UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA
DA COMPUTAÇÃO

PREVISÃO DE SETORE E ÍNDICE BOVESPA POR MEIO DE NOTÍCAS ECONÔMICAS E SUAS REPERCUSSÕES EM REDES SOCIAIS

JOSÉ GILDO DE ARAÚJO JÚNIOR

CAMPINA GRANDE – PB
2016

Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Previsão de Setores e Índice Bovespa por meio de Notícias Econômicas e suas Repercussões em Redes Sociais

José Gildo de Araújo Júnior

Tese submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Doutor em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Linha de Pesquisa

Leandro Balby Marinho
(Orientador)

Campina Grande, Paraíba, Brasil

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL DA UFCG

A668p Araújo Júnior, José Gildo de.

Previsão de setores e índice Bovespa por meio de notícias econômicas e suas repercussões em redes sociais / José Gildo de Araújo Júnior – Campina Grande, 2017.

177 f. : il. color.

Tese (Doutorado em Ciência da Computação) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2016.

"Orientação: Prof. Dr. Leandro Balby Marinho".

Referências.

1. Mercado acionário. 2. Análise de dados. 3. Aprendizagem de máquina. I. Marinho, Leandro Balby. II. Título.

CDU 004.65(043)

"PREVISÃO DE SETORES E ÍNDICE BOVESPA POR MEIO DE NOTÍCIAS ECONÔMICAS E SUAS REPERCUSSÕES EM REDES SOCIAIS"

JOSÉ GILDO DE ARAÚJO JÚNIOR

TESE APROVADA EM 13/12/2016

LEANDRO BALBY MARINHO, Dr., UFCG
Orientador(a)

NAZARENO FERREIRA DE ANDRADE, Dr., UFCG
Examinador(a)

CLAUDIO ELIZIO CALAZANS CAMPELO, PhD., UFCG
Examinador(a)

ADRIANO CESAR MACHADO PEREIRA, Dr., UFMG
Examinador(a)

PAULO SALGADO GOMES DE MATTOS NETO, Dr., UFPE
Examinador(a)

CAMPINA GRANDE - PB

Resumo

Há algum tempo pesquisadores e analistas de mercado vêm apresentando indícios da previsibilidade de mercados acionários. Embora acredite-se que o mercado de ações seja imprevisível, análises de previsibilidade realizadas em bolsas da China, Turquia, Hong Kong, Itália, Teerã e EUA vêm mostrando o contrário. O fato é que a hipótese de eficiência de mercado foi planteada em 1970, e não se poderia prever as mudanças culturais e tecnológicas que impactaram o mundo, como o aumento da capacidade de processamento dos computadores, o desenvolvimento de técnicas de aprendizagem de máquina, a publicação de notícias online e a exposição em tempo real da opinião de investidores em redes sociais, por exemplo. A combinação destes elementos passaram a potencializar o lucro de investidores à medida que simplificaram o monitoramento e a gestão do risco, a compreensão do cenário econômico e até a realização de análises complexas sobre setores, índices e ações em poucos minutos.

Este trabalho se propôs a lançar luz sobre relações e impactos que as notícias econômicas publicadas em jornais brasileiros, online, mantêm com o mercado acionário nacional em dois níveis de análise: Índice Bovespa e setores. Inicialmente, foram coletadas notícias econômicas publicadas em jornais de alta circulação no Brasil entre 2000 e 2015, seus comentários e suas repercussões nas redes sociais Twitter, Facebook, LinkedIn e GooglePlus. A análise de correlação entre o índice Bovespa e a quantidade de compartilhamento de notícias em redes sociais revelam uma correlação negativa de 48%. Além disso, por meio da análise de sentimento das notícias coletadas, verificou-se que a quantidade de notícias positivas publicadas é, em média, 4.5 vezes superior ao de negativas, e que, apesar disso, as notícias negativas são mais repercutidas nas redes sociais que as positivas. Para os setores, verificou-se que o setor mais previsível apenas por meio de notícias econômicas é o setor de Petróleo, Gás e Biocombustíveis enquanto o menos previsível é o setor Bens Industriais.

Por fim, as variáveis extraídas das notícias foram utilizadas como base no desenvolvimento de modelos de predição tanto para o Índice Bovespa quanto para os setores da BM&FBOVESPA. De forma geral, os resultados encontrados superaram estatisticamente *baselines* comumente utilizados em $\sim 20\%$.

Abstract

For some time researchers and market analysts have shown evidence of predictability of stock markets. Although many investors believe that the stock market is unpredictable, predictability analysis in China, Turkey, Hong Kong, Italy, Tehran and the US stock markets has shown the opposite situation. The Efficient-Market Hypothesis (EMH) was designed in 1970 and could not anticipate the cultural and technological changes that affected the world, such as the increased processing power of computers, the development of machine learning techniques, real time publication of news and opinions of investors in social media platforms, such as twitter and facebook, for example. The combination of these elements enabled investors to perform more complex analysis of sectors, indices and stocks in almost real time, thus increasing their understanding of the stock market dynamics and improving their likelihood of success.

This study aimed to shed light on the relationships and impacts that economic news published in online Brazilian newspapers, have with the national stock market in two levels of analysis: Bovespa Index and sectors. Initially, we collected economic news published in high-circulation newspapers in Brazil between 2000 and 2015, their comments and their repercussions on social medias like Twitter, Facebook, LinkedIn and GooglePlus. The correlation analysis between the Bovespa index and the amount of news shared on social networks showed a negative correlation of 48%. Furthermore, using sentiment analysis it was found that the amount of positive news reported is in average of 4.5 times higher than the negative, and, nonetheless, the negative news are more rebound on the social media than positive news. For the sectors, it was found that the most predictable sector by economic news is the Oil, Gas and Biofuels while the less predictable is the Industrial Goods sector.

Finally, the variables drawn from the news were used as as input for the definition of prediction models for both the Bovespa Index and for the sectors of BM& FBOVESPA. In general, the results overperformed baselines such as the random classifier in $\sim 20\%$.

Agradecimentos

Ao professor Leandro Balby por sua nobreza e seu exemplo que vão muito além de sua orientação.

Aos meus amigos e companheiros de laboratório pelos inúmeros ensinamentos.

Ao CNPQ por financiar meus estudos.

À minha família por sempre compreender...

"Au milieu de l'hiver, j'ai découvert en moi un invincible été."

