.2687	0.4127	-0.5279	-0.3928	-0.5279
%Op. Lucrativas	40.32%	43.55%	46.77%	58.06%	56.45%	53.23%	48.78%	51.22%	53.66%	51.22%	48.78%	46.34%	
3%	Quantidade Operações	64	64	64	64	64	64	44	44	44	44	44	44
	Dias Comprado	197	197	197	197	197	197	164	164	164	164	164	164
	Lucro Médio/Op	0.07%	0.76%	0.82%	-0.07%	-0.76%	-0.82%	1.12%	2.37%	3.44%	-1.12%	-2.37%	-3.44%
	Lucro Acumulado	-5.79%	31.43%	23.81%	-13.98%	-54.28%	-59.92%	53.11%	118.77%	234.97%	-43.20%	-76.44%	-85.96%
	Sharpe Ratio	-0.0490	0.0977	0.0884	-0.2201	-0.2153	-0.2201	0.2201	0.2697	0.3938	-0.4962	-0.3775	-0.4962
%Op. Lucrativas	42.19%	45.31%	50.00%	54.69%	54.69%	50.00%	50.00%	47.73%	54.55%	50.00%	52.27%	45.45%	
5%	Quantidade Operações	62	62	62	62	62	62	42	42	42	42	42	42
	Dias Comprado	197	197	197	197	197	197	164	164	164	164	164	164
	Lucro Médio/Op	0.16%	0.86%	1.01%	-0.16%	-0.86%	-1.01%	1.40%	2.64%	4.01%	-1.40%	-2.64%	-4.01%
	Lucro Acumulado	-0.88%	38.29%	37.07%	-18.39%	-56.49%	-64.07%	66.08%	127.51%	277.86%	-47.68%	-77.31%	-87.81%
	Sharpe Ratio	-0.0305	0.1140	0.1195	-0.2628	-0.2385	-0.2628	0.2644	0.2885	0.4298	-0.5393	-0.4049	-0.5393
%Op. Lucrativas	40.32%	46.77%	50.00%	58.06%	53.23%	50.00%	48.78%	48.78%	56.10%	51.22%	51.22%	43.90%	
8%	Quantidade Operações	62	62	62	62	62	62	42	42	42	42	42	42
	Dias Comprado	198	198	198	198	198	198	168	168	168	168	168	168
	Lucro Médio/Op	0.15%	0.86%	1.00%	-0.15%	-0.86%	-1.00%	1.28%	2.60%	3.85%	-1.28%	-2.60%	-3.85%
	Lucro Acumulado	-1.16%	37.90%	36.30%	-18.13%	-56.38%	-63.85%	60.22%	128.93%	266.79%	-45.90%	-77.51%	-87.54%
	Sharpe Ratio	-0.0319	0.1131	0.1177	-0.2607	-0.2373	-0.2607	0.2607	0.2862	0.4248	-0.5347	-0.4020	-0.5347
%Op. Lucrativas	40.32%	46.77%	50.00%	58.06%	53.23%	50.00%	47.62%	50.00%	54.76%	52.38%	50.00%	45.24%	
10%	Quantidade Operações	62	62	62	62	62	62	41	41	41	41	41	41
	Dias Comprado	200	200	200	200	200	200	168	168	168	168	168	168
	Lucro Médio/Op	0.11%	0.82%	0.93%	-0.11%	-0.82%	-0.93%	1.34%	2.69%	4.01%	-1.34%	-2.69%	-4.01%
	Lucro Acumulado	-3.40%	34.77%	30.19%	-15.99%	-55.45%	-62.11%	62.43%	132.07%	276.95%	-46.33%	-77.81%	-87.87%
	Sharpe Ratio	-0.0430	0.1056	0.1031	-0.2437	-0.2280	-0.2437	0.2706	0.2910	0.4340	-0.5429	-0.4067	-0.5429
%Op. Lucrativas	40.32%	46.77%	48.39%	58.06%	53.23%	51.61%	48.78%	51.22%	56.10%	51.22%	48.78%	43.90%	

Tabela 6.4: Resultado da simulação por análise da multidão
- Twitter para VALE5 com e sem limiar de tweets.

VALES - ANÁLISE DA MULTIDÃO - TWITTER													
Ob. Lucro	Medições	Limiar=0						Limiar=15					
		TENDÊNCIA			Contra-tendência			tendência			Contra-tendência		
		Comprado	Vendido	C/Vendido	Comprado	Vendido	C/Vendido	Comprado	Vendido	C/Vendido	Comprado	Vendido	C/Vendido
0%	Quantidade Operações	71	71	71	71	71	71	60	60	60	60	60	60
	Dias Comprado	277	277	277	277	277	277	269	269	269	269	269	269
	Lucro Médio/Op	0.51%	1.19%	1.70%	-0.51%	-1.19%	-1.70%	0.40%	1.22%	1.67%	-0.40%	-1.22%	-1.67%
	Lucro Acumulado	25.93%	119.71%	196.33%	-35.51%	-59.41%	-74.01%	21.23%	97.49%	139.42%	-25.50%	-54.58%	-68.87%
	Sharpe Ratio	0.1201	0.5400	0.5333	-0.7744	-0.8670	-0.7744	0.0420	0.4309	0.4167	-0.6530	-0.7185	-0.6530
%Op. Lucrativas	47.89%	53.52%	53.52%	49.30%	46.48%	46.48%	48.33%	58.33%	50.00%	50.00%	41.67%	50.00%	
3%	Quantidade Operações	78	78	78	78	78	78	65	65	65	65	65	65
	Dias Comprado	259	259	259	259	259	259	253	253	253	253	253	253
	Lucro Médio/Op	0.51%	1.16%	1.68%	-0.51%	-1.16%	-1.68%	0.51%	1.32%	1.85%	-0.51%	-1.32%	-1.85%
	Lucro Acumulado	41.09%	129.84%	224.28%	-37.44%	-62.81%	-77.23%	34.65%	119.35%	195.37%	-30.95%	-60.73%	-73.75%
	Sharpe Ratio	0.1693	0.6057	0.6588	-0.9142	-0.9457	-0.9142	0.1376	0.5382	0.5937	-0.8473	-0.8285	-0.8473
%Op. Lucrativas	47.44%	51.28%	52.56%	50.00%	47.44%	47.44%	47.69%	55.38%	50.77%	50.77%	43.08%	49.23%	
5%	Quantidade Operações	74	74	74	74	74	74	62	62	62	62	62	62
	Dias Comprado	275	275	275	275	275	275	269	269	269	269	269	269
	Lucro Médio/Op	0.36%	1.02%	1.38%	-0.36%	-1.02%	-1.38%	0.31%	1.11%	1.42%	-0.31%	-1.11%	-1.42%
	Lucro Acumulado	23.10%	100.53%	146.85%	-28.99%	-55.70%	-68.56%	16.06%	89.06%	119.43%	-20.64%	-52.63%	-62.81%
	Sharpe Ratio	0.0545	0.4645	0.4793	-0.7395	-0.8005	-0.7395	0.0023	0.4159	0.4300	-0.7283	-0.7218	-0.7283
%Op. Lucrativas	45.95%	51.35%	50.00%	51.35%	48.65%	50.00%	48.39%	56.45%	50.00%	50.00%	43.55%	50.00%	
8%	Quantidade Operações	72	72	72	72	72	72	61	61	61	61	61	61
	Dias Comprado	277	277	277	277	277	277	269	269	269	269	269	269
	Lucro Médio/Op	0.44%	1.11%	1.54%	-0.44%	-1.11%	-1.54%	0.32%	1.14%	1.46%	-0.32%	-1.14%	-1.46%
	Lucro Acumulado	28.99%	110.13%	171.06%	-32.06%	-57.64%	-71.22%	16.46%	89.71%	120.93%	-20.92%	-52.79%	-63.07%
	Sharpe Ratio	0.0900	0.5211	0.5416	-0.7975	-0.8722	-0.7975	0.0055	0.4189	0.4330	-0.7302	-0.7247	-0.7302
%Op. Lucrativas	47.22%	52.78%	52.78%	50.00%	47.22%	47.22%	49.18%	57.38%	50.82%	49.18%	42.62%	49.18%	
10%	Quantidade Operações	72	72	72	72	72	72	61	61	61	61	61	61
	Dias Comprado	277	277	277	277	277	277	269	269	269	269	269	269
	Lucro Médio/Op	0.44%	1.11%	1.54%	-0.44%	-1.11%	-1.54%	0.32%	1.14%	1.46%	-0.32%	-1.14%	-1.46%
	Lucro Acumulado	28.99%	110.13%	171.06%	-32.06%	-57.64%	-71.22%	16.46%	89.71%	120.93%	-20.92%	-52.79%	-63.07%
	Sharpe Ratio	0.090	0.521	0.542	-0.798	-0.872	-0.798	0.006	0.419	0.433	-0.730	-0.725	-0.730
%Op. Lucrativas	47.22%	52.78%	52.78%	50.00%	47.22%	47.22%	49.18%	57.38%	50.82%	49.18%	42.62%	49.18%	

6.5.2 Twitter com cruzamento de Médias Móveis Exponenciais - MME

As Tabelas 6.7 e 6.8 apresentam os resultados das simulações de compra e venda de ações a partir de indicadores do Twitter para o dia t ou $t - 1$ e $t - 2$ e cruzamento de médias móveis exponenciais de 5 e 20 períodos.

Das simulações realizadas para PETR4, 46,6% foram lucrativas. O maior lucro acumulado alcançado foi de 91% para uma operação que seguiu a tendência atuando como COMPRADO/VENDIDO e com objetivo de 10% de lucro. O maior valor de *sharpe ratio* de toda a tabela da PETR4 também foi observado nesta operação 0,31.

Em relação aos resultados da VALE5, de todas as operações realizadas para MME com Twitter 55% foram lucrativas. Os dois maiores valores para lucro acumulado alcançado foram de 74,68% e 70,22%, ambos obtidos em uma simulação que seguiu a tendência como COMPRADO e com objetivo de 3% de lucro, sendo o primeiro com, e o segundo sem aplicação de limiar para tweets. Os maiores valores para *sharpe ratio* da tabela também foram obtidos nessas operações 0,43 na primeira e 0,39 na segunda.

Em média, o sistema realizou 15 operações em 180 dias para PETR4 e 12 operações em 145 dias para a VALE5 sem a adoção de limiar para tweets. Com o limiar em média foram realizadas 11 operações em 178 dias e 12 em 145, respectivamente para PETR4 e VALE5.

Para a ação da Petrobrás, as simulações de Twitter com MME renderam mais operações lucrativas do que a técnica Twitter com MACD, entretanto, o maior lucro acumulado foi obtido na segunda técnica.

6.6 Comentários sobre os dados

Durante a realização de vários testes com os dados de entrada percebeu-se vários detalhes que podem a partir de agora explorados. O primeiro a ser comentado é o valor da variável Limiar. Nas simulações acima foram adotados os valores de zero e $(\frac{1}{n} \sum_{t=1}^n B_t)/2$ com n sendo a quantidade total de amostras. Esses valores foram tomados aleatoriamente, mas enquanto os testes eram realizados, verificou-se a possibilidade de empregar outros valores.

Entretanto, percebeu-se que para a forma como os dados se encontravam, o valor adotado inicialmente produzia resultados satisfatórios. Porém, é possível que outros valores para o Limiar ajudem a gerar melhores resultados e, portanto, é um ponto a ser explorado.

Outro detalhe interessante é a limpeza dos dados realizada no estágio de pré-processamento - Figura 3.1 do Capítulo 5. No decorrer da pesquisa, descobriu-se que à medida que se removiam tweets da base de dados, a partir de rotinas de limpeza mais severas, menores eram os lucros obtidos nas simulações para análise da multidão. Para averiguar tal percepção, foram realizados alguns testes com a mesma base de dados descrita na Seção 6.2 deste capítulo e utilizada nos testes relatados acima.

Tabela 6.5: Resultado da simulação MACD com Twitter para PETR4.