– Albert Camus

Sumário

Lista de Símbolos

SVM - *Máquina de Vetores de Suporte*

BM&FBOVESPA - *Bolsa de Valores, Mercadorias e Futuros Bovespa*

ON - *Ações ordinárias nominativas*

PN - *Ações preferenciais nominativas*

G1 - *Jornal relacionado a Rede Globo de Comunicações*

SFN - *Sistema Financeiro Brasileiro*

BACEN - *Banco central do Brasil*

CNSP - *onselho Nacional de Seguros Privados*

CNPC - *Conselho Nacional de Previdência Complementar*

HTF - *High-Frequency Trading*

CVM - Comissão de Valores Mobiliários

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A cada segundo, especialistas e investidores de todo mundo analisam, correlacionam e especulam exaustivamente sobre uma grande quantidade de informações na tentativa de obterem respostas ou indícios sobre a dinâmica da bolsa de valores. Algumas perguntas frequentes são: O que impactará no mercado de ações? De que maneira? Quais empresas serão afetadas? Quando? Positiva ou negativamente? Obviamente, em um mercado que movimenta bilhões de dólares diariamente, respostas para estas perguntas não são simples e a interpretação adequada de informações e eventos podem determinar o êxito ou síncope de um investimento. Neste cenário, uma das possíveis fontes de informação sobre o mercado de ações são os jornais de domínio público. Estes jornais publicam regularmente notícias que relatam sobre fatos políticos e econômicos da atualidade. É possível que as notícias sendo publicadas por estes jornais ofereçam informações relevantes de modo a impactarem a opinião de seus leitores. Como alguns desses leitores estão inseridos no mercado acionário é bastante provável que suas decisões reflitam sua interpretação das notícias e, como consequência, o mercado acabe por sentir os efeitos das notícias publicadas. Dessa forma, é possível que as informações, não apenas a análise matemática tecnicista, possam ser utilizadas como diferencial estratégico para realização de bons negócios. A Figura 1.1, por exemplo, apresenta a valorização das empresas de armamentos *Raytheon Company*[1] e *Lockheed Corporation*[2] no dia seguinte de negociações após o atentado de Paris de 13/11/2015. Isto é, após a ampla divulgação da mídia sobre o fato.

[1] http://www.raytheon.com/

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

Figura 1.1: Valorização de algumas empresas do ramo de armamentos após os ataques em Paris 13/11/2015.

Um outro exemplo do impacto de notícias sobre os mercados acionários, ocorreu logo após o resultado presidencial dos Estados Unidos quando Donald Trump venceu a disputa presidencial contrariando todas as projeções estatísticas. Logo após a divulgação da notícia por meio dos jornais online houve queda em bolsas de valores em todo o mundo. Especificamente para a BM&FBOVESPA o impacto levou a uma desvalorização de 3% das empresas.

Entender a complexidade que envolve a dinâmica do mercado acionário é crucial para a lucratividade e, em geral, cada investidor segue diferentes estratégias para alcançá-la. De forma simples, é possível dividir os investidores em dois perfis: os guiados por análises técnicas ou grafistas [Bulkowski, 2011], que operam baseados em comportamento de índices técnicos derivados de preços de ações e investidores guiados por informações extraídas de notícias e balancetes de companhias, ditos fundamentalistas. Enquanto os analistas técnicos defendem a hipótese de eficiência de mercado [Fama, 1970] na qual o preço da ação por si só já resume toda informação disponível sobre ela, trabalhos recentes rebatem essa afirmação sustentando que por meio da análise de informações contidas em notícias e nas redes sociais é possível obter diferencial estratégico potencialmente útil em gerar riqueza [Chan, 2003; Tetlock et al., 2008; Schumaker and Chen, 2010]. Afirmar categoricamente sobre qual análise é mais lucrativa (técnica vs fundamentalista) ainda será motivo de amplo debate e mesmo não sendo comprovada vantagem estratégica para a análise de notícias, é fato que inúmeros serviços jornalísticos lucram com a venda de notícias econômicas em tempo-real, e.g., Reu-

ter, Bloomberg e Folha de São Paulo.

A predição de tendências no mercado de ações baseadas em índices técnicos, tais como, MA (Medias Móveis), Linhas de Suporte e Resistência, IFR (Índices de Força Relativa), *Bollinger Bands* e Modelos Estocástico têm sido extensivamente exploradas e constituem, de fato, a base matemática sobre a qual muitos investidores de varejo apoiam suas decisões [Bulkowski, 2011]. Porém, é ingênuo equacionar o passado com variáveis que não possuem causalidade com preço. Apoiar-se apenas nessa ideia para tomar decisões no mercado financeiro apresenta-se bastante arriscado. Visto por esse ângulo, os indivíduos mais exitosos são aqueles que buscam compreender as reais causas que influenciam os preços, como por exemplo, taxas de juros, renovações de contratos, vendas de filiais, mudanças na política econômica, entre outros.

Cientes das limitações da análise gráfica, tanto a comunidade científica quanto as empresas[3] buscaram investigar o impacto que as notícias econômicas publicadas diariamente nos jornais, em conjunto com as manifestações dos leitores, causavam sobre as decisões dos investidores no mercado e, em seguida, passaram a incorporar essas informações em novas análises[4].

A Bolsa de Valores, Mercadorias e Futuros de São Paulo (BM&FBOVESPA) é uma das mais influentes bolsas de valores do mundo, movendo 1,8 trilhões de dólares em 2014 e obtendo no primeiro semestre de 2015 um lucro líquido de 392 milhões de dólares associado a um crescimento de 17% comparado a 2011[5]. Apesar de menor que outras bolsas de valores como a bolsa de New York e a Nasdaq, muitos fatores a fazem atraente, incluindo: constante tendência de crescimento do mercado, alta volatilidade de preços de ações, índices e dólar. Mesmo com essas características, esse mercado é timidamente explorado pelos brasileiros, sendo dominado pelo investimento estrangeiro[6], bancos e robôs de alta frequência [Hagströmer and Norden, 2013]. Além disso, foram encontrados poucos trabalhos na literatura revisada que exponham o impacto que a mídia nacional causa sobre seu mercado acionário [Chan and Franklin, 2011].

[3]https://www.wintoncapital.com - A Winton Capital é uma empresa britânica que desenvolve modelos computacionais hábeis a realizarem investimentos em mercados acionários.