PETR4 - TWITTER (t,t-1,t-2) com MACD													
Ob. Lucro	Medições	Limiar=0						Limiar=40					
		Tendência			Contra-tendência			Tendência			Contra-tendência		
		Comprado	Vendido	C/Vendido	Comprado	Vendido	C/Vendido	Comprado	Vendido	C/Vendido	Comprado	Vendido	C/Vendido
0%	Quantidade Operações	19	18	18	19	18	18	16	16	16	16	16	16
	Dias Comprado	248	247	247	248	247	247	216	216	216	216	216	216
	Lucro Médio/Op	-1.58%	-0.15%	-4.14%	1.58%	0.15%	4.14%	-4.99%	-1.47%	-6.69%	4.99%	1.47%	6.69%
	Lucro Acumulado	-35.91%	-10.87%	-58.34%	13.81%	-8.67%	87.76%	-60.47%	-24.83%	-70.29%	100.86%	20.25%	162.31%
	Sharpe Ratio	-0.2005	-0.0954	-0.5063	0.3404	-0.0630	0.3404	-0.5108	-0.3861	-0.6610	0.5046	0.1167	0.5046
%Op. Lucrativas	36.84%	33.33%	33.33%	63.16%	66.67%	66.67%	31.25%	43.75%	31.25%	62.50%	56.25%	68.75%	
3%	Quantidade Operações	28	27	27	28	27	27	23	22	22	23	22	22
	Dias Comprado	210	210	210	210	210	210	219	218	218	219	218	218
	Lucro Médio/Op	-0.09%	0.93%	0.83%	0.09%	-0.93%	-0.83%	-0.73%	-0.31%	-1.41%	0.73%	0.31%	1.41%
	Lucro Acumulado	-17.57%	14.64%	-5.79%	-16.88%	-33.39%	-43.47%	-31.88%	-12.23%	-44.63%	0.22%	0.22%	-0.60%
	Sharpe Ratio	-0.0915	0.0444	-0.0068	-0.1711	-0.1955	-0.1711	-0.1149	-0.2016	-0.1747	0.0484	-0.0800	0.0484
%Op. Lucrativas	46.43%	37.04%	40.74%	53.57%	62.96%	59.26%	39.13%	45.45%	45.45%	56.52%	54.55%	54.55%	
5%	Quantidade Operações	25	24	24	25	24	24	22	21	21	22	21	21
	Dias Comprado	233	232	232	233	232	232	225	224	224	225	224	224
	Lucro Médio/Op	-0.72%	0.30%	-0.53%	0.72%	-0.30%	0.53%	-0.79%	-0.35%	-1.55%	0.79%	0.35%	1.55%
	Lucro Acumulado	-29.88%	-4.79%	-31.83%	-3.20%	-18.46%	-16.81%	-32.37%	-12.87%	-45.43%	-6.09%	0.95%	1.02%
	Sharpe Ratio	-0.1588	-0.0458	-0.1473	-0.0458	-0.1228	-0.0458	-0.1190	-0.2077	-0.1817	0.0533	-0.0735	0.0533
%Op. Lucrativas	36.00%	33.33%	37.50%	64.00%	66.67%	62.50%	36.36%	42.86%	42.86%	59.09%	57.14%	57.14%	
8%	Quantidade Operações	21	20	20	21	20	20	21	20	20	21	20	20
	Dias Comprado	227	226	226	227	226	226	227	226	226	227	226	226
	Lucro Médio/Op	-0.72%	0.30%	-0.53%	0.72%	-0.30%	0.53%	-0.66%	-0.19%	-1.28%	0.66%	0.19%	1.28%
	Lucro Acumulado	-29.88%	-2.48%	-31.83%	-3.20%	-18.46%	-16.81%	-29.77%	-9.52%	-41.15%	-9.49%	-2.65%	-5.44%
	Sharpe Ratio	-0.1588	-0.0458	-0.1473	-0.0458	-0.1228	-0.0458	-0.1037	-0.1837	-0.1569	0.0300	-0.1120	0.0300
%Op. Lucrativas	36.00%	33.33%	37.50%	64.00%	66.67%	62.50%	38.10%	45.00%	45.00%	57.14%	55.00%	55.00%	
10%	Quantidade Operações	25	24	24	25	24	24	20	20	20	20	20	20
	Dias Comprado	236	235	235	236	235	235	228	228	228	228	228	228
	Lucro Médio/Op	-0.77%	0.28%	-0.58%	0.77%	-0.28%	0.58%	-1.37%	-0.43%	-1.77%	1.37%	0.43%	1.77%
	Lucro Acumulado	-30.75%	-3.69%	-33.52%	-2.10%	-18.49%	-17.79%	-38.44%	-14.36%	-47.28%	4.34%	1.23%	2.09%
	Sharpe Ratio	-0.1694	-0.0503	-0.1612	-0.0475	-0.1182	-0.0475	-0.1648	-0.2166	-0.2066	0.0622	-0.0611	0.0622
%Op. Lucrativas	36.00%	33.33%	33.33%	64.00%	66.67%	66.67%	35.00%	45.00%	35.00%	60.00%	55.00%	65.00%	

Tabela 6.6: Resultado da simulação MACD com Twitter para VALE5 .

VALE5 - TWITTER(t,t-1,t-2) com MACD													
Ob. Lucro	Medições	Limiar=0						Limiar=15					
		Tendência			Contra-tendência			Tendência			Contra-tendência		
		Comprado	Vendido	C/Vendido	Comprado	Vendido	C/Vendido	Comprado	Vendido	C/Vendido	Comprado	Vendido	C/Vendido
0%	Quantidade Operações	17	17	17	17	17	17	16	16	16	16	16	16
	Dias Comprado	210	210	210	210	210	210	206	206	206	206	206	206
	Lucro Médio/Op	-3.54%	0.21%	-3.26%	3.54%	-0.21%	3.26%	-3.26%	0.95%	-2.26%	3.26%	-0.95%	2.26%
	Lucro Acumulado	-46.92%	1.39%	-46.18%	77.37%	-5.62%	62.79%	-42.22%	14.39%	-33.91%	64.18%	-15.70%	35.93%
	Sharpe Ratio	-0.8779	-0.1176	-0.5111	0.3135	-0.2022	0.3135	-0.8276	0.0376	-0.4008	0.1960	-0.3787	0.1960
%Op. Lucrativas	17.65%	47.06%	29.41%	82.35%	52.94%	70.59%	18.75%	50.00%	31.25%	81.25%	43.75%	68.75%	
3%	Quantidade Operações	20	20	20	20	20	20	18	18	18	18	18	18
	Dias Comprado	198	198	198	198	198	198	194	194	194	194	194	194
	Lucro Médio/Op	-2.47%	0.71%	-1.68%	2.47%	-0.71%	1.68%	-2.41%	1.33%	-1.02%	2.41%	-1.33%	1.02%
	Lucro Acumulado	-41.26%	12.20%	-34.09%	58.01%	-15.64%	28.58%	-37.25%	24.24%	-22.04%	49.61%	-23.31%	11.88%
	Sharpe Ratio	-0.6252	0.0070	-0.2759	0.1275	-0.2997	0.1275	-0.5855	0.1397	-0.1885	0.0467	-0.4416	0.0467
%Op. Lucrativas	30.00%	50.00%	35.00%	70.00%	50.00%	65.00%	27.78%	50.00%	33.33%	72.22%	44.44%	66.67%	
5%	Quantidade Operações	19	19	19	19	19	19	17	17	17	17	17	17
	Dias Comprado	199	199	199	199	199	199	202	202	202	202	202	202
	Lucro Médio/Op	-2.58%	0.78%	-1.71%	2.58%	-0.78%	1.71%	-2.55%	1.42%	-1.06%	2.55%	-1.42%	1.06%
	Lucro Acumulado	-40.94%	12.82%	-33.36%	57.73%	-16.10%	27.10%	-37.18%	24.38%	-21.86%	49.97%	-23.40%	11.50%
	Sharpe Ratio	-0.6201	0.0133	-0.2684	0.1202	-0.3036	0.1202	-0.5849	0.1403	-0.1859	0.0446	-0.4391	0.0446
%Op. Lucrativas	26.32%	52.63%	36.84%	73.68%	47.37%	63.16%	23.53%	52.94%	35.29%	76.47%	41.18%	66.67%	
8%	Quantidade Operações	18	18	18	18	18	18	16	16	16	16	16	16
	Dias Comprado	201	201	201	201	201	201	202	202	202	202	202	202
	Lucro Médio/Op	-2.69%	0.84%	-1.70%	2.69%	-0.84%	1.70%	-2.72%	1.48%	-1.12%	2.72%	-1.48%	1.12%
	Lucro Acumulado	-40.73%	13.21%	-32.90%	56.19%	-16.44%	22.75%	-37.45%	23.84%	-22.53%	49.68%	-23.07%	9.42%
	Sharpe Ratio	-0.6215	0.0167	-0.2694	0.0990	-0.3110	0.0990	-0.5970	0.1361	-0.1965	0.0346	-0.4389	0.0346
%Op. Lucrativas	22.22%	50.00%	33.33%	77.78%	50.00%	66.67%	18.75%	50.00%	31.25%	81.25%	43.75%	68.75%	
10%	Quantidade Operações	17	17	17	17	17	17	16	16	16	16	16	16
	Dias Comprado	206	206	206	206	206	206	203	203	203	203	203	203
	Lucro Médio/Op	-3.03%	0.72%	-2.16%	3.03%	-0.72%	2.16%	-2.66%	1.54%	-0.99%	2.66%	-1.54%	0.99%
	Lucro Acumulado	-42.43%	9.96%	-36.69%	61.37%	-13.98%	30.67%	-36.90%	24.92%	-21.18%	48.10%	-23.86%	6.63%
	Sharpe Ratio	-0.6459	-0.0141	-0.2969	0.1376	-0.2744	0.1376	-0.5746	0.1455	-0.1783	0.0206	-0.4439	0.0206
%Op. Lucrativas	17.65%	47.06%	29.41%	82.35%	52.94%	70.59%	18.75%	50.00%	31.25%	81.25%	43.75%	68.75%	

Tabela 6.7: Resultado da simulação MME de 5 e 20 períodos com Twitter para PETR4.

PETR4 - TWITTER(t,t-1,t-2) com MME 5 E 20													
Ob Lucro	Medições	Limiar=0						Limiar=40					
		Tendência			Contra-tendência			Tendência			Contra-tendência		
		Comprado	Vendido	C/Vendido	Comprado	Vendido	C/Vendido	Comprado	Vendido	C/Vendido	Comprado	Vendido	C/Vendido
0%	Quantidade Operações	11	11	11	11	11	11	9	9	9	9	9	9
	Dias Comprado	178	178	178	178	178	178	176	176	176	176	176	176
	Lucro Médio/Op	-2.84%	5.31%	1.55%	2.84%	-5.31%	-1.55%	-3.20%	6.80%	2.53%	3.20%	-6.80%	-2.53%
	Lucro Acumulado	-29.44%	34.55%	-5.06%	31.72%	-79.04%	-45.94%	-27.32%	38.21%	0.45%	29.36%	-79.64%	-48.92%
	Sharpe Ratio	-0.4683	0.1790	0.0405	-0.1015	-0.2377	-0.1015	-0.7233	0.2205	0.0737	-0.1457	-0.2657	-0.1457
%Op. Lucrativas	36.36%	27.27%	36.36%	63.64%	72.73%	63.64%	22.22%	22.22%	33.33%	77.78%	77.78%	66.67%	
3%	Quantidade Operações	18	18	18	18	18	18	13	13	13	13	13	13
	Dias Comprado	174	174	174	174	174	174	176	176	176	176	176	176.00
	Lucro Médio/Op	0.52%	3.51%	3.81%	-0.52%	-3.51%	-3.81%	1.86%	5.79%	7.09%	-1.86%	-5.79%	-7.09%
	Lucro Acumulado	-1.94%	36.81%	34.15%	-22.18%	-81.39%	-76.25%	15.27%	60.36%	84.85%	-32.18%	-82.21%	-78.03%
	Sharpe Ratio	-0.0137	0.1476	0.1582	-0.2124	-0.2103	-0.2124	0.0841	0.2160	0.2844	-0.3661	-0.2840	-0.3661
%Op. Lucrativas	50.00%	27.78%	44.44%	50.00%	72.22%	55.56%	46.15%	30.77%	46.15%	53.85%	69.23%	53.85%	
5%	Quantidade Operações	16	16	16	16	16	16	13	13	13	13	13	13
	Dias Comprado	179	179	179	179	179	179	177	177	177	177	177	177
	Lucro Médio/Op	1.13%	4.44%	5.34%	-1.13%	-4.44%	-5.34%	1.71%	5.65%	6.82%	-1.71%	-5.65%	-6.82%
	Lucro Acumulado	6.46%	48.53%	58.13%	-29.30%	-82.76%	-80.10%	12.98%	57.18%	77.59%	-30.93%	-81.91%	-77.34%
	Sharpe Ratio	0.0308	0.1708	0.2040	-0.2668	-0.2377	-0.2668	0.0717	0.2092	0.2693	-0.3501	-0.2769	-0.3501
%Op. Lucrativas	43.75%	25.00%	50.00%	56.25%	75.00%	50.00%	46.15%	30.77%	46.15%	53.85%	69.23%	53.85%	
8%	Quantidade Operações	15	15	15	15	15	15	12	12	12	12	12	12
	Dias Comprado	184	184	184	184	184	184	180	180	180	180	180	180
	Lucro Médio/Op	0.88%	4.20%	4.58%	-0.88%	-4.20%	-4.58%	1.73%	6.04%	7.17%	-1.73%	-6.04%	-7.17%
	Lucro Acumulado	1.47%	41.56%	43.63%	-25.53%	-79.89%	-72.59%	11.48%	55.10%	72.91%	-29.66%	-81.78%	-76.72%
	Sharpe Ratio	0.0084	0.1634	0.1836	-0.2513	-0.2321	-0.2513	0.0699	0.2169	0.2742	-0.3538	-0.2828	-0.3538
%Op. Lucrativas	40.00%	26.67%	46.67%	60.00%	73.33%	53.33%	41.67%	33.33%	50.00%	58.33%	66.67%	50.00%	
10%	Quantidade Operações	15	15	15	15	15	15	12	12	12	12	12	12
	Dias Comprado	185	185	185	185	185	185	182	182	182	182	182	182
	Lucro Médio/Op	1.22%	4.50%	5.26%	-1.22%	-4.50%	-5.26%	2.19%	6.45%	8.07%	-2.19%	-6.45%	-8.07%
	Lucro Acumulado	6.28%	48.27%	57.58%	-29.83%	-80.78%	-75.56%	17.17%	63.01%	91.00%	-33.95%	-82.64%	-79.35%
	Sharpe Ratio	0.0364	0.1784	0.2148	-0.2818	-0.2473	-0.2818	0.1038	0.2351	0.3134	-0.3930	-0.3013	-0.3930
%Op. Lucrativas	40.00%	33.33%	46.67%	60.00%	66.67%	53.33%	41.67%	41.67%	50.00%	58.33%	58.33%	50.00%	

Tabela 6.8: Resultado da simulação MME de 5 e 20 períodos para VALE5.

VALE5 - TWITTER(t,t-1,t-2) com MME 5 E 20													
Ob Lucro	Medições	Limiar=0						Limiar=15					
		Tendência			Contra-tendência			Tendência			Contra-tendência		
		Comprado	Vendido	C/Vendido	Comprado	Vendido	C/Vendido	Comprado	Vendido	C/Vendido	Comprado	Vendido	C/Vendido
0%	Quantidade Operações	11	11	11	11	11	11	10	10	10	10	10	10
	Dias Comprado	154	154	154	154	154	154	147	147	147	147	147	147
	Lucro Médio/Op	-1.74%	4.16%	2.31%	1.74%	-4.16%	-2.31%	-1.57%	5.01%	3.31%	1.57%	-5.01%	-3.31%
	Lucro Acumulado	-18.17%	46.68%	20.04%	19.95%	-44.69%	-31.10%	-15.10%	52.18%	29.20%	16.17%	-47.69%	-36.76%
	Sharpe Ratio	-0.7961	0.2965	0.1266	-0.2247	-0.4266	-0.2247	-0.8615	0.3463	0.1900	-0.2838	-0.4654	-0.2838
%Op. Lucrativas	27.27%	54.55%	54.55%	72.73%	45.45%	63.64%	20.00%	50.00%	40.00%	80.00%	50.00%	60.00%	
3%	Quantidade Operações	14	14	14	14	14	14	12	12	12	12	12	12
	Dias Comprado	134	134	134	134	134	134	131	131	131	131	131	131
	Lucro Médio/Op	-0.35%	4.29%	3.90%	0.35%	-4.29%	-3.90%	-0.13%	5.38%	5.25%	0.13%	-5.38%	-5.25%
	Lucro Acumulado	-5.91%	68.66%	58.70%	3.74%	-52.13%	-49.25%	-2.55%	74.68%	70.22%	0.41%	-55.13%	-54.55%
	Sharpe Ratio	-0.2527	0.3933	0.3403	-0.4796	-0.5332	-0.4796	-0.1799	0.4317	0.3958	-0.5102	-0.5566	-0.5102
%Op. Lucrativas	35.71%	71.43%	50.00%	64.29%	28.57%	50.00%	33.33%	66.67%	50.00%	66.67%	33.33%	50.00%	
5%	Quantidade Operações	13	13	13	13	13	13	12	12	12	12	12	12
	Dias Comprado	138	138	138	138	138	138	131	131	131	131	131	131
	Lucro Médio/Op	-0.66%	4.32%	3.61%	0.66%	-4.32%	-3.61%	-0.13%	5.38%	5.25%	0.13%	-5.38%	-5.25%
	Lucro Acumulado	-9.40%	62.41%	47.15%	7.63%	-50.21%	-45.15%	-2.55%	74.68%	70.22%	0.41%	-55.13%	-54.55%
	Sharpe Ratio	-0.3037	0.3592	0.2742	-0.4047	-0.4975	-0.4047	-0.1799	0.4317	0.3958	-0.5102	-0.5566	-0.5102
%Op. Lucrativas	30.77%	69.23%	46.15%	69.23%	30.77%	53.85%	33.33%	66.67%	50.00%	66.67%	33.33%	50.00%	
8%	Quantidade Operações	12	12	12	12	12	12	11	11	11	11	11	11
	Dias Comprado	147	147	147	147	147	147	139	139	139	139	139	139
	Lucro Médio/Op	-0.95%	4.44%	3.46%	0.95%	-4.44%	-3.46%	-0.38%	4.79%	5.27%	0.38%	-5.62%	-5.27%
	Lucro Acumulado	-12.02%	57.69%	38.73%	10.33%	-48.80%	-38.73%	-5.31%	69.73%	60.72%	2.77%	-53.97%	-52.66%
	Sharpe Ratio	-0.3652	0.3324	0.2274	-0.3608	-0.4695	-0.3608	-0.2395	0.4028	0.3448	-0.4617	-0.5262	-0.4617
%Op. Lucrativas	25.00%	58.33%	33.33%	75.00%	41.67%	33.33%	27.27%	40.83%	35.17%	7.24%	-40.83%	-35.17%	
10%	Quantidade Operações	12	12	12	12	12	12	11	11	11	11	11	11
	Dias Comprado	150	150	150	150	150	150	144	144	144	144	144	144
	Lucro Médio/Op	-1.42%	3.97%	2.47%	1.42%	-3.97%	-2.47%	-1.17%	4.79%	3.53%	1.17%	-4.79%	-3.53%
	Lucro Acumulado	-16.64%	49.42%	24.56%	17.17%	-45.72%	-24.56%	-12.92%	56.09%	35.93%	12.52%	-49.04%	-35.93%
	Sharpe Ratio	-0.5442	0.2892	0.1489	-0.2903	-0.4269	-0.2903	-0.4881	0.3404	0.2319	-0.3622	-0.4672	-0.3622
%Op. Lucrativas	25.00%	58.33%	33.33%	75.00%	41.67%	33.33%	18.18%	35.55%	25.85%	26.95%	-35.55%	25.85%	

Ao arquivo de expressões e palavras para limpeza de tweets utilizado nos testes acima se acrescentou mais itens como: 'corinthians'; 'deus'; 'curiosidade'; 'lento'; 'fã'; 'seguro'; 'linda'; 'carai'; 'orgulho'; 'burro'; 'burra'; 'recalque'; 'campanha da dilma'; 'Petr4 -' (este era o início de uma postagem com notícias sobre a Petrobrás) e mais 16 termos classificados como palavrões na língua portuguesa, tornando a limpeza um pouco mais severa. O arquivo de tweets para PETR4 passou de 38.070 tweets (resultado da primeira limpeza) para 27.648 com o acréscimo de expressões, e de 13.218 para 12.809 tweets da VALE5.

Como a diferença de quantidade de tweets para VALE5 é pequena da primeira limpeza para esta mais severa, serão apresentados os resultados da simulação apenas para a PETR4, cuja diferença e volume de dados é maior.

As Tabelas 6.9, 6.10 e 6.11 apresentam os resultados de simulação de compra e venda da ação PETR4 para análise da multidão, MACD com Twitter e MME 5 e 20 períodos com Twitter, respectivamente.

Ao observar e comparar os resultados obtidos para análise da multidão com limpeza e com limpeza mais severa nas Tabelas 6.3 e 6.9, é possível visualizar uma queda nos valores de lucro médio por operação e lucro acumulado obtidos da primeira para a segunda tabela. Os maiores valores obtidos para lucro médio por operação e lucro acumulado em simulações disponíveis na Tabela 6.3 foram de 4,01% e 277,86%, e na Tabela 6.9 foram de 2,16% e 110,81%, ou seja, o maior lucro caiu mais de 100% em relação ao da primeira tabela. A simulação é a mesma, a única diferença está nos dados oriundos do Twitter, que passaram por uma rotina de pré-processamento com limpeza mais severa.

Ao aplicar Twitter com MACD, o maior lucro acumulado passou de 162,31% (Tabela 6.5) para 173,04% (Tabela 6.10). No caso da simulação de MME com Twitter, o maior lucro acumulado passou de 91% (Tabela 6.7) para 172,55% (Tabela 6.11). Essa situação traz a reflexão de que um equilíbrio na limpeza dos dados das redes sociais é necessário, a remoção severa de tweets com palavrões e expressões nem sempre é a melhor opção quando se deseja obter indicadores de tendência e sentimento a partir de mensagens das redes sociais. No caso das simulações apresentadas, a diferença nos lucros acumulados em alguns dos casos foi de mais de 100% para menos, valor bastante significativo em se tratando de lucros.

Outro fato interessante a ser comentado nessa experiência é que os resultados da análise da multidão, quando aplicada uma limpeza mais severa nos dados do Twitter, foram piores do que os das análises técnica e da multidão em conjunto. O contrário aconteceu quando a limpeza aplicada foi mais branda, ou seja, os valores obtidos da análise da multidão foram superiores.

Tabela 6.9: Resultados da simulação para PETR4 por análise da multidão com limpeza severa de tweets.

PETR4 - ANÁLISE DA MULTIDÃO - TWITTER													
Ob Lucro	Medições	Limiar=0						Limiar=29					
		TENDÊNCIA			Contra-tendência			tendência			Contra-tendência		
		Comprado	Vendido	C/Vendido	Comprado	Vendido	C/Vendido	Comprado	Vendido	C/Vendido	Comprado	Vendido	C/Vendido
0%	Quantidade Operações	63	63	63	63	63	63	46	46	46	46	46	46
	Dias Comprado	206	206	206	206	206	206	180	180	180	180	180	180
	Lucro Médio/Op	-0.29%	0.35%	0.03%	0.29%	-0.35%	-0.03%	0.51%	1.64%	2.08%	-0.51%	-1.64%	-2.08%
	Lucro Acumulado	-24.42%	5.44%	-20.31%	9.92%	-33.34%	-24.90%	19.56%	66.80%	99.42%	-25.84%	-66.84%	-73.05%
	Sharpe Ratio	-0.1843	0.0129	-0.0611	-0.0752	-0.1566	-0.0752	0.0464	0.1955	0.2678	-0.4374	-0.3322	-0.4374
%Op. Lucrativas	39.68%	41.27%	44.44%	60.32%	58.73%	55.56%	41.30%	39.13%	45.65%	54.35%	60.87%	54.35%	
3%	Quantidade Operações	65	65	65	65	65	65	49	49	49	49	49	49
	Dias Comprado	198	198	198	198	198	198	173	173	173	173	173	173
	Lucro Médio/Op	-0.19%	0.42%	0.19%	0.19%	-0.42%	-0.19%	0.40%	1.44%	1.77%	-0.40%	-1.44%	-1.77%
	Lucro Acumulado	-20.37%	11.10%	-11.53%	2.60%	-36.50%	-32.83%	14.30%	59.47%	82.28%	-22.92%	-65.23%	-70.32%
	Sharpe Ratio	-0.1349	0.0352	-0.0126	-0.0929	-0.1658	-0.0929	0.0104	0.1691	0.2101	-0.3524	-0.2921	-0.3524
%Op. Lucrativas	41.54%	41.54%	49.23%	58.46%	58.46%	50.77%	42.86%	36.73%	51.02%	53.06%	63.27%	48.98%	
5%	Quantidade Operações	63	63	63	63	63	63	47	47	47	47	47	47
	Dias Comprado	198	198	198	198	198	198	173	173	173	173	173	173
	Lucro Médio/Op	-0.12%	0.51%	0.37%	0.12%	-0.51%	-0.37%	0.52%	1.61%	2.07%	-0.52%	-1.61%	-2.07%
	Lucro Acumulado	-16.22%	16.89%	-2.06%	-2.66%	-39.57%	-40.42%	20.27%	67.79%	101.79%	-26.86%	-66.91%	-73.68%
	Sharpe Ratio	-0.1164	0.0523	0.0153	-0.1326	-0.1920	-0.1326	0.0454	0.1871	0.2453	-0.3980	-0.3171	-0.3980
%Op. Lucrativas	39.68%	42.86%	49.21%	60.32%	57.14%	50.79%	40.43%	38.30%	51.06%	55.32%	61.70%	48.94%	
8%	Quantidade Operações	63	63	63	63	63	63	47	47	47	47	47	47
	Dias Comprado	200	200	200	200	200	200	168	168	168	168	168	168
	Lucro Médio/Op	-0.13%	0.50%	0.35%	0.13%	-0.50%	-0.35%	0.57%	1.66%	2.16%	-0.57%	-1.66%	-2.16%
	Lucro Acumulado	-16.86%	15.99%	-3.57%	-1.82%	-39.14%	-39.48%	22.92%	71.50%	110.81%	-28.11%	-67.64%	-74.64%
	Sharpe Ratio	-0.1203	0.0496	0.0109	-0.1272	-0.1886	-0.1272	0.0616	0.1948	0.2606	-0.4136	-0.3230	-0.4136
%Op. Lucrativas	39.68%	42.86%	47.62%	60.32%	57.14%	52.38%	42.55%	40.43%	48.94%	53.19%	59.57%	51.06%	
10%	Quantidade Operações	63	63	63	63	63	63	46	46	46	46	46	46
	Dias Comprado	201	201	201	201	201	201	177	177	177	177	177	177
	Lucro Médio/Op	-0.16%	0.47%	0.29%	0.16%	-0.47%	-0.29%	0.57%	1.68%	2.17%	-0.57%	-1.68%	-2.17%
	Lucro Acumulado	-18.34%	13.92%	-6.98%	0.16%	-38.12%	-37.22%	22.32%	70.66%	108.75%	-27.75%	-67.48%	-74.21%
	Sharpe Ratio	-0.1297	0.0434	0.0010	-0.1139	-0.1800	-0.1139	0.0649	0.2008	0.2678	-0.4240	-0.3279	-0.4240
%Op. Lucrativas	39.68%	42.86%	46.03%	60.32%	57.14%	53.97%	41.30%	39.13%	47.83%	54.35%	60.87%	52.17%	

Tabela 6.10: Resultado da simulação para PETR4 por Twitter com MACD com limpeza severa de tweets.