[4]https://www.wintoncapital.com/en/research-and-insights/research-papers-briefs/research-papers-archive

[5]http://ri.bmfbovespa.com.br/ptb/2402/BVMF - 1T15 Apresentao do Resultado_15.05.2015_Final.pdf

[6]http://www.bmfbovespa.com.br/renda-variavel/BuscarParticipacaoInvestimento.aspx?Idioma=pt-br

Este trabalho se propôs a investigar de que maneira as notícias econômicas, oriundas dos jornais de domínio público de maior repercussão nacional[7], se relacionam com a BM&FBOVESPA, o mercado de ações nacional. Para isso, coletaram-se notícias econômicas disponíveis nos jornais G1, Folha de São Paulo e Estadão entre os anos de 2000 e 2015, analisaram-se os aspectos de quantidade de notícias publicadas, comentários e a repercussão dessas notícias nas redes sociais Twitter, Facebook, LinkedIn, GooglePlus e o sentimento associado a cada uma delas em relação ao cenário nacional (i.e. referem-se a algo positivo ou negativo?). Essas informações foram correlacionadas tanto com o indicador de desempenho médio das cotações dos ativos (IBOVE) quanto com variáveis associadas a cada setor da BM&FBOVESPA. Por fim, foram investigados modelos de predição para ambos os níveis de agrupamentos de ações, para o IBOVE e o para os setores, baseados puramente em informações oriundas de notícias. É importante enfatizar que ambas as análises são complementares. Enquanto a análise do índice Bovespa escrutina os dados da BM&FBOVESPA desde uma perspectiva mais ampla, a análise setorial desmembra essas informações mais amplas em setores e analisa suas relações com as notícias separadamente.

No futuro, a construção de um sistema com base nos resultados desta pesquisa permitirá que pequenos investidores do mercado, para os quais os recursos são insuficientes para investirem em informações especializadas (Reuters, Infomoney), realizem melhores negócios à medida em que terão mais condições para apoiarem suas decisões. É provável que a popularização de ferramentas como essa permitam ainda uma maior participação dos próprios brasileiros no mercado nacional que é amplamente dominado por estrangeiros e timidamente explorado por pessoas físicas.

1.1 Objetivos

A seguir apresentam-se os objetivos gerais e específicos deste trabalho.

1.1.1 Objetivo Geral

Analisar o potencial das notícias publicadas nos jornais nacionais (brasileiros) online e suas repercussões nas mídias sociais na predição de estados futuros da BM&FBOVESPA, seus

[7]http://www.anj.org.br/maiores-jornais-do-brasil/

setores e seu índice mais importante.

1.1.2 Objetivos Específicos

- Avaliar a quantidade e o conteúdo de publicações de notícias econômicas e sua correlação com o mercado acionário nacional;
- Investigar o sentimento das notícias econômicas em relação ao cenário econômico nacional (positivo ou negativo);
- Avaliar a repercussão de notícias econômicas em redes sociais, suas características e sua correlação com o mercado acionário nacional, especificamente em relação ao Índice Bovespa e aos principais setores do mercado acionário brasileiro;
- Analisar causas de potenciais *outliers*. Isto é, dias para os quais a quantidade de publicação é bastante superior a maioria e notícias para as quais há uma quantidade de repercussões, igualmente, superior a maioria;
- Extrair atributos das notícias e repercussões das mídias sociais de modo a relacioná-los com características setoriais e do índice Bovespa;
- Investigar quais algoritmos de predição são mais acurados para prever setores e o índice Bovespa por meio de notícias de jornais e repercussões das mídias sociais;
- Investigar quais setores são mais previsíveis mediante a análise de notícias;
- Investigar qual a configuração ideal dos algoritmos para alcançar os melhores valores de acurácia para a previsão dos setores e do índice Bovespa;
- Verificar a eficiência do modelo criando ao prever índices e setores da BM&BOVESPA.

1.1.3 Contribuições

Este trabalho apresenta as seguintes contribuições:

- Desenvolvimento de uma medida de sensibilidade aplicada aos setores e ao índice Bovespa em relação ao impacto causado mediante a publicação de notícias econômicas.

- Avaliação de diversos modelos de aprendizagem de máquina e a definição dos melhores modelos e suas configurações ideais para predição de tendências para cada setor da BM&FBOVESPA e para o índice Bovespa.
- Definição do algoritmo estado-da-arte para análise de sentimento das notícias de jornais nacional online.
- Análise da estabilidade das predições ao longo do tempo.
- Análise descritiva do comportamento de publicação dos jornais e da forma como estas publicações são repercutidas nas mídias sociais.
- Medição das correlações entre atributos extraídos das notícias e atributos técnicos extraídos dos setores e do índice Bovespa.
- Desenvolvimento de modelos de predição de setores e do índice Bovespa que recebem como parâmetros de entrada informações extraídas de notícias econômicas oriundas de jornais online de modo a compor ferramentas que ajudem pequenos investidores a tomarem decisões mais conscientes sobre o mercado e obterem maior retorno financeiro.

1.1.4 Considerações Finais

Este trabalho está inserido em um contexto relativo a CVM no que abrange os mercados à vista, seus setores e o índice Bovespa.

Nele, busca-se explorar, sobre vários pontos de vista, a hipótese de que o processamento de notícias publicadas em jornais de domínio público influenciam no preço das ações negociadas no mercado acionário nacional. Nessa perspectiva, foram realizadas várias análises (análise descritiva, de sentimento, de extremos e de correlação) de modo a compreender o cenário e explorar características que permitissem prever tendências gerais que pudessem ser aplicadas em vários níveis da análise da BM&FBOVESPA. Neste trabalho, essas informações foram aplicadas sobre índices amplos e em setores, porém, não estão presos a eles. No futuro, almeja-se que este trabalho auxilie pequenos investidores (*Market Markers*) de curto e médio prazo a obterem maiores lucros por meio da análise fundamentalista automática do

mercado nacional e que essas informações possam impulsionar, em certa medida, a inclusão de mais brasileiros em seu próprio mercado acionário.

Capítulo 2

Fundamentação e Formalização do Problema

Este capítulo trará informações referentes a fundamentação teórica necessária para compreensão de todo trabalho como também da formalização do problema.

2.1 Conceitualização

A conceitualização buscará esclarecer ao leitor conceitos técnicos fundamentais para compreensão deste trabalho e o posicionamento de todas as suas contribuições. Informações mais detalhadas sobre a composição do mercado e formas de acesso podem ser encontradas, ainda, no apêndice A.