PETR4 - TWITTER (t,t-1,t-2) com MACD													
Ob Lucro	Medições	Limiar=0						Limiar=29					
		TENDÊNCIA			Contra-tendência			tendência			Contra-tendência		
		Comprado	Vendido	C/Vendido	Comprado	Vendido	C/Vendido	Comprado	Vendido	C/Vendido	Comprado	Vendido	C/Vendido
0%	Quantidade Operações	19	18	18	18	19	18	16	15	15	16	15	15
	Dias Comprado	249	248	248	249	248	248	237	236	236	237	236	236
	Lucro Médio/Op	-1.22%	0.12%	-3.86%	1.22%	-0.12%	3.86%	-2.87%	-1.66%	-7.50%	2.87%	1.66%	7.50%
	Lucro Acumulado	-32.67%	-6.36%	-56.23%	1.27%	-13.16%	78.48%	-47.67%	-27.22%	-72.22%	30.01%	18.63%	173.04%
	Sharpe Ratio	-0.15418	-0.06676	-0.47091	0.307489	-0.09096	0.3074887	-0.2360	-0.36904	-0.67646	0.512	0.097898	0.511591
%Op. Lucrativas	36.84%	38.89%	38.89%	63.16%	61.11%	61.11%	37.50%	33.33%	20.00%	62.50%	66.67%	80.00%	
3%	Quantidade Operações	25	24	24	25	24	24	21	20	20	21	20	20
	Dias Comprado	203	202	202	203	202	202	199	198	198	199	198	198
	Lucro Médio/Op	0.47%	1.61%	2.09%	-0.47%	-1.61%	-2.09%	-0.62%	0.37%	-0.32%	0.62%	-0.37%	0.32%
	Lucro Acumulado	-6.33%	30.28%	21.66%	-31.31%	-42.13%	-60.04%	-29.01%	-1.27%	-30.13%	-9.77%	-16.13%	-23.83%
	Sharpe Ratio	-0.0157	0.1073	0.0938	-0.2238	-0.2629	-0.2238	-0.0917	-0.0726	-0.0706	-0.0227	-0.1762	-0.0227
%Op. Lucrativas	52.00%	45.83%	45.83%	48.00%	54.17%	54.17%	47.62%	45.00%	45.00%	52.38%	55.00%	55.00%	
5%	Quantidade Operações	23	22	22	23	22	22	20	19	19	20	19	19
	Dias Comprado	230	229	229	230	229	229	219	218	218	219	218	218
	Lucro Médio/Op	-0.29%	0.80%	0.42%	0.29%	-0.80%	-0.42%	-1.26%	-0.30%	-1.70%	1.26%	0.30%	1.70%
	Lucro Acumulado	-22.59%	7.67%	-16.91%	-17.72%	-26.64%	-37.33%	-37.39%	-12.93%	-45.65%	1.64%	-3.48%	-0.59%
	Sharpe Ratio	-0.0932	0.0082	-0.0419	-0.1064	-0.1826	-0.1064	-0.1535	-0.1937	-0.1766	0.0628	-0.0797	0.0628
%Op. Lucrativas	39.13%	40.91%	45.45%	60.87%	59.09%	54.55%	40.00%	36.84%	36.84%	60.00%	63.16%	63.16%	
8%	Quantidade Operações	20	19	19	20	19	19	20	19	19	20	19	19
	Dias Comprado	219	218	218	219	218	218	219	218	218	219	218	218
	Lucro Médio/Op	-0.29%	0.80%	0.42%	0.29%	-0.80%	-0.42%	-1.26%	-0.30%	-1.70%	1.26%	0.30%	1.70%
	Lucro Acumulado	-22.59%	7.67%	-16.91%	-17.72%	-26.64%	-37.33%	-37.39%	-12.93%	-45.65%	1.64%	-3.48%	-0.59%
	Sharpe Ratio	-0.0932	0.0082	-0.0419	-0.1064	-0.1826	-0.1064	-0.1535	-0.1937	-0.1766	0.0628	-0.0797	0.0628
%Op. Lucrativas	39.13%	40.91%	45.45%	60.87%	59.09%	54.55%	40.00%	36.84%	36.84%	60.00%	63.16%	63.16%	
10%	Quantidade Operações	23	22	22	23	22	22	20	19	19	20	19	19
	Dias Comprado	233	232	232	233	232	232	221	220	220	221	220	220
	Lucro Médio/Op	-0.34%	0.77%	0.37%	0.34%	-0.77%	-0.37%	-1.33%	-0.34%	-1.77%	1.33%	0.34%	1.77%
	Lucro Acumulado	-23.55%	6.32%	-18.97%	-16.79%	-26.66%	-38.07%	-38.24%	-14.11%	-47.12%	2.92%	-3.41%	-1.53%
	Sharpe Ratio	-0.1015	0.0029	-0.0527	-0.1072	-0.1766	-0.1072	-0.1633	-0.1969	-0.1905	0.0609	-0.0728	0.0609
%Op. Lucrativas	39.13%	40.91%	40.91%	60.87%	59.09%	59.09%	40.00%	36.84%	31.58%	60.00%	63.16%	68.42%	

Tabela 6.11: Resultado da simulação para PETR4 por Twitter com MME 5 e 20 com limpeza severa de tweets.

PETR4 - WITTER(t,t-1,t-2) com MME 5 E 20													
Obj. Lucro	Medições	Limiar=0						Limiar=29					
		TENDÊNCIA			Contra-tendência			tendência			Contra-tendência		
		Comprado	Vendido	C/Vendido	Comprado	Vendido	C/Vendido	Comprado	Vendido	C/Vendido	Comprado	Vendido	C/Vendido
0%	Quantidade Operações	11	11	11	11	11	11	9	9	9	9	9	9
	Dias Comprado	178	178	178	178	178	178	167	167	167	167	167	167
	Lucro Médio/Op	-2.84%	5.31%	1.55%	2.84%	-5.31%	-1.55%	-2.21%	7.77%	4.44%	2.21%	-7.77%	-4.44%
	Lucro Acumulado	-29.44%	34.55%	-5.06%	31.72%	-79.04%	-45.94%	-20.47%	51.66%	20.62%	18.45%	-81.26%	-56.64%
	Sharpe Ratio	-0.4683	0.1790	0.0405	-0.1015	-0.2377	-0.1015	-0.3891	0.2396	0.1503	-0.2192	-0.2966	-0.2192
%Op. Lucrativas	36.36%	27.27%	36.36%	63.64%	72.73%	63.64%	33.33%	33.33%	44.44%	66.67%	66.67%	55.56%	
3%	Quantidade Operações	16	16	16	16	16	16	12	12	12	12	12	12
	Dias Comprado	167	167	167	167	167	167	164	164	164	164	164	164.00
	Lucro Médio/Op	1.36%	4.80%	5.89%	-1.36%	-4.80%	-5.89%	2.68%	7.02%	9.03%	-2.68%	-7.02%	-9.03%
	Lucro Acumulado	12.33%	56.72%	76.05%	-30.49%	-83.81%	-81.38%	25.46%	75.03%	119.59%	-37.11%	-83.76%	-81.19%
	Sharpe Ratio	0.0655	0.1978	0.2467	-0.3007	-0.2591	-0.3007	0.1499	0.2620	0.3727	-0.4418	-0.3244	-0.4418
%Op. Lucrativas	56.25%	31.25%	50.00%	43.75%	68.75%	50.00%	50.00%	33.33%	58.33%	50.00%	66.67%	41.67%	
5%	Quantidade Operações	14	14	14	14	14	14	11	11	11	11	11	11
	Dias Comprado	172	172	172	172	172	172	164	164	164	164	164	164
	Lucro Médio/Op	2.18%	6.06%	7.99%	-2.18%	-6.06%	-7.99%	3.45%	8.12%	10.92%	-3.45%	-8.12%	-10.92%
	Lucro Acumulado	21.96%	70.15%	107.51%	-36.84%	-85.00%	-84.72%	32.17%	84.40%	143.73%	-41.02%	-84.55%	-83.59%
	Sharpe Ratio	0.1132	0.2232	0.2984	-0.3528	-0.2871	-0.3528	0.1844	0.2817	0.4145	-0.4822	-0.3461	-0.4822
%Op. Lucrativas	50.00%	28.57%	57.14%	50.00%	71.43%	42.86%	45.45%	36.36%	63.64%	54.55%	63.64%	36.36%	
8%	Quantidade Operações	13	13	13	13	13	13	11	11	11	11	11	11
	Dias Comprado	177	177	177	177	177	177	164	164	164	164	164	164
	Lucro Médio/Op	1.97%	5.90%	7.31%	-1.97%	-5.90%	-7.31%	3.45%	8.12%	10.92%	-3.45%	-8.12%	-10.92%
	Lucro Acumulado	16.24%	62.17%	88.49%	-33.48%	-82.51%	-78.96%	32.17%	84.40%	143.73%	-41.02%	-84.55%	-83.59%
	Sharpe Ratio	0.0932	0.2204	0.2905	-0.3483	-0.2858	-0.3483	0.1844	0.2817	0.4145	-0.4822	-0.3461	-0.4822
%Op. Lucrativas	46.15%	30.77%	53.85%	53.85%	69.23%	46.15%	45.45%	36.36%	63.64%	54.55%	63.64%	36.36%	
10%	Quantidade Operações	13	13	13	13	13	13	11	11	11	11	11	11
	Dias Comprado	180	180	180	180	180	180	167	167	167	167	167	167
	Lucro Médio/Op	2.45%	6.32%	8.25%	-2.45%	-6.32%	-8.25%	4.01%	8.61%	12.03%	-4.01%	-8.61%	-12.03%
	Lucro Acumulado	22.92%	71.48%	110.78%	-38.03%	-83.42%	-81.63%	39.77%	95.00%	172.55%	-45.06%	-85.35%	-85.67%
	Sharpe Ratio	0.1264	0.2397	0.3310	-0.3887	-0.3059	-0.3887	0.2192	0.3031	0.4617	-0.5299	-0.3684	-0.5299
%Op. Lucrativas	46.15%	38.46%	53.85%	53.85%	61.54%	46.15%	45.45%	45.45%	63.64%	54.55%	54.55%	36.36%	

6.7 Comentários sobre transações no mercado de bolsa de valores

Como informado na introdução deste trabalho, os resultados alcançados ao final dessa pesquisa contemplam o objetivo traçado inicialmente, que era o de investigar se seria possível obter indicadores, a partir de dados de redes sociais em língua portuguesa, que pudessem auxiliar o processo de tomada de decisão de um sistema computacional. No caso deste trabalho, foi adotada a simulação de compra e venda de uma ação da cada possibilidade de negociação na bolsa de valores para validação do sistema de tomada de decisão baseado em análise da multidão.

Entretanto, para que a modalidade de compra e venda de ações baseada em análise da multidão pudesse efetivamente ser utilizada por investidores, um novo estudo sobre gastos com transações, ou seja valores pagos a corretoras de ativos e impostos, deveria ser realizado para averiguar a efetividade dos lucros apontados pelo simulador. Para operar na bolsa de valores é necessário passar por uma instituição que seja autorizada a comprar e vender ações e demais ativos na bolsa de valores. Essas instituições são habilitadas a executar suas operações pelo Banco Central e pela Comissão de Valores Mobiliários (CVM).

Existem várias corretoras no mercado de ações brasileiro, cada uma com várias opções de preços disponíveis para investidores dependendo do seu tipo de atuação no mercado. De acordo com a opção escolhida pelo investidor, um custo a mais lhe será cobrado por suas negociações. Além dos

valores com a corretora, existem também custos relacionados aos impostos cobrados no Brasil para quem opera na bolsa dependendo dos lucros obtidos.

Assim, conforme os resultados apresentados nas seções anteriores, o simulador provou que é possível obter informações de dinâmica do mercado em redes sociais e essa informação pode ser usada para compor estratégias de compra e venda de ações. Entretanto, o investidor deve avaliar, de acordo com os custos inerentes às negociações na bolsa, se estes compensam o investimento.

Como estudo futuro, dados sobre gastos em relação a corretoras mais utilizadas por investidores brasileiros podem ser levantados para estimar o lucro efetivo e real obtido com as transações realizadas a partir do simulador baseado em análise da multidão.

Tabela 6.12: Custos com corretora para negociações na bolsa de valores (valores pesquisados em outubro de 2015).

Corretora	Valor líquido	ISS(5%)	Valor bruto
Mirae (http://corretora.miraeasset.com.br)	R\$ 1,50	R\$ 0,075	R\$ 1,57
Tov(http://www.tov.com.br)	R\$ 1,99	R\$ 0,10	R\$ 2,09

Na Tabela6.12 estão disponíveis os custos cobrados por duas corretoras nacionais, cujos valores adotados estão entre os mais acessíveis, para que se possa ter uma ideia sobre os gastos efetivos no processo de adquirir ou vender ações no mercado brasileiro de bolsa de valores. Na tabela a sigla ISS se refere ao Imposto sobre serviços de qualquer natureza praticado no Brasil. É importante salientar que cada corretora tem sua forma de trabalhar oferecendo aos seus clientes pacotes de corretagem que possuem limites para compra de ações mensal. Isso significa que o investidor, adquirindo um pacote qualquer X, poderá realizar negociações até um limite de valor Y. Outras cobram a cada compra ou venda um valor Z pela transação.

Capítulo 7

Conclusão

7.1 Introdução

Nesse capítulo, serão apresentadas as conclusões finais sobre a pesquisa concluída e também considerações a respeito de possibilidades de trabalhos futuros a serem realizados na área de análise de conteúdo de redes sociais para previsão e auxílio à tomada de decisão humana.

7.2 Conclusões

Sobre os resultados obtidos das simulações demonstradas no Capítulo 6 para compra e venda de ações PETR4 da Petrobrás e VALE5 da Vale S.A., foi possível perceber o quanto a análise da multidão é promissora e desafiadora para o campo da pesquisa. O universo de indicadores a serem formados a partir de mensagens postadas nas redes sociais é imenso. Nesse trabalho, foram utilizadas apenas mensagens da rede social Twitter, entretanto, existem várias outras, inclusive específicas que podem ser exploradas por sistemas de auxílio à tomada de decisão por humanos.

Além das mensagens em redes sociais, os *links* de sites postados nas mensagens podem possuir um conteúdo que pode agregar valor ao sentimento da mensagem postada, pois muitas mensagens são compostas apenas de *links* que usuários colocam e que refletem sobre o que o usuário está pensando no momento da postagem.

Em relação aos lucros acumulados, indicadores interessantes para comentar sobre as simulações realizadas, nas Figuras 7.1 e 7.2, estão disponibilizados gráficos que permitem a comparação, quando existirem, dos cinco maiores lucros acumulados obtidos nas simulações para PETR4 e VALE5 do Capítulo 6. É visível que os maiores lucros acumulados alcançados foram com a análise da multidão. As técnicas de cruzamento de médias móveis MME e MACD obtiveram os menores valores, em algumas simulações não foi possível obter operações com os dados de entrada, por isso encontram-se sem valores nos gráficos.

Outro valor interessante para reflexão que avalia a relação entre o retorno e o risco de um investimento, tendo por base um ativo livre de risco (foi adotado o CDI) é o *sharpe ratio*. Para

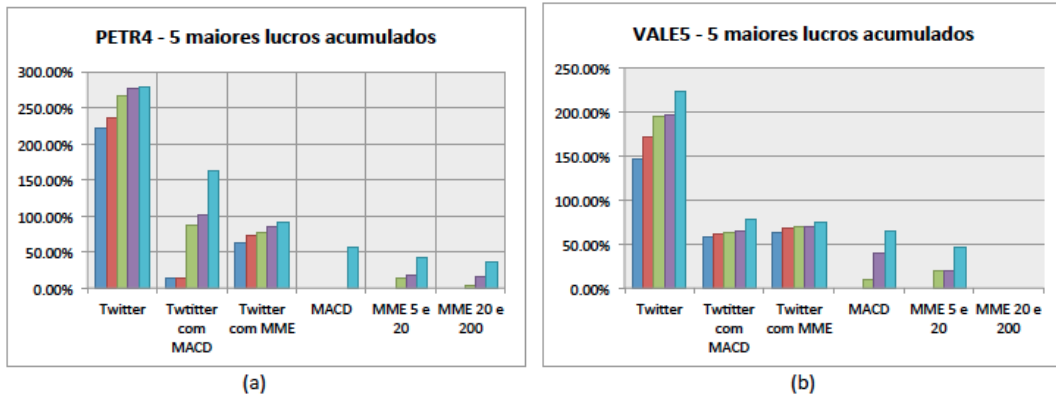


Figura 7.1: Cinco maiores lucros acumulados obtidos com as simulações realizadas, em (a) PETR4 e em (b) VALE5.

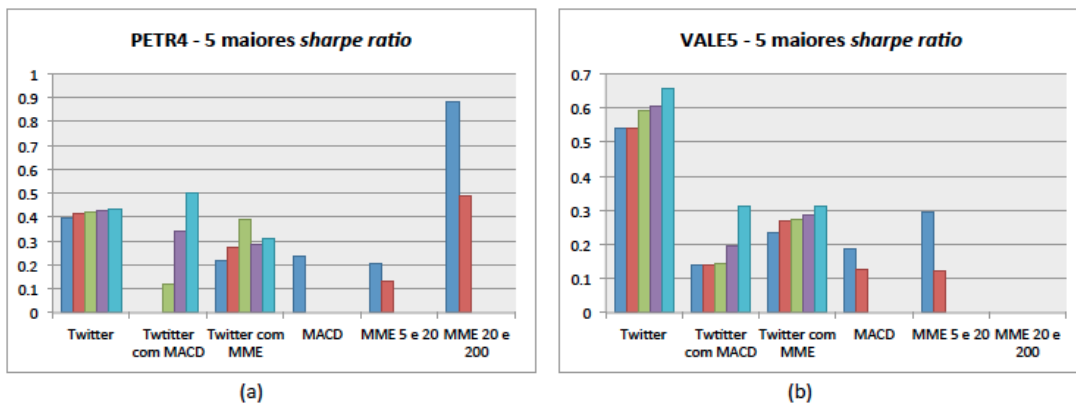


Figura 7.2: Cinco maiores valores de *sharpe ratio* alcançados com as simulações realizadas, em (a) PETR4 e em (b) VALE5.

PETR4, os maiores valores foram alcançados na simulação da análise técnica MME de 20 e 200 períodos e análise da multidão - Twitter com análise técnica MACD, Figura 7.2 (a). É importante notar que o maior valor de *sharpe ratio* alcançado de 0,88, foi realizado em uma simulação na qual apenas uma operação foi realizada, conforme a descrição da Tabela 6.1 para estratégia MME 20 e 200 períodos, na contra tendência atuando como comprado. Os maiores valores para VALE5 foram obtidos com análise da multidão, veja Figura 7.2 (b).

Os valores apresentados evidenciam o quão promissores são os indicadores obtidos a partir da rede social Twitter para avaliar compra e venda de ações no mercado brasileiro de bolsa de valores. Ambas as Figuras 7.1 e 7.2 resumem o quanto é interessante a aplicação de indicadores de redes sociais tanto isoladamente quanto quando em conjunto com outros indicadores de outras fontes. Comparando os resultados obtidos para os métodos de análise técnica com os demais que utilizam das redes sociais, o segundo conquistou melhores resultados.

É oportuno salientar que essa pesquisa não levou em consideração as flutuações dos preços das ações durante o dia, aqui foram considerados apenas o valor de preço de fechamento ajustado e todos os tweets trafegados no dia para simulação, ou seja, não foram utilizados nem preços, nem tweets de hora em hora ou de minuto em minuto como fariam os investidores normalmente.

Entretanto, mesmo em condições restritivas à análise da multidão se mostrou vantajosa em relação às demais investigadas.

Sobre a natureza dos dados obtidos das redes sociais, é importante ressaltar quão grande é o desafio de trabalhar com o pensamento humano expresso em palavras. A análise de sentimento herda todos os problemas, insolúveis até o momento, que são alvos atuais de pesquisa na área de processamento natural de linguagem tais como ambiguidade, mistura de idiomas, frases com erros ortográficos e gramaticais, sarcasmo e outros. Não é objeto de estudo de essa pesquisa produzir melhores sistemas para análise de sentimentos, no momento foi escolhido um *software* livre e de uso já comprovado em outras pesquisas e que aceitasse a língua portuguesa para treinamento. Entretanto, o problema de obtenção do sentimento do conteúdo dos tweets não foi completamente sanado. Os treinamentos realizados obtiveram em torno de 55% de acerto que é factível para as conclusões desse trabalho, porém pode ser explorado em trabalhos futuros para melhorar essa percentagem.

Outra questão é a limpeza dos dados. No decorrer da pesquisa, várias questões foram vivenciadas, uma delas foi a leitura do banco de tweets. Inicialmente, muitas mensagens (contendo *links*, palavras, repetição de mensagens) foram consideradas espúrias e no transcorrer da pesquisa, após revisões e análises revelaram-se válidas. Exemplo são os retweets, que inicialmente foram removidos da base e após reflexão foram readmitidos por reforçam uma ideia postada por outro usuário da rede social. Esses problemas foram experimentados no momento em que foram realizadas as simulações de compra e venda. Ao limpar completamente a base de tweets removendo palavras, retweets e *links*, muito do pensamento da multidão também foi removido, fazendo com que os resultados obtidos fossem menos interessantes, situação apresentada na Seção 6.6 do Capítulo 6. Dessa forma, percebeu-se uma relação muito importante da limpeza adequada dos dados e bons resultados obtidos para a análise da multidão.

Um fator também interessante em relação aos dados é o limiar adotado para restringir as operações de compra e venda. Para PETR4, nas operações nas quais foram adotadas o limiar observou-se a obtenção dos melhores valores para lucro acumulado e *sharpe ratio*. Isto significa que para efetivar uma compra ou venda é necessário ter uma quantidade mínima de comentários sobre a ação no dia em questão. Quando não há limiar, se no dia avaliado houver um tweet, esse único será o responsável pela tomada de decisão. A Figura 7.3 apresenta os melhores resultados para VALE5 e PETR4 com e sem limiar para efeitos de comparação. No entanto, os maiores lucros foram para VALE5 obtidos sem o uso de limiar. Uma questão que pode ser avaliada é a quantidade total de tweets, que para PETR4 é bem maior que a VALE5. Vários valores de limiar foram testados, todavia sem nenhum critério. Essa é uma característica que também pode ser explorada, ou seja, a obtenção de valores de limiar que proporcionem maiores lucros.

7.2.1 Comparativo simplificado de rendimentos da simulação com poupança e CDI

Uma comparação bastante simplificada pode ser realizada entre os melhores rendimentos da simulação de compra e venda de ações para PETR4 e VALE5 e a poupança e CDI, esses dois últimos

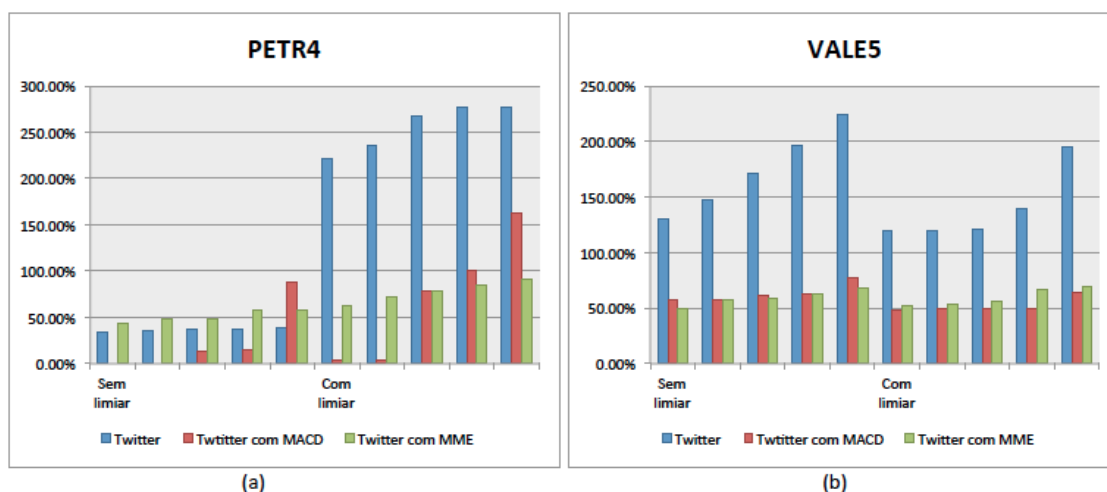


Figura 7.3: Cinco maiores lucros acumulados com e sem limiar, em (a) PETR4 e em (b) VALE5.

são considerados modalidades de investimento conservadores. Para realizar essa comparação, foi utilizada uma ferramenta de simulação disponibilizada pelo Banco Central ao cidadão brasileiro para cálculo de rendimentos de poupança, CDI e outros, a Calculadora do Cidadão¹.

No dia 13/08/2013, data inicial de coleta de dados, o preço de fechamento ajustado para uma ação PETR4 e VALE5 era de R\$16,3 e R\$36,56, respectivamente. Utilizando a Calculadora do Cidadão é possível ter uma noção de quanto seria a porcentagem de correção sobre o valor de uma ação até 04/05/2015, data final de coleta de dados da pesquisa. As Figuras 7.4 e 7.5 apresentam o cálculo feito no *site* do Banco Central para correção do valor pelo CDI e pela poupança.

Se o investidor estivesse aplicado R\$16,37 ou R\$36,56 referentes ao preço de fechamento ajustado das ações PETR4 e VALE5 no período de 13/08/2013 a 04/05/2015 obteria 19,01% de lucro acumulado em CDI (Figura 7.4) e 11,99% em poupança (7.5).

Caso o investimento fosse feito em uma ação PETR4 da bolsa de valores, e seguindo as regras de compra e venda determinadas pelo simulador apresentado no Capítulo 5, obteria 277,86% de lucro acumulado em uma análise da multidão com valor de limiar de tweets igual à 40, objetivo de lucro de 5%, atuando na tendência como COMPRADO/VENDIDO – Tabela 6.3. Sendo esse o maior lucro acumulado obtido para a PETR4 de todas as simulações realizadas.

Se optasse por uma ação VALE5, seguindo o simulador de compra e vendas (Capítulo 5), seria obtido 224,28% de lucro acumulado em uma análise da multidão, sem aplicação de limiar de tweets, objetivando lucro de 3%, atuando na tendência como COMPRADO/VENDIDO – Tabela 6.4. Sendo esse também o maior lucro acumulado obtido para a VALE5 de todas as simulações realizadas.

Dessa forma, mesmo que as aplicações sejam diferentes em termos de risco, sendo CDI e poupança conservadores e a bolsa de valores de alto risco, é possível comparar, mesmo que de maneira simplificada, o valor dos lucros obtidos com ambas opções de investimento.

¹Disponível em: <http://www.bcb.gov.br/?CALCULADORA>

BANCO CENTRAL DO BRASIL		Calculadora do cidadão	Acesso público
		08/09/2015 - 16:17	
Calculadora	Ajuda		
Início → Calculadora do cidadão → Correção de valores			[CALFW0306]
Resultado da Correção pelo CDI			
Dados básicos da correção pelo CDI			
Dados informados			
Data inicial	13/08/2013		
Data final	04/05/2015		
Valor nominal	R\$ 16,37 (REAL)		
Dados calculados			
Índice de correção no período	1,190149821048645		
Valor percentual correspondente	19,014982104864489 %		
Valor corrigido na data final	R\$ 19,48 (REAL)		
<input type="button" value="Fazer nova pesquisa"/>			

(a)

BANCO CENTRAL DO BRASIL		Calculadora do cidadão	Acesso público
		08/09/2015 - 16:16	
Calculadora	Ajuda		
Início → Calculadora do cidadão → Correção de valores			[CALFW0306]
Resultado da Correção pelo CDI			
Dados básicos da correção pelo CDI			
Dados informados			
Data inicial	13/08/2013		
Data final	04/05/2015		
Valor nominal	R\$ 36,56 (REAL)		
Dados calculados			
Índice de correção no período	1,190149821048645		
Valor percentual correspondente	19,014982104864489 %		
Valor corrigido na data final	R\$ 43,51 (REAL)		
<input type="button" value="Fazer nova pesquisa"/>			

(b)

Figura 7.4: Cálculo do rendimento do valor de uma ação no CDI durante o período simulação adotado, (a) PETR4 e (b) VALE5.

BANCO CENTRAL DO BRASIL		Calculadora do cidadão	Acesso público
		08/09/2015 - 16:13	
Calculadora	Ajuda		
Início → Calculadora do cidadão → Correção de valores			[CALFW0304]
Resultado da Correção pela Poupança			
Dados básicos da correção pela Poupança			
Dados informados			
Data inicial	13/08/2013		
Data final	04/05/2015		
Valor nominal	R\$ 16,37 (REAL)		
Regra de correção	Nova		
Dados calculados			
Índice de correção no período	1,1199177		
Valor percentual correspondente	11,9917700%		
Valor corrigido na data final	R\$ 18,33 (REAL)		
<input type="button" value="Fazer nova pesquisa"/>			

(a)

BANCO CENTRAL DO BRASIL		Calculadora do cidadão	Acesso público
		08/09/2015 - 16:10	
Calculadora do cidadão	Ajuda		
Início → Calculadora do cidadão → Correção de valores			[CALFW0304]
Resultado da Correção pela Poupança			
Dados básicos da correção pela Poupança			
Dados informados			
Data inicial	13/08/2013		
Data final	04/05/2015		
Valor nominal	R\$ 36,56 (REAL)		
Regra de correção	Nova		
Dados calculados			
Índice de correção no período	1,1199177		
Valor percentual correspondente	11,9917700%		
Valor corrigido na data final	R\$ 40,94 (REAL)		
<input type="button" value="Fazer nova pesquisa"/>			

(b)

Figura 7.5: Cálculo do rendimento do valor de uma ação na poupança durante o período simulação adotado, (a) PETR4 e (b) VALE5.