2.1.1 Ações

Em um contexto econômico, ações são títulos de renda variável, emitidas por empresas constituídas na forma de sociedades anônimas e que representam a menor fração do capital da empresa emissora. Ações são convertidas em dinheiro a qualquer momento por meio da bolsa de valores.

As ações podem ser de dois tipos principais:

- **Ordinárias - (ON):** Confere participação nos resultados da empresa e direito a voto em assembleias gerais.

- **Preferenciais - (PN):** Não possuem direito à voto, entretanto, têm preferência no pagamento de dividendos.

É possível identificar o tipo da ação por meio do código dado a empresa após a abertura de capital. Os código 3, 4 referem-se a ações ordinárias e preferenciais respectivamente. Por exemplo, ações VALE3 referem-se a empresa Vale do tipo ordinária nominativa.

Proventos com Ações

É possível obter proventos mediante a posse de ações em cinco situações principais:

1. **Dividendos:** Distribuição de parte dos lucros aos acionistas. Por lei, as empresas devem dividir no mínimo 25% do lucro líquido. Atualmente não há tributação de imposto de renda sobre os dividendos.

2. **Juros sobre Capital Próprio:** Também constitui parte da distribuição dos lucros entre os acionistas, embora esse pagamento seja tratado como despesa da empresa. Dessa forma, os acionistas devem arcar com imposto de renda. Os juros sobre o capital próprio é benéfico para a empresa pois repassa aos acionistas o ônus do pagamento dos tributos.

3. **Bonificações:** As bonificações são distribuições gratuitas dadas aos acionistas que vão desde ações (proporcionais as já possuídas) a dinheiro (quando os dividendos superam os 25% previstos em lei).

4. **Aluguel:** Alugar as ações faz com que seus proprietários ganhem a taxa de locação enquanto os locatários ganhem a variação de preço do período acordado do aluguel.

5. **Subscrição:** Direito de aquisição de novos lotes de ações pelos acionistas (prioridade). Com isso, muitos acionistas podem vender o seu direito de subscrição, ou seja, sua preferência pelos novos lotes.

2.1.2 Liquidez

Em economia entende-se por liquidez a propriedade daquilo que é facilmente convertido em dinheiro. De acordo com sua liquidez, as ações do mercado de capitais são classificadas da seguinte forma:

1. **Primeira linha (*Blue Chips*):** São as empresas mais negociadas no mercado de ações e reconhecidas como de maior liquidez.

2. **Segunda linha:** São as empresas bem conceituadas junto aos investidores porém menos negociadas em comparação com as de primeira linha.

3. **Terceira linha (*Small Caps*):** São ações de pouca liquidez em geral companhias de médio e pequeno porte.

4. **Quarta linha (*Penny Stocks*):** São ações abaixo de 1 real geralmente vulgarmente reconhecidas como "apostas".

2.1.3 Índices

Índices são indicadores de desempenho de uma carteira de ações. O desempenho do índice reflete uma média geométrica, conforme fatores definidos por regulamento, do desempenho das ações que o compõem. Há diversos tipos de índices, como por exemplo:

- **Amplos:** Visam refletir de forma geral o desempenho das principais ações listadas na BM&FBOVESPA, como por exemplo: Ibovespa, IBrX 50, IBrX 100 e IBrA.

- **Setoriais:** Diferente de índices amplos, refletem o desempenho de setores como energia elétrica (IEE), imobiliário (IMOB), financeiro (IFCN) e utilidade pública (UTIL).

- **De Sustentabilidade:** Refletem o desempenho de aspectos sustentáveis como o índice de sustentabilidade empresarial (ISE) e o carbono eficiente (ICO_2).

O mercado de índices apresenta vencimento nas quartas-feiras mais próximas ao dia 15 dos meses pares.

Índice Bovespa

De forma técnica o índice Bovespa (Ibovespa, IBOVE) representa o resultado de uma carteira teórica de ativos onde todas as ações que a compõem cumprem essencialmente 3 (três) critérios: presença em no mínimo 95% dos pregões (fácil converter ações em dinheiro), participação de um volume financeiro maior ou igual a 0,1% do mercado à vista (não é pequena) e não ser classificada como "Penny Stock"(não é aposta).

Em suma, o Ibovespa indica o desempenho médio das cotações dos ativos de maior negociabilidade e representatividade do mercado de ações brasileiro.

Lucrar com Índices

Ao comprar um contrato de índice, o comprador acredita que o valor da carteira hipotética que o compõe irá subir até o vencimento. Ou seja, há uma expectativa futura[8] de sua valorização que resultará em ganho financeiro.

Como o índice é uma composição teórica de uma carteira de ações, a análise detalhada de informações e tendências de todas as ações que o compõe constitui tarefa muito trabalhosa. Via de regra, sua compra é motivada mediante informações macroeconômicas relevantes como taxa de juros, selic, desemprego, entre outros. Essas informações além de imprecisas, incorporam-se rapidamente ao preço dos ativos.

Formalmente define-se o valor diário obtido com índices da seguinte forma:

VL : valor do ajuste diário, em reais, referente à data atual.

$\triangle P$: variação da quantidade de pontos[9] na data atual.

C : valor em reais de cada ponto de índice, estabelecido pela BM&F – para o Ibovespa, por exemplo, temos R$1.00 (um real) quando o contrato é padrão e R$0,20 (vinte centavos de real) para uma fração do contrato padrão (minicontrato).

N : o número de contratos.

$$VL = \triangle P \times C \times N \tag{2.1}$$

A equação 2.1 define o valor pago diariamente aos proprietários de contratos de índices como sendo igual a variação de pontos do dia multiplicado pela quantidade de contratos e o valor correspondente ao tipo de cada contrato. Em um dia em que o índice Bovespa variou positivamente de 48.000 para 48.100 pontos, por exemplo, um investidor que possui 5 contratos padrão terá: $VL = (48.100 - 48.000) \times R\$1.00 \times 5 = R\$500$ de lucro.

[8] O mercado futuro é um acordo fixado atualmente para ser consumado em momento futuro. A bolsa apenas intermedia as transações. Neste tipo de mercado sempre que alguém ganha outro alguém perde recursos em mesma proporção.

[9] Os pontos representam uma abstração para o preço de uma carteira teórica contendo um lote de ações padrão das empresas que compõem o índice. Em geral, cada variação de ponto representa R$1.00 a mais ou a menos na valorização dessa carteira.

2.1.4 Setores

Setores são grupos de ações que desempenham atividades econômicas semelhantes entre si. A BM&FBOVESPA possui essencialmente 11 setores:

- **Construção e Transporte:** Pertencem a esse setor empresas relacionadas à construção, transporte e logística de produtos e serviços. Exemplos: Gol, Gafisa, Ecorodovias, entre outras.

- **Consumo Cíclico:** Correspondem a empresas que dependem de ciclos econômicos para obterem ganhos mais expressivos como por exemplo o setor de varejo, hotéis, tecidos e lazer em geral. Exemplos: Lojas Renner, Hering, Lojas Americanas, Arezzo, Saraiva, entre outras.

- **Consumo não Cíclico:** Pertencem a esse setor as empresas cuja necessidade de seus produtos são constantes, independente de ciclos, como por exemplo a indústria de alimentos, bebidas e saúde. Fazem parte deste setor: Ambev, BRFoods, Souza Cruz, entre outras.

- **Utilidade Pública:** Fazem parte deste setor todas as empresas relacionadas ao fornecimento de serviços de necessidade básica para a população em geral como energia elétrica, gás natural e abastecimento de água. Por exemplo: Copel, Cemig e AES Tiete.

- **Bens Industriais:** Este setor corresponde à espinha dorsal da economia. São empresas que fabricam maquinários, equipamentos, instrumentos, materiais e partes de componentes para serem utilizadas por outras indústrias ou firmas. São empresas que fazem parte desse setor: Embraer, Marcopolo e a Baumer.

- **Telecomunicações:** Este setor engloba os serviços assim definidos pela regulamentação vigente desempenhados por agentes que possuam concessão ou autorização para a prestarem serviços de telecomunicação. Telefonia fixa, móvel, comunicação multimídia, TV por assinatura, são exemplos de serviços prestados por empresas deste setor. São exemplos de empresas desse setor: Oi, Tim e Jereissati Participações.

- **Materiais Básicos:** As empresas desse setor estão relacionadas a mineração e refinamento de materiais, produção química e silvicultura. Como exemplo podemos sitar a Guerdau, Klabin e Brasken.

- **Tecnologia da Informação:** Este setor compreende as atividades de serviços que incluem empresas voltadas para o desenvolvimento e a comercialização de software, suporte técnico e manutenção de dispositivos. Algumas empresas que compreendem esse setor são: TOTVS, Positivo e a Linx.

- **Petróleo, Gás e Biocombustíveis:** Este setor compreende as empresas que atual potencialmente no segmento de Petróleo, Gás, Naval e Offshore. Seja pelo fornecimento de equipamentos, serviços de exploração, refino ou distribuição de derivados. Empresas como Petrobrás, Cosan e Ultrapar são exemplos de empresas que pertencem a esse setor.

- **Financeiro e Outros:** Fazem parte desse setor as empresas fornecedoras de produtos e serviços relativos a exploração de imóveis, holdings, intermediários financeiros, previdência e seguros, securitizadoras e outros serviços financeiros diversos. Banco do Brasil, Itaú e Caixa Seguradora são exemplos de empresas desse setor.

- **Não Classificados:** Algumas empresas inicialmente por não possuir sua classificação setorial definida podem ser classificadas como sem classificação.

Em setembro de 2016 a BM&FBOVESPA mudou a classificação de algumas empresas e a nomenclatura de alguns setores, este trabalho não contempla essas modificações.[10]

2.1.5 Relação dos Ativos com a Economia

De certa forma, é razoável crer que todos os ativos tendem a refletir mudanças econômicas, tais como, taxa selic, taxa de juros, PIB, taxa de desemprego e políticas públicas. A alteração na taxa de juros, por exemplo, impacta diretamente no poder de compra da população. Com juros baixos, mais pessoas passam a ter acesso a produtos o que aumenta a receita das empresas e a expectativa de lucro. E assim, sucessivamente.

[10]http://www.bmfbovespa.com.br/pt_br/produtos/listados-a-vista-e-derivativos/renda-variavel

O preço das ações vai mudando ao longo do tempo justamente porque novas informações e novos eventos geram nos investidores nova compreensão do futuro da empresa e essa expectativa de futuro motiva a compra ou a venda de títulos. Renovar ou não renovar o contrato? Vender ou não vender a filial? Nova política econômica? Ou manutenção da atual? Respostas à essas perguntas ou a expectativa de respostas geram o impacto e as mudanças nos preços constantemente.

2.1.6 Participantes do Mercado

Há diferentes perfis de participantes (*players*) do mercado. Entre eles, encontram-se os bancos, fundos de investimento, não financeiras, hedgers, pessoas físicas, estrangeiros, robôs de alta frequência (HTFs), entre outros. Eles distinguem-se entre si por basicamente 4 (quatro) características principais: tamanho, propósito de atuação, janela de atuação e estilo operacional.

Tamanho

Os participantes do mercado podem ser classificados de acordo com o volume de recursos que possuem:

- *Price Makers*: Participantes que possuem uma quantidade de recursos financeiros tão grande que a decisão de comprar ou vender afeta o preço das ações para baixo ou para cima.
- *Price Takers*: *Traders* de varejo. Participantes do mercado cuja decisão de comprar ou vender não impacta no preço das ações.

Propósito

Muitos participantes buscam alcançar objetivos diferentes no mercado. Enquanto muitos investidores buscam lucro por meio da valorização do preço das ações, outros participantes do mercado, como grandes empresas, buscam apenas eliminar riscos de grandes variações de preços (via mercado futuro). Por fim, temos o próprio BACEN que utiliza o mercado como alternativa de equilibrar o mercado monetário de títulos e o câmbio – oferta e demanda de moeda estrangeira em relação ao real.

Janela de Atuação

Quanto a janela de atuação os participantes do mercado podem ser divididos em:

- **Giro:** realizam várias operações diárias em diferentes mercados.
- **Day trade:** realiza ao menos uma operação de compra e venda no mesmo dia.
- **Swing trade:** retém ações compradas por alguns dias buscando realizar os melhores negócios.
- **Position:** retém as ações compradas por alguns meses ou anos.