7.3 Trabalhos Futuros

Sobre os resultados obtidos para a análise estatística inicial apresentada no Capítulo 4, o modelo de regressão linear adotado, conforme comentado anteriormente, é o simples, de maneira que os resultados alcançados foram os iniciais para a pesquisa. Desejava-se, primeiramente, verificar o relacionamento estatístico entre os indicadores a serem utilizados posteriormente. Entretanto, no capítulo foram pontuadas algumas alternativas que podem ser aplicadas em relação a obter melhores resultados quanto ao relacionamento estatístico como:

- Aplicação de pesos que favoreçam um conjunto de pontos em relação a outro;
- Adoção de modelo de ajuste de uma curva ampliando a quantidade de parâmetros a serem ajustados com a finalidade de captar um intervalo de confiança maior;
- Adoção de modelos não-lineares e modelos para análise de séries temporais muito utilizados em economia como ARMA e ARIMA;
- Aplicação de indicadores de Twitter juntamente com vários indicadores de mercado para o enriquecimento de modelos de predição e obtenção de estimadores mais robustos;
- Adoção de modelos de predição baseados no burburinho e indicadores do mercado do dia $t-1$, estimar o comportamento do mercado no dia t .

Com relação aos resultados finais desta pesquisa alcançados com o simulador de compra e venda de ações, abre-se um leque de oportunidades para estudos e melhoramento de resultados. Com relação aos dados obtidos do Twitter, aqui foram apresentados estudos iniciais de análise de sentimento de tweets em língua portuguesa para o campo de mercado de bolsa de valores brasileiro. Entretanto, muito ainda pode ser desenvolvido nesse sentido, como, por exemplo:

- Verificação de limpeza de dados, qual a medida aproximada de limpeza nos dados que favoreça melhores resultados de predição. Durante os testes realizados, verificou-se, superficialmente, que quanto mais limpeza fosse aplicada aos tweets, pior era o resultado do simulador baseado apenas em análise da multidão e melhor eram os resultados obtidos quando análise técnica era empregada com análise da multidão. É interessante verificar qual seria o equilíbrio para remoção de tweets detectados como espúrios da base de dados;
- Estudos de ferramentas de análise de sentimento de documentos para a língua portuguesa. A grande maioria está disponível para a língua inglesa, entretanto, o volume de dados trafegado em língua portuguesa nas redes sociais é gigantesco e promissor não só para mercado de ações, mas também para várias outras aplicações tais como: análise de sentimento sobre aceitação de produtos e campanhas lançadas por empresas e governantes, sobre difamação de pessoas como celebridades, políticos e empresas, sobre eventos esportivos, epidemias, problemas sociais, e muitas outras possibilidades;
- Estudos sobre a veracidade da dinâmica humana, ou seja, se o conteúdo compartilhado nas redes sociais pode ser tomado como verdadeiro ou falso. No caso da pesquisa apresentada,

nas condições de dados adotados, os resultados obtidos com o simulador para análise da multidão e que seguiam a tendência do burburinho das redes, obtiveram resultados de lucros acumulados superiores as demais técnicas. Todas as operações realizadas seguindo contra a tendência para a análise da multidão obtiveram lucros negativos, como comprovados nas Figuras 6.3 e 6.4 do Capítulo 6. Isso prova que, no contexto adotado para a pesquisa, o povo fala a verdade na rede social Twitter sobre as ações PETR4 e VALE5 do mercado de ações brasileiro;

- Estudos sobre a evolução do pensamento nas redes sociais. No caso dessa pesquisa, é interessante acompanhar a evolução das postagens na rede Twitter sobre ações do mercado brasileiro de bolsa de valores, avaliando: quantidade de postagens, se o investidor está ou não interessado em continuar postando seus pensamentos a respeito do assunto nas redes, se este pensamento ao longo do tempo continua verdadeiro, se mais pessoas se interessam pelo tema, e se o pensamento do povo tende a melhorar os resultados do simulador de compra e venda ao longo de anos;
- Outra possibilidade de estudo é realizar uma fusão de indicadores do Twitter, de outras redes sociais que o brasileiro investidor tem adotado [24], e da bolsa de valores com aplicação de pesos nos mais confiáveis para obter previsões sobre o mercado do dia seguinte. No caso desse trabalho, foram realizadas simulações de compra e venda de ações baseadas apenas em indicadores para técnicas de cruzamento de médias móveis e indicadores obtidos de volume de postagens e análise de tendências e sentimento do Twitter;
- Outros indicadores podem ser adotados, como, por exemplo, um robô pode ser criado para vasculhar os *links* postados em tweets. O conteúdo apontado por esses *links* pode ser analisado por sistemas de análise de sentimentos para documentos de forma que um tweet, que aparentemente não contribuiria como indicador por conter um *link*, passaria a representar um sentimento depois que o conteúdo apontado por esse fosse analisado;
- Outra possibilidade que também pode ser investigada é a adoção de indicadores de sentimento disponibilizados por agências, como, por exemplo, Reuters² e empresas como a PsychSignal³ comentadas no Capítulo 2.
- Uma característica interessante observada nos tweets sobre as ações do mercado brasileiro é a quantidade de mensagens postadas em inglês, talvez seja interessante explorar a análise do conteúdo dessas mensagens com o intuito de verificar se elas agregariam valor aos indicadores de sentimento já obtidos com a língua portuguesa;
- Aplicação de análise de sentimento de postagens em redes sociais e modelos de previsão em outras áreas como as já mencionadas (produtos e campanhas lançadas por empresas e políticas públicas por governantes, difamação de pessoas como celebridades, políticos e empresas, eventos esportivos, epidemias, problemas sociais) e outras.

²<http://www.reuters.com>

³<https://psychsignal.com>

Uma das possibilidades de estudo na área de finanças, como apontado na Seção 6.7 do Capítulo, é a investigação sobre gastos em relação a corretoras nacionais para aquisição e venda de ações brasileiro, para que haja uma estimativa do lucro efetivo e real obtido com as transações realizadas a partir do simulador baseado em análise da multidão.

7.4 Comentários finais

Conforme todos os experimentos realizados nessa pesquisa, abordando o problema do uso de dados obtidos através de redes sociais online como fonte de informação para previsão de comportamento do mercado de ações brasileiro, os resultados apresentados se mostraram promissores para o emprego em sistemas de auxílio a tomada de decisão.

Diante de tais resultados, abre-se um amplo espaço de possibilidades. Sobre os dados utilizados nos experimentos, foram mais de oito milhões de tweets coletados da rede social Twitter e 626 dias dados históricos obtidos da bolsa de valores brasileira - Bovespa.

Inicialmente, foi realizada uma análise estatística do relacionamento entre esses dados. Posteriormente, foram selecionados os casos de teste, as ações PETR4 e VALE5 das empresas Petrobrás e Vale S.A, sobre as quais um sistema simulador efetuou operações de compra e venda de ações baseadas em tweets e preço de fechamento ajustado.

Com relação ao simulador desenvolvido para auxílio a tomada de decisão de compra e venda de ações, através da execução dos experimentos foi possível perceber o quão desafiador é trabalhar com dados obtidos de redes sociais e o quanto o caminho é promissor tanto para a pesquisa quanto para o uso fora da academia.

Há muitas possibilidades em termos de dados e sistema para serem exploradas e amadurecidas. Sobre o banco de tweets, é possível acompanhar o aumento de seu volume diariamente, consequência do crescimento da disposição do cidadão brasileiro em comentar sobre o ambiente da bolsa de valores do país em redes sociais. Isso só confirma que o estudo na área é de grande potencial no Brasil, tendo em vista que o interesse pelo mercado de ações é crescente por parte da pessoa física [24].

Há também a possibilidade de aplicação de sistemas de auxílio à tomada de decisão em diversas outras áreas conforme citado acima e em língua portuguesa, sistemas esses que poderão produzir resultados interessantes para o comércio de produtos e serviços para empresas, governo e pessoas públicas.

REFERÊNCIAS BIBLIOGRÁFICAS

- [1] RICHTER, F. *Social networking is the No. 1 online activity*. 2013. Disponível em: <<http://www.statista.com/chart/1238/digital-media-use-in-the-us/>>.
- [2] SALLOWICZ, M. *Acesso à internet no Brasil cresce, mas 53% da população ainda não usa a rede*. 2013. Disponível em: <<http://www1.folha.uol.com.br/mercado/2013/05/1279552-acesso-a-internet-no-brasil-cresce-mas-53-da-populacao-ainda-nao-usa-a-rede.shtml>>.
- [3] EBC. *Acesso a internet chega a 49 por cento da população brasileira*. 2015. Disponível em: <<http://www.ebc.com.br/tecnologia/2015/04/acesso-internet-chega-494-da-populacao-brasileira>>.
- [4] OLIVEIRA, N.; CORTEZ, P.; AREAL, N. Some experiments on modeling stock market behavior using investor sentiment analysis and posting volume from Twitter. In: *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics - WIMS '13*. New York, New York, USA: ACM Press, 2013. p. 1. ISBN 9781450318501. Disponível em: <<http://dl.acm.org/citation.cfm?doid=2479787.2479811>>.
- [5] ASUR, S.; HUBERMAN, B. Predicting the future with social media. In: *... Agent Technology (WI-IAT), 2010 IEEE ...* IEEE Computer Society, 2010. p. 492–499. Disponível em: <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5616710>.
- [6] DONDIO, P. Predicting Stock Market Using Online Communities Raw Web Traffic. In: *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*. Ieee, 2012. p. 230–237. ISBN 978-1-4673-6057-9. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6511889>>.
- [7] BOLLEN, J.; MAO, H.; ZENG, X. Twitter mood predicts the stock market. *Journal of Computational Science*, Elsevier B.V., v. 2, n. 1, p. 1–8, mar. 2011. ISSN 18777503. Disponível em: <<http://linkinghub.elsevier.com/retrieve/pii/S187775031100007X>>.
- [8] DENG, S. et al. Combining Technical Analysis with Sentiment Analysis for Stock Price Prediction. In: *2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing*. Ieee, 2011. p. 800–807. ISBN 978-1-4673-0006-3. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6118898>>.
- [9] QIAN, B.; RASHEED, K. Stock market prediction with multiple classifiers. *Applied Intelligence*, v. 26, n. 1, p. 25–33, 2007.

- [10] MAO, H.; COUNTS, S.; BOLLEN, J. *Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data*. 2011, arXiv:1112.1051 [q-fin.ST] p. Disponível em: <<http://arxiv.org/pdf/1112.1051.pdf>>.
- [11] John R. Nofsinger. Social Mood and Financial Economics. *Journal of Behavioral Finance*, v. 6, n. 3, p. 144–160, 2005.
- [12] ANTWEILER, W.; FRANK, M. Z. Is all that talk just noise - The information content of internet stock message boards. *The Journal of finance*, v. 59, n. 3, p. 1259–1294, 2004.
- [13] CAO, H. H.; COVAL, J. D.; HIRSHLEIFER, D. Sidelined investors, trade-generated news, and security returns. *Review of Financial Studies*, v. 15, n. 2, p. 615–648, 2002.
- [14] ARIAS, M.; ARRATIA, A.; XURIGUERA, R. Forecasting with twitter data. *ACM Transactions on Intelligent Systems and Technology (TIST) - Special Section on Intelligent Mobile Knowledge Discovery and Management Systems and Special Issue on Social Web Mining*, v. 5, n. 1, 2013. Disponível em: <<http://dl.acm.org/citation.cfm?id=2542190>>.
- [15] BOYD, D. M.; ELLISON, N. B. Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-mediated communication*, v. 13, n. 1, p. 210–230, 2007.
- [16] LIU, B. *Sentiment Analysis and Opinion Mining*. 1. ed. Toronto: Morgan & Claypool Publishers, 2012. ISBN 9781608458844.
- [17] Myspace. Disponível em: <<http://en.wikipedia.org/wiki/Myspace>>.
- [18] Google+. Disponível em: <<http://en.wikipedia.org/wiki/Google+>>.
- [19] HERNANDEZ, B. A. *Instagram reaches 150million monthly active users*. 2013. Disponível em: <<http://mashable.com/2013/09/08/instagram-150-million-monthly-active-users/>>.
- [20] Twitter. Disponível em: <<http://en.wikipedia.org/wiki/Twitter>>.
- [21] Facebook. Disponível em: <<http://en.wikipedia.org/wiki/Facebook>>.
- [22] MIYANO, S. *Youtube says has 1 billion monthly active users*. 2013. Disponível em: <<http://www.reuters.com/article/2013/03/21/us-youtube-users-idUSBRE92K03O20130321>>.
- [23] BENYON, D. *Interação humano-computador*. 2. ed. São Paulo: Pearson Education do Brasil, 2011. 276–284 p. ISBN 978-85-7936-109-8.
- [24] AGUILHAR, L. *Redes sociais de investimento atraem brasileiros*. 2014. Disponível em: <<http://blogs.estadao.com.br/link/redes-sociais-de-investimento-atraem-brasileiros/>>.
- [25] LERMAN, K.; GHOSH, R. Information Contagion: An Empirical Study of the Spread of News on Digg and Twitter Social Networks. *ICWSM- International AAAI Conference on Weblogs and social Media*, p. 90–97, 2010. Disponível em: <<http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewPDFInterstitial/1509/1839>>.