Estilo/Conceito operacional

O mercado de capitais possui inúmeros estilos de atuação, não limitando-se apenas a análise gráfica e fundamentalista. Dentre os principais estilos de atuação no mercado destacam-se:

- **Fundamentalista:** Busca aproximar-se da compreensão do negócio da empresa, de sua contabilidade e possui um modelo para estimar como a dinâmica econômica do país irá impactar na lucratividade da empresa.
- **Abordagem Intermercados:** Busca investir mediante a interpretação de eventos econômicos que ocorram em bolsas que operaram antes da abertura da BM&FBOVESPA.
- **Reconhecimento de Padrões:** Acredita que o mercado segue padrões e que o lucro pode ser obtido ao conseguir percebê-los e/ou prevê-los. Muitas análises como as de *candlestick*, *M*, *W* [Bulkowski, 2011], são fruto da investigação desse segmento de análise do mercado. Em certa medida, o trabalho sendo proposto busca o reconhecimento de padrões por meio da análise quantitativa de notícias e suas repercussões em mídias sociais.
- **Estatístico e Indicadores Técnicos:** Analisa como e quais ativos serão afetados pela mudança dos indicadores técnicos da economia e investe baseado em interpretações dessas mudanças.

- **Econofísico:** Busca explicar o comportamento do mercado por meio de modelos matemáticos que descrevem a natureza como ondas de Elliott [Machado et al., 2015], Fibonacci, gás [de Mattos Neto et al., 2013], entre outros.

- **Algoritmos:** Visa reproduzir qualquer estratégia lucrativa para ser executada de forma automática por computadores.

- **Market Maker:** Compra para vender em seguida, oportunista de curto prazo, ganha na diferença do negócio.

- **Análise de Fluxo de Ordens:** Acompanha o mercado e tenta imitar as operações realizadas por grandes *players* analisando o volume de compra e venda fora do convencional (agressão) ao fluxo de ordens.

2.1.7 Mídia Brasileira e Informações Diferenciadas

Grande parte das teorias econômicas consideram que os agentes atuantes nos mercados possuem acesso as mesmas informações. No entanto, é bastante provável que tal afirmação seja uma simplificação bastante ingênua da realidade. Por exemplo, quando o Grupo Oi de telefonia decide injetar R$ 10 bilhões no Brasil, necessariamente deve fazê-lo por meio da contratação de um banco. O banco que o Grupo Oi contratou para realizar o serviço, por sua vez, é participante do mercado e sabe previamente que serão injetados R$ 10 bilhões no Brasil e a data na qual esses valores entrarão no Brasil. Apesar de não saber quem planeja retirar dinheiro do Brasil, nessa situação o banco possui muito mais informações que muitos *players* e pode alavancar grandes lucros utilizando essa informação na hora de operar o mercado[11].

Tomando por base situações reais semelhantes ao exemplo que foi mencionado, é fato que existe uma assimetria do nível de informação entre todos os participantes no mercado. Apoiada sobre esta hipótese surgiu recentemente uma nova linha de estudos econômicos denominada de economia de microestruturas. Nessa nova linha, são sugeridos modelos econômicos que passam a considerar *players* que possuem informações privilegiadas sobre o mercado.

[11] Parte dessa análise foi estudada no curso introdutório ao trader ministrado gratuitamente pela empresa Scalper Trader: http://scalpertrader.com.br/

A Figura 2.1 apresenta uma divisão didática, porém informal, sobre os níveis de informação de participantes do mercado acionário.

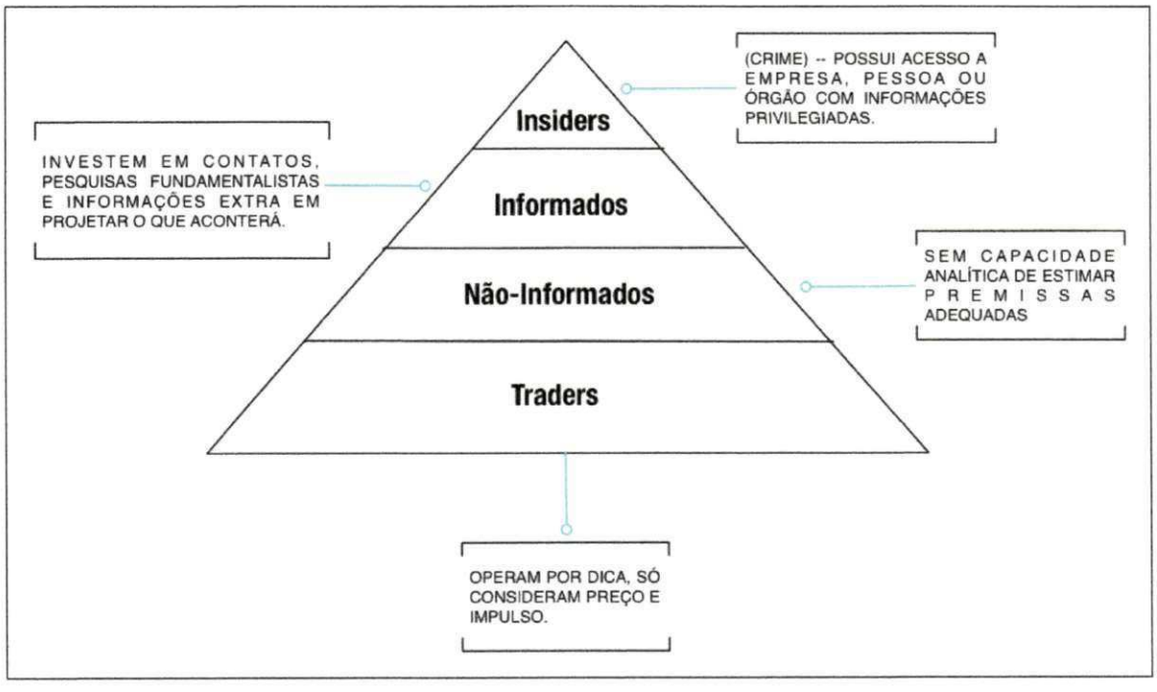

Figura 2.1: Hierarquia de informações no Mercado Acionário.

Apesar da hipótese de eficiência do mercado afirmar que o preço de um ativo já resume por si só toda a informação disponível sobre ele, muitos trabalhos atuais apresentam evidências contrárias à essa hipótese.

Mídia Brasileira

Apesar de o Brasil ser um país de dimensões continentais, com uma população aproximada de 200 milhões de cidadãos, surpreende que um pequeno grupo de famílias – Marinho (Globo), Frias (Folha), Mesquita (Estadão), Civita (Abril), Abravanel (SBT), Levy (Gazeta), Nascimento Brito (Jornal do Brasil), Saad (Band) e Bloch (Antiga Manchete), Macedo (Record) – detenham o monopólio de toda imprensa nacional [12]. É de se supor que as políticas de trabalho em cada um desses veículos de comunicação condicionem tanto a produção quanto a publicação de notícias de modo a não ferir os interesses particulares de seus proprietários e, dessa forma, apresentem visões tendenciosas sobre determinado fato. Neste caso,

[12] http://www.fndc.org.br/noticias/midia-brasileira-e-controlada-por-apenas-11-familias-924625/

político-econômico.