- [26] BENEVENUTO, F. et al. Characterizing user navigation and interactions in online social networks. *Information Sciences*, v. 195, p. 1–24, jul. 2012. ISSN 00200255. Disponível em: <<http://linkinghub.elsevier.com/retrieve/pii/S0020025511006372>>.
- [27] *Brasileiro com acesso a rede fica mais tempo na internet do que na TV*. 2014. Disponível em: <<http://www1.folha.uol.com.br/poder/2014/03/1422088-jornais-impresos-sao-a-fonte-mais-confiavel-a-populacao-diz-pesquisa.shtml>>.
- [28] ADNEWS. *Estudo analisa hábitos do internauta brasileiro*. 2013. Disponível em: <<http://exame.abril.com.br/tecnologia/noticias/estudo-analisa-habitos-do-internauta-brasileiro>>.
- [29] BBC. *IBGE Metade dos brasileiros estão conectados a internet, Norte litorânea em acesso por celular*. 2015. Disponível em: <http://www.bbc.com/portuguese/noticias/2015/04/150429_divulgacao_pnad_ibge_lgb>.
- [30] BBC. *Brasil deve fechar 2014 como o quarto país com mais acesso a internet diz consultoria*. 2014. Disponível em: <http://www.bbc.com/portuguese/noticias/2014/11/141124_brasil_internet_pai>.
- [31] FELDMAN, R. Techniques and applications for sentiment analysis. *Communications of the ACM*, v. 56, n. 4, p. 82, abr. 2013. ISSN 00010782.
- [32] GONCALVES, P.; FABRÍCIO, B.; CHA, M. *Panas-t: A Psychometric Scale for Measuring Sentiments on Twitter*. 2013. Disponível em: <<http://arxiv.org/pdf/1308.1857.pdf>>.
- [33] HU, X. et al. Unsupervised Sentiment Analysis with Emotional Signals. In: *International Conference on World Wide Web*. [S.l.: s.n.], 2013. p. 607–617. ISBN 9781450320351.
- [34] PALTOGLOU, G.; THELWALL, M. Twitter, MySpace, Digg: Unsupervised Sentiment Analysis in Social Media. *ACM Transactions on Intelligent Systems and Technology*, v. 3, n. 4, p. 1–19, set. 2012. ISSN 21576904.
- [35] DAS, A.; GAMBÄCK, B. Sentimantics: conceptual spaces for lexical sentiment polarity representation with contextuality. In: *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*. [s.n.], 2012. p. 38–46. Disponível em: <<http://dl.acm.org/citation.cfm?id=2392974>>.
- [36] PASSONNEAU, R. Sentiment Analysis of Twitter Data. n. June, p. 30–38, 2011.
- [37] HU, X. et al. Exploiting Social Relations for Sentiment Analysis in Microblogging. In: *Proceedings of the sixth ACM international conference on Web search and data mining*. [S.l.: s.n.], 2013. p. 537–546.
- [38] FIAIDHI, J. et al. Mining twitterspace for information: Classifying sentiments programmatically using Java. In: *Seventh International Conference on Digital Information Management (ICDIM 2012)*. Ieee, 2012. p. 303–308. ISBN 978-1-4673-2430-4. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6360089>>.

- [39] BATOOL, R. et al. Precise tweet classification and sentiment analysis. In: *2013 IEEE/ACIS 12th International Conference on Computer and Information Science (ICIS)*. [S.l.]: Ieee, 2013. p. 461–466. ISBN 978-1-4799-0174-6.
- [40] HASSAN, A.; ABBASI, A.; ZENG, D. Twitter Sentiment Analysis - A Bootstrap Ensemble Framework. In: *2013 International Conference on Social Computing*. [S.l.]: IEEE Computer Society, 2013. p. 357–364. ISBN 978-0-7695-5137-1.
- [41] BAHRAINIAN, S.-A.; DENGEL, A. Sentiment Analysis Using Sentiment Features. In: *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*. Ieee, 2013. p. 24–28. ISBN 978-0-7695-5145-6. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6690688>>.
- [42] NIU, Z.; YIN, Z.; KONG, X. Sentiment Classification for Microblog by Machine Learning. In: *2012 Fourth International Conference on Computational and Information Sciences*. [S.l.]: Ieee, 2012. p. 286–289. ISBN 978-1-4673-2406-9.
- [43] CAMBRIA, E. et al. Statistical Approaches to Concept-Level Sentiment Analysis. *IEEE Intelligent Systems*, v. 28, n. 3, p. 1541–1672, 2013.
- [44] PANG, B.; LEE, L. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, v. 2, p. 1–135, 2008. ISSN 1554-0669. Disponível em: <<http://www.nowpublishers.com/product.aspx?product=INR&doi=1500000011>>.
- [45] LIU, B. *Sentiment Analysis and subjectivity - Handbook of Natural Language Processing*. [S.l.: s.n.], 2010.
- [46] CAMBRIA, E. et al. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, v. 28, n. 2, p. 15–21, 2013. Disponível em: <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6468032>.
- [47] NANLI, Z. et al. Sentiment analysis: A literature review. In: *2012 International Symposium on Management of Technology (ISMOT)*. Ieee, 2012. p. 572–576. ISBN 978-1-4673-4593-4. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6679538>>.
- [48] GONCALVES, P. et al. Comparing and Combining Sentiment Analysis Methods Categories and Subject Descriptors. In: *Proceedings of the first ACM conference on Online social networks*. New York, New York, USA: [s.n.], 2013. p. 27–38. ISBN 9781450320849.
- [49] GHAG, K.; SHAH, K. Comparative analysis of the techniques for Sentiment Analysis. In: *International Conference on Advances in Technology and Engineering*. [s.n.], 2013. p. 1–7. Disponível em: <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6524752>.
- [50] NASCIMENTO, P. et al. Análise de sentimento de tweets com foco em notícias. In: *I Brazilian Workshop on Social Network Analysis and Mining*. [S.l.: s.n.], 2012.
- [51] ALEM, A. C. Análise de Sentimentos e Mineração de Links em uma Rede de Co-ocorrência de Hashtags. 2013.

- [52] Patricia L. V. Ribeiro, Li Weigang, T. L. A unified approach for domain-specific tweet sentiment analysis. In: *Proceedings of the International Conference on Information Fusion*. [S.l.: s.n.], 2015.
- [53] LIANG, P.-W.; DAI, B.-R. Opinion Mining on Social Media Data. In: *2013 IEEE 14th International Conference on Mobile Data Management*. Ieee, 2013. p. 91–96. ISBN 978-0-7695-4973-6. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6569069>>.
- [54] FIAIDHI, J. et al. Opinion mining over twitterspace: Classifying tweets programmatically using the R approach. In: *Seventh International Conference on Digital Information Management (ICDIM 2012)*. Ieee, 2012. p. 313–319. ISBN 978-1-4673-2430-4. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6360095>>.
- [55] WILLIAMS, C.; GULATI, G. What is a Social Network Worth? Facebook and Vote Share in 2008 Presidential Primaries. In: *Annual Meeting of the American Political Science Association*. Boston, MA, USA.: [s.n.], 2008. p. 1–17.
- [56] TUMASJAN, A. et al. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In: *Proceedings of the Fourth international AAAI conference on Weblogs and Social Media*. [S.l.: s.n.], 2010. p. 178–185.
- [57] CONOVER, M. D. et al. Political polarization on Twitter. In: *International AAAI Conference on Weblogs and Social Media*. [S.l.: s.n.], 2011.
- [58] GLASGOW, K.; FINK, C. From push brooms to prayer books: social media and social networks during the London riots. In: *iConference 2013*. [S.l.]: iSchools, 2013. p. 155–169.
- [59] AHMED, S.; JAIDKA, K. Protests against delhigangrape on Twitter: Analyzing Indias Arab Spring. *JeDEM Journal Of Democracy*, v. 1, n. 5, p. 28–58, 2013. Disponível em: <<http://www.jedem.org>>.
- [60] TAN, L. et al. Analyzing the impact of social media on social movements: a computational study on Twitter and the occupy wall street movement. In: *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. Niagara Falls, Canada.: [s.n.], 2013. p. 1259–1266.
- [61] SRINIVASAN, M. S.; SRINIVASA, S.; THULASIDASAN, S. Exploring Celebrity Dynamics on Twitter. In: *Proceedings of the 5 th IBM Collaborative Academia Research Exchange Workshop*. [S.l.: s.n.], 2013.
- [62] MENG, X. et al. Entity-centric topic-oriented opinion summarization in twitter. In: *Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining*. Beijing, China: [s.n.], 2012. p. 379–387.
- [63] SAEZ-TRUMPER, D. et al. Finding trendsetters in information networks. In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*. Beijing, China: ACM Press, 2012. p. 1014–1022. ISBN 9781450314626. Disponível em: <<http://dl.acm.org/citation.cfm?doid=2339530.2339691>>.

- [64] GOORHA, S.; UNGAR, L. Discovery of Significant Emerging Trends. In: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [S.l.: s.n.], 2010. p. 57–64.
- [65] LI, Y.-M.; LI, T.-Y. Deriving Marketing Intelligence over Microblogs. In: *44th International Conference on System Sciences*. Hawaii: [s.n.], 2011. p. 1–10.
- [66] KARAMIBEKER, M.; GHORBANI, A. a. Sentiment Analysis of Social Issues. In: *2012 International Conference on Social Informatics*. Ieee, 2012. p. 215–221. ISBN 978-1-4799-0234-7. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6542443>>.
- [67] ZHOU, X. et al. Sentiment analysis on tweets for social events. In: *IEEE 17th International Conference on Computer Supported Cooperative Work in Design*. [s.n.], 2013. p. 557–562. Disponível em: <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6581022>.
- [68] OELKE, D. et al. Visual opinion analysis of customer feedback data. In: *IEEE Symposium on Visual Analytics Science and Technology*. Atlantic City, New Jersey, USA: [s.n.], 2009. p. 187–194.
- [69] GOMIDE, J. et al. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. In: *Proceedings of the ACM SIGWEB Web Science Conference*. Koblenz, Germany.: [s.n.], 2011.
- [70] GHOSH, S.; SHARMA, N.; BENEVENUTO, F. Cognos: crowdsourcing search for topic experts in microblogs. In: *Proceedings of the 35th International ACM SIGIR conference on Research and development in information retrieval*. Portland, Oregon, USA.: [s.n.], 2012. p. 575–590. ISBN 9781450314725.
- [71] COSTA, H.; BENEVENUTO, F.; MERSCHMANN, L. Detecting tip spam in location-based social networks. In: *Proceeding of the 28th Annual ACM Symposium on Applied Computing*. Coimbra, Portugal.: [s.n.], 2013. p. 724–729. ISBN 9781450316569. Disponível em: <<http://dl.acm.org/citation.cfm?id=2480501>>.
- [72] GHOSH, S. et al. Understanding and combating link farming in the twitter social network. In: *Proceedings of the 21st international conference on World Wide Web - WWW '12*. Lyon, France: ACM Press, 2012. p. 61–70. ISBN 9781450312295. Disponível em: <<http://dl.acm.org/citation.cfm?doid=2187836.2187846>>.
- [73] GAO, H. et al. Detecting and characterizing social spam campaigns. In: *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. [S.l.]: ACM Press, 2010. p. 35–47.
- [74] LEE, K.; CAVERLEE, J.; WEBB, S. Uncovering social spammers: social honeypots + machine learning. In: *Proceedings of the 33th international ACM SIGIR conference on Research and development in information retrieval*. Geneva, Switzerland: ACM Press, 2010. p. 435–442.
- [75] WANG, D.; IRANI, D.; PU, C. A social-spam detection framework. In: *Proceeding of the 8th Annual Collaboration, Eletronic messaging, Anti-Abuse and Spam Conference*. Redmond, Washington, USA: ACM Press, 2011. p. 46–54.

- [76] CUNHA, E. et al. A gender based study of tagging behavior in twitter. In: *Proceedings of the 23rd ACM conference on Hypertext and social media - HT '12*. Milwaukee, WI. USA.: ACM Press, 2012. p. 323. ISBN 9781450313353. Disponível em: <<http://dl.acm.org/citation.cfm?doid=2309996.2310055>>.
- [77] WALTON, S. C.; RICE, R. E. Mediated disclosure on Twitter: The roles of gender and identity in boundary impermeability, valence, disclosure and stage. *Journal of Computers in Human Behavior*, v. 29, n. 4, p. 1465–1474, 2013.
- [78] FERREIRA, W. et al. Comer, comentar e compartilhar: Análise de uma rede de ingredientes. In: *Proceedings of the Simpósio Brasileiro de Sistemas Colaborativos*. Salvador, Brasil.: [s.n.], 2013.
- [79] BURKE, M.; MARLOW, C.; LENTO, T. Social network activity and social well-being. In: *Proceedings of the SIGCHI Conference on Human Factor in Computing Systems*. Atlanta - USA: ACM Press, 2010. p. 1909–1912.
- [80] ELLISON, N. B. et al. Cultivating Social Resources on Social Network Sites: Facebook Relationship Maintenance Behaviors and Their Role in Social Capital Processes. *Journal of Computer-Mediated Communication*, v. 19, n. 2, 2014.
- [81] GUADAGNO, R. E.; MUSCANELL, N. L.; David E. Pollio. The homeless use Facebook? Similarities of social network use between college students and homeless young adults. *Journal of Computers in Human Behavior*, v. 29, n. 1, p. 86–89, 2013.
- [82] JUNG, Y. et al. Favours from facebook friends: unpacking dimensions of social capital. In: *Proceedings of the SIGCHI Conference on Human Factor in Computing Systems*. Paris - França: ACM Press, 2013. p. 11–20.
- [83] YANG, C.-c.; BROWN, B. B. Motives for Using facebook, patterns of facebook activities, and late adolescents social adjustment to college. *Journal of Youth and Adolescence*, v. 42, n. 3, p. 403–416, 2013.
- [84] VALAFAR, M.; REJAIE, R.; WILLINGER, W. Beyond friendship graphs: a study of user interaction in Flickr. In: *Proceedings of the 2nd ACM workshop on Online social networks*. Barcelona, Espanha: ACM Press, 2009. p. 25–30.
- [85] MCGUIRE, M. *Commemoration in 140 characters: How twitter is remediating how we commemorate resonant events*. 171 p. Tese (PH.D. Thesis) — New Mexico State University, 2013.
- [86] STADDON, J.; ACQUISTI, A.; LEFEVRE, K. Self-Reported Social Network Behavior: Accuracy Predictors and Implications for the Privacy Paradox. In: *IEEE International conference on Social Computing*. Washington, DC, USA: [s.n.], 2013. p. 295–302.
- [87] SCHNEIDER, F. et al. Understanding online social network usage from a network. In: *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*. [S.l.]: ACM Press, 2009. p. 35–48.