É possível que notícias tendenciosas sejam capazes de gerar "marés"psicológicas sociais, influenciando inúmeras pessoas em várias esferas de sua vida (Em quem votar? O que comprar para o natal? Em quem acreditar nas disputas políticas?). A Figura 2.2 apresenta um exemplo de notícia tendenciosa publicada no jornal Folha de São Paulo, onde a imagem em destaque induz o leitor a entender que o prefeito de que trata a matéria principal é o da foto, o que não acontece.

Mesmo sem comprovação científica do impacto causado por notícias ao mercado acionário nacional, o empresário Eike Baptista foi processado por divulgar em maio de 2010 notícias superestimadas sobre a empresa OGX em um programa de televisão, o que teria influenciado inúmeros investidores a comprarem ações da companhia, inflando os preços das ações e lucrando em seguida[13].


C6 ribeirão

MÁFIA DO ISS

Auditor preso diz que 'prefeito' tinha ciência de tudo

Subsecretário de Kassab, Ronilson Rodrigues fez afirmação em conversa gravada em setembro com autorização judicial

O ex-prefeito diz que o fiscal deu informação falsa e que repudia tentativa de envolvê-lo em irregularidades

Afirmações de auditor fiscal são falsas, diz Kassab

O prefeito de São Paulo, Fernando Haddad (PT), em coletiva no auditório da prefeitura

Haddad afirma que demitirá 'vários' servidores municipais

Kassab deixa evento antes de petista chegar

FOLHA DE S.PAULO

CINEMA SINOPSES

CINEMA ENDE

Figura 2.2: A forma como a imagem foi disposta ao lado da notícia gera uma interpretação tendenciosa. A imagem ao lado da manchete principal não corresponde ao sujeito da notícia.

Ao longo deste trabalho são apresentadas análises sobre as publicações econômicas ao longo dos anos e sua correlação com a BM&FBOVESPA. Os resultados obtidos mostram

[13]http://epoca.globo.com/tempo/noticia/2014/02/como-beike-batistab-turbinou-acoes-das-proprias-empresas.html

que, em certa medida, a previsibilidade da BM&FBOVESPA contraria hipótese de eficiência de mercado.

2.2 Computação Inteligente

Este trabalho utiliza inúmeras técnicas da inteligência artificial de modo a aprender padrões por meio de exemplos seja para descrever comportamentos ou prevê-los. A Inteligência computacional consiste de um conjunto de conceitos, paradigmas e algoritmos que sustentam em teoria a existência de comportamentos inteligentes em ambientes computacionais. Grande parte dos artefatos produzido são influenciados por analogias ao mundo natural ou ao menos inspirados pela natureza. Estes métodos possuem a vantagem de serem robustos diante do impreciso e incerto, facilitando, dessa forma, o encontro de soluções que são aproximações, viáveis e robustas ao mesmo tempo [Kruse et al., 2013].

2.3 Formalização

Nesta seção será apresentada uma formalização geral que será instanciada tanto para a análise setorial quanto para a análise da BM&FBOVESPA.

Todas as variáveis analisadas são numéricas e, em essência, têm-se um problema de regressão. Entretanto, na prática, grande parte dos investidores está apenas interessada em aspectos relacionados às tendências dos setores e não seus valores exatos de crescimento ou queda. Neste ponto, é possível simplificar o problema de regressão que necessariamente tentaria precisar o valor médio exato no futuro de uma determinada análise para a classificação de sua tendência futura. Dito isto, cada uma das variáveis alvo foi classificada em três níveis de tendências futura, são elas: *crescimento*, *consolidação*, e *queda*. Dado um *timestamp* t, para cada variável alvo, deseja-se predizer se a dita variável irá crescer, manter-se estável (consolidação) ou decrescer no *timestamp* $t + 1$.

De maneira formal, têm-se para cada variável um problema de classificação multi-classe. Seja:

$X \in \mathbb{R}^m$: um conjunto de vetores de características m-dimensionais.

Y : o conjunto de classes consideradas no problema.

$D^{train} = \{(\vec{x}_1,y_1),\ldots,(\vec{x}_n,y_n)\}$: o conjunto de treino onde $\vec{x} \in X$ é um vetor de atributos e $y_i \in Y$ representa a classe para a qual $\vec{x}_i$ está associado.

A ideia é encontrar a função de classificação $\hat{y} : X \rightarrow Y$ que minimiza o erro no conjunto de teste $D^{test} = \{(\vec{x}_1,y_1),\ldots,(\vec{x}_p,y_p)\}$, que não é conhecido durante a etapa de treinamento. Isto é, $D^{test} \cap D^{train} = \emptyset$.

Dessa forma tem-se:

$$err(\hat{y}; D^{test}) = \frac{1}{|D^{test}|} \sum_{(\vec{x},y)\in D^{test}} l(y, \hat{y}(\vec{x})) \tag{2.2}$$

onde $l : Y \times Y \rightarrow \mathbb{R}$ é função que mede, para alguma instância de teste $(\vec{x}, y) \in D^{test}$, a diferença entre o valor ideal y e o valor previsto $\hat{y}(\vec{x})$.

Pode-se instanciar esta definição geral do problema presente na seção para o problema de previsão de setores da BM&FBOVESPA da seguinte forma: Para cada setor da BM&FBOVESPA tem-se quatro alvos, isto é, $Y_1, Y_2, \ldots, Y_4$, cada um deles relacionado com uma das variáveis resposta *média de preços*, *quantidade de negociações*, *quantidade de contratos negociados* e *volume financeiro* respectivamente. Sendo assim, cada $Y_i = \{0, 1, 2\}$ representa o i-ésimo alvo segundo os níveis de classificação (i.e., queda (0), consolidação (1) e crescimento (2)). É importante perceber que agora faz-se necessário a construção de classificadores multi-classe separados para cada uma das variáveis alvos em cada setor da BM&BOVESPA. Ou seja, dado que existem 9 setores ativos na BM&BOVESPA e 4 alvos de interesse que podem ser analisados em cada um deles, são necessários 36 classificadores multi-classe.