- [88] CHEUNG, C. M. K.; CHIU, P.-Y.; LEE, M. K. O. Online Social networks: Why do students use facebook? *Journal of Computers in Human Behavior*, v. 27, n. 4, p. 1337–1343, 2011.
- [89] ZHOU, T. Understanding online community user participation: a social influence perspective. *Internet Research*, v. 21, n. 1, p. 67–81, 2011.
- [90] COMARELA, G.; CROVELLA, M.; ALMEIDA, V. Understanding Factors that Affect Response Rates in Twitter Categories and Subject Descriptors. In: *Proceedings os the 23rd ACM conference on Hypertext and social media*. [S.l.: s.n.], 2012. p. 123–132. ISBN 9781450313353.
- [91] KRISHNAMURTHY, B.; GILL, P.; ARLITT, M. A few chirps about twitter. In: *Proceedings of the first workshop on online social networks*. Seattle, WA, USA.: [s.n.], 2008. p. 19–24.
- [92] HUBERMAN, B. A.; ROMERO, D. M.; WU, F. Social networks that matter: Twiter under the microscope. *First Monday*, v. 14, n. 1, 2009. Disponível em: <<http://firstmonday.org/ojs/index.php/fm/article/view/2317/2063>>.
- [93] GUERRA, P. H. C.; JR., W. M. Combinando Interações de Endosso e Comunicação em Redes Sociais Multipolarizadas. In: *II BraSNAM - Brazilian Workshop on social network analysis and mining*. Maceió, Brasil.: [s.n.], 2013.
- [94] ALVES, B. L.; LAENDER, A. H. F.; FABRÍCIO, B. The Role of Research Leaders on the Evolution of Scientific Communities Categories and Subject Descriptors. In: *Proceedings of the 22nd international conference on world wide web companion*. Rio de Janeiro, RJ, Brazil.: ACM Press, 2103. p. 649–656. ISBN 9781450320382.
- [95] LEY, M. DBLP: some lessons learned. *Proceedings of VLDB Endowment*, v. 2, n. 2, p. 1493–1500, 2009.
- [96] BIRYUKOV, M.; DONG, C. Analysis of Computer Science Communities based on DBLP. *Research and Advanced Technology for Digital Libraries*, v. 6273, p. 228–235, 2010.
- [97] XIAO, X. et al. Scientific Communities Found Based on the Path Structure of Citation network. *Computer Science & Communications*, v. 2, n. 1, p. 16–21, 2012.
- [98] SUN, J. et al. A novel Approach for personalized article recommendation in online scientific communities. In: *46th International conference on System Sciences*. Wailea, Maui, HI: [s.n.], 2013. p. 1543–1552.
- [99] MIYATA, B. K. O.; KANO, V. Y.; DIGIAMPIETRI, L. A. Combinando mineração de textos e análise de redes sociais para a identificação das áreas de atuação de pesquisadores. In: *II BraSNAM - Brazilian Workshop on social network analysis and mining*. Maceió, Brasil.: [s.n.], 2013.
- [100] DIAS, T. M. R. et al. Modelagem e caracterização de redes científicas: um estudo sobre a plataforma Lattes. In: *II BraSNAM - Brazilian Workshop on social network analysis and mining*. Maceió, Brasil.: [s.n.], 2013.

- [101] SOARES, P. R. S.; PRUDÊNCIO, R. B. C. Predição de relacionamento baseada em eventos temporais. In: *II BraSNAM - Brazilian Workshop on social network analysis and mining*. Maceió, Brasil.: [s.n.], 2013.
- [102] CUNHA, M. d. V. et al. Rede de títulos de artigos científicos variáveis no tempo. In: *II BraSNAM - Brazilian Workshop on social network analysis and mining*. Maceió, Brasil.: [s.n.], 2013.
- [103] GILBERT, E.; KARAHALIOS, K. Widespread worry and the stock market. In: *Proceedings of the International Conference on Weblogs and Social*. [S.l.: s.n.], 2010.
- [104] OH, C.; SHENG, O. R. L. Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement. In: *Proceedings of ICIS 2011*. Shanghai, China.: [s.n.], 2011.
- [105] SPRENGER, T. O.; WELPE, I. M. *Tweets and Trades: The Information Content of Stock Microblogs*. 2010. Disponível em: <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1702854>.
- [106] RAO, T.; SRIVASTAVA, S. Modeling movements in oil, gold, forex and market indices using search volume index and Twitter sentiments. In: *Proceedings of the 5th Annual ACM Web Science Conference*. [s.n.], 2013. I, p. 336–345. ISBN 9781450318891. Disponível em: <<http://dl.acm.org/citation.cfm?id=2464521>>.
- [107] ZHANG, L. *Sentiment analysis on Twitter with stock price and significant keyword correlation*. Tese (Doutorado) — University of Texas, 2013. Disponível em: <<http://hdl.handle.net/2152/20057>>.
- [108] ALDAHAWI, H. a.; ALLEN, S. M. Twitter Mining in the Oil Business: A Sentiment Analysis Approach. In: *2013 International Conference on Cloud and Green Computing*. Ieee, 2013. p. 581–586. ISBN 978-0-7695-5114-2. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6686093>>.
- [109] KARABULUT, Y. *Can facebook predict stock market activity*. [S.l.], 2013. Disponível em: <<http://ssrn.com/abstract=1919008>>.
- [110] LOPES, T. J. P. et al. Mineração de opiniões aplicada a análise de investimentos. In: *Proceedings of the Brazilian Symposium on Multimedia and the web*. [S.l.: s.n.], 2008.
- [111] BASSAGANAS, J. *Construindo uma startup em tecnologia social - Parte I- Tuite seus sinais do metatrader 5*. 2014. Disponível em: <<https://www.mql5.com/pt/articles/925>>.
- [112] LEE, S. *Long-only Trading Strategy with NLP derives social Media Sentiment*. 2015. Disponível em: <<https://www.quantopian.com/posts/long-only-trading-strategy-with-nlp-derived-social-media-sentiment-tear-sheet-attached?c=1>>.
- [113] GINSBERG, J. et al. Detecting influenza epidemics using search engine query data. *Nature - International weekly journal of science*, v. 457, p. 1012–1014, 2009.

- [114] LAMB, A.; PAUL, M. J.; DREDZE, M. Separating Fact from fear: Tracking Flu Infections on Twitter. In: *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. [S.l.: s.n.], 2013. p. 789–795.
- [115] LAZER, D. et al. The parable of Google Flu: Traps in Big Data Analysis. *Science*, v. 343, n. 6176, p. 1203–1205, 2014.
- [116] LOHR, S. *Google Flu Trends: The Limits of Big Data*. 2014. Disponível em: <http://bits.blogs.nytimes.com/2014/03/28/google-flu-trends-the-limits-of-big-data/?_php=true&_type=blogs&_r=0>.
- [117] ARTHUR, C. *Google Flu Trends is no longer good at predicting flu, scientists find*. 2014. Disponível em: <<http://www.theguardian.com/technology/2014/mar/27/google-flu-trends-predicting-flu>>.
- [118] CULOTTA, A. Towards detecting influenza epidemics by analyzing Twitter messages. In: *Proceedings of the First Workshop on Social Media Analytics - SOMA '10*. New York, New York, USA: ACM Press, 2010. p. 115–122. ISBN 9781450302173. Disponível em: <<http://portal.acm.org/citation.cfm?doid=1964858.1964874>>.
- [119] QUINCEY, E. de; KOSTKOVA, P. Early warning and outbreak detection using social networking websites: the potencial of twitter, eletronic healthcare. In: *eHealth 2nd International Conference*. Istanbul, Tirkey.: [s.n.], 2009.
- [120] LERMAN, K.; REY, M.; HOGG, T. Using a Model of Social Dynamics to Predict Popularity of News. In: *Proceedings of the 19th international conference on World Wide Web*. Raleigh, NC, USA: [s.n.], 2010. p. 621–630. ISBN 9781605587998.
- [121] BOGDANOV, P. et al. The social media genome: modeling individual topic-specific behavior in social media. In: *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. Niagara Falls, Canada.: [s.n.], 2013. p. 236–242.
- [122] T. Nguyen, L. et al. Predicting collective sentiment dynamics from time-series social media. In: *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*. Beijing China: [s.n.], 2012.
- [123] GAYO-AVELLO, D. *A Balanced Survey on Election Prediction using Twitter Data*. 2012, arXiv:1204.6441v1 p.
- [124] MEYER, R. *A new Study Says Twitter can Predict US Elections*. 2013. Disponível em: <<http://www.theatlantic.com/technology/archive/2013/08/a-new-study-says-twitter-can-predict-us-elections/278612/>>.
- [125] Alessandro Kraemer, André Luiz Satoshi Kawamoto, M. A. G. Predicao de resultado de eleicao para reitor de universidade usando tweets como fonte de pesquisa. In: *Conferencia IADIS Ibero Americana WWW Internet*. [S.l.: s.n.], 2012.

- [126] SAKAKI, T.; OKAZAKI, M.; MATSUO, Y. Earthquake shakes Twitter users: real-time event detection by social sensors. In: *Proceedings of the 19th International Conference on World Wide Web*. Raleigh, NC, USA: [s.n.], 2010. p. 851–869.
- [127] EARLE, P. S.; BOWDEN, D. C.; GUY, M. *Twitter earthquake detection: earthquake monitoring in a social world*. 2012. Disponível em: <<http://www.annalsofgeophysics.eu/index.php/annals/article/view/5364>>.
- [128] ALBUQUERQUE, D. W. et al. Estudo do uso do Twitter como ferramenta de análise de opinião durante as eleições municipais de João Pessoa. In: *II BraSNAM - Brazilian Workshop on social network analysis and mining*. [S.l.: s.n.], 2013.
- [129] GONCALVES, P.; FABRICIO, B.; ALMEIDA, V. O que tweets contendo emoticons podem revelar sobre sentimentos coletivos? In: *II BraSNAM - Brazilian Workshop on social network analysis and mining*. Maceió, Brasil.: [s.n.], 2013.
- [130] Tiago C. de Franca, J. O. Análise de Sentimento de tweets relacionados aos protestos que ocorreram no Brasil entre junho de agosto de 2013. In: *Brazilian Workshop on Social Networking Analysis and Mining*. [S.l.: s.n.], 2014.
- [131] Leonardo Augusto Sapiras, K. B. Identificação de aspectos de candidatos eleitorais em comentários de notícias. In: *Brazilian Workshop on Social Networking Analysis and Mining*. [S.l.: s.n.], 2014.
- [132] FERREIRA, E. d. B. A. *ANÁLISE DE SENTIMENTO EM REDES SOCIAIS UTILIZANDO INFLUÊNCIA DAS PALAVRAS*. Recife, Pernambuco, Brasil., 2010.
- [133] OLIVEIRA, F. W. C. de. *Análise de sentimentos de comentários em português utilizando SENTIWORDNET*. Maringá, Brasil., 2013.
- [134] MATSUURA, S. *Brasileiros acham que o Twitter deixa a TV mais legal, diz estudo*. 2014. Disponível em: <<http://oglobo.globo.com/sociedade/tecnologia/brasileiros-acham-que-twitter-deixa-tv-mais-legal-diz-estudo-12216556>>.
- [135] TWITTER. *Developer Rules of the Road*. Disponível em: <<https://dev.twitter.com/terms/api-terms>>.
- [136] PostgreSQL Global Development Group. *PostgreSQL 9.4beta2 Documentation*. [S.l.].
- [137] *Banco do Brasil*. Disponível em: <http://pt.wikipedia.org/wiki/Banco_do_Brasil>.
- [138] BOVESPA. *Comprar e vender ações*. [S.l.], 2004. Disponível em: <<http://www.bmfbovespa.com.br/Pdf/mercqvist080604.pdf>>.
- [139] CVM - Comissão de Valores Mobiliários. *Aluguel de Ações*. Disponível em: <http://www.portaldoinvestidor.gov.br/menu/Menu_Investidor/funcionamento_mercado/aluguel_acoes.html>.

- [140] NASCIMENTO, P. et al. Análise de sentimento de tweets com foco em notícias. *lbd.dcc.ufmg.br*, 2009. Disponível em: <<http://www.lbd.dcc.ufmg.br/colecoes/brasnam/2012/007.pdf>>.
- [141] DAS, A.; BANDYOPADHYAY, S. SentiWordNet for Indian Languages. In: *Proceedings of the 8th workshop on asian Language Resources*. Beijing, China: [s.n.], 2010. p. 56–63. Disponível em: <<http://www.aclweb.org/anthology/W10-3208>>.
- [142] LINGPIPE. *Logistic Regression Tutorial*. Disponível em: <<http://alias-i.com/lingpipe/demos/tutorial/logistic-regression/read-me.html>>.
- [143] SWEENEY, D. J.; WILLIAMS, T. A.; ANDERSON, D. R. *Estatística aplicada a administração e economia*. 6. ed. São Paulo: Cengage Learning, 2013. 497–526 p.
- [144] YAU, C. *R Tutorial - An R Introduction to Statistics*. 2009. Disponível em: <<http://www.r-tutor.com/>>.
- [145] RCORETEAM. *An introduction to R*. [S.l.], 2013. Disponível em: <<http://cran.r-project.org/doc/manuals/R-intro.html>>.
- [146] FORTUNA, E. *Mercado Financeiro - Produtos e Serviços*. 17 ed.. ed. Rio de Janeiro, RJ, Brazil.: Qualitymark, 2008.
- [147] MATSURA, E. *Comprar ou Vender - Como investir na bolsa utilizando análise gráfica*. 4 edição.. ed. São Paulo: Editora Saraiva, 2006.
- [148] Portal Bussola do Investidor. *Bussola do Investidor*. Disponível em: <<http://www.bussoladoinvestidor.com.br>>.
- [149] ADVFN. Disponível em: <<http://br.advfn.com/educacional/analise-tecnica>>.
- [150] Infomoney. Disponível em: <<http://www.infomoney.com.br/educacao/guias/guias-de-analise-tecnica>>.