De forma semelhante, pode-se instanciar a definição geral para o problema de previsão de tendências do IBOVE. Cada característica referente ao IBOVE corresponde a um dos quatro alvos desejados, isto é, $Y_1, Y_2, \ldots, Y_4$, cada um deles relacionado com uma das variáveis resposta *média de preços*, *quantidade de negociações*, *quantidade de contratos negociados* e *volume financeiro*. Cada $Y_i = \{0, 1, 2\}$ representa o i-ésimo alvo classificando segundo os níveis de classificação (i.e., queda (-), consolidação (1) e crescimento(2)). Considera-se queda quando o valor em $(t + 1)$ é inferior ao valor de t em mais 0.5%. Considera-se crescimento, a situação oposta, ou seja, quando o valor em $(t + 1)$ é superior ao valor de t em mais de 0.5%. Por fim, considera-se consolidação o estado onde a diferença de valores entre $t + 1$ e t é inferior a 0.5% seja positivamente ou negativamente. Do ponto de

vista prático, pode-se afirmar que devido as taxas envolvidas com as negociações em bolsa, nem toda variação positiva ou negativa resulta em bons negócios. Faz-se necessário que a variação de valores seja suficiente para superar essas taxas e dessa forma, após experimentos preliminares o valor de 0.5% foi escolhido como limiar para a classificação. É importante perceber que agora faz-se necessário, também, a construção de classificadores multi-classe separados para cada uma das variáveis alvos relacionadas ao IBOVE.

É importante ressaltar que ambas as análises são distintas. A análise do índice Bovespa investiga os dados da BM&FBOVESPA desde uma perspectiva mais ampla, deste ponto de vista, cada registro contém a sumarização das empresas mais negociadas na BM&FBOVESPA. Por sua vez, a análise setorial lança luz sobre outra perspectiva, explorando as mesmas informações, porém, em um nível de granularidade mais específico. Na análise setorial as empresas passam a ser agrupadas em setores e cada setor é analisado individualmente conforme especificado na formalização.

Capítulo 3

Trabalhos Relacionados

Este capítulo apresenta uma análise atualizada do estado da arte no tocante aos principais temas de contribuição desse trabalho. Por óbvio, há uma vasta literatura a ser explorada que se estende desde a matemática aplicada, econometria, até a análise de sentimentos e aprendizagem de máquina. Aqui serão relacionados apenas os aspectos que têm relação direta com as contribuições propostas por este trabalho como a importância da previsão de índices nos mercados acionários, a utilização da aprendizagem de máquina como forma de construir modelos de previsão de preço e magnitude de riscos e a utilização de notícias de jornais e redes sociais como fontes de sinais informativos sobre a tendência dos mercados.

Por fim, são apresentadas e discutidas algumas ferramentas de previsão do mercado norte americano baseadas em notícias e redes sociais que vem sendo utilizadas por acionistas como forma de obter vantagem estratégica.

3.1 Previsão de Índices

A previsão de índices do mercado acionário vem sendo um problema desafiador não só para os analistas de mercado como também para matemáticos, economistas e cientistas da computação. O índice, de maneira geral, representa o desempenho médio das principais empresas que compõem o mercado acionário, não apenas o reflexo de uma ação específica. Prevê-lo dá ao investidor a vantagem de saber como o mercado irá reagir, e assim, gerir seus recursos de forma a sempre potencializar seus ganhos.

Trabalhos como os de [Shi et al., 2016] [Talarposhti et al., 2016] e [Awasthi and Mala-

feyev, 2015] analisaram índices por meio de suas séries temporais e constataram evidências empíricas quanto a previsibilidade do mercado acionário.

Enquanto alguns trabalhos de previsão de índices buscam prevê-lo utilizando a implementação de regras baseadas apenas em preços ou em medidas derivadas dele [Slivka and Biryol, 2015] um novo segmento de trabalhos busca prevê-lo por meio da combinação entre análise de dados e algoritmos de aprendizagem de máquina como [Yin et al., 2015] e o trabalho aqui sendo proposto.

Alguns trabalhos buscam prever índices de um determinado mercado por meio de sua correlação com índices de outros mercados. [González, 2016], por exemplo, ao investigar essas correlações encontrou evidências de causalidade entre variações de pares de índices. Ou seja, variação de um determinado índice implica necessariamente na variação de outro com o qual está relacionado. Neste estudo foi encontrada relação entre o índice indiano *SENSEX*[14] e o chinês *Shanghai Stock Exchange*[15]. Também foi verificado que o índice italiano *MIB 30* é o principal causador de variação para as bolsas Européias assim como o índice americano *DOW JONES 100*[16] para as bolsas da América. Já o trabalho de [Chmielewski et al., 2015] que utilizou o processo de clusterização aplicada aos índices concluiu que há grupos de índices que influenciam outros e que poderiam ser utilizados em decisões estratégicas. Por fim, uma pequena parcela de trabalhos como o de [Da Fonseca and Wang, 2015] buscam criar modelos que expliquem o comportamento de mercados. Neste caso, concluiu-se que por meio do modelo Markoviano é possível correlacionar índices de diferentes mercados baseado no processo de transmissão de informação entre eles.

3.2 Aprendizagem de Máquina e o Mercado Financeiro

A utilização da aprendizagem de máquina como forma de obter vantagem estratégica no mercado financeiro por meio da previsão de preços ainda continua sendo amplamente explorada por investidores e cientistas. Recentemente trabalhos como os de [Zhao and Wang, 2015] e [Li, 2015] contrariam a hipótese de eficiência de mercado afirmando que alguns investidores de posse de robôs construídos com base em técnicas de regressão linear e árvores de deci-

[14]http://www.bseindia.com/sensexview

[15]http://english.sse.com.cn/

[16]http://www.djindexes.com/dividend/

são implementadas sob estratégias oriundas da análise técnica vêm superando diariamente estratégias *buy-and-hold* nas bolsas de Hong Kong e China obtendo lucros consistentes.

Da mesma forma que os anteriores, mas sem a implementação em robôs [Li et al., 2014] e [Rao et al., 2015] propuseram modelos capazes de prever os movimentos de preços de ações baseados na análise quantitativa de dados do mercado. Essas informações, obtidas apenas em períodos *intra-day*[17] e *tick-by-tick*[18], foram submetidas a técnicas de aprendizagem de máquina como máquina de vetor de suporte (SVM) e redes neurais com *back-propagation* (BP-NN). Os resultados mostraram que a estratégia foi suficiente para gerar lucros e minimizar riscos. Já [Asad, 2015] foi mais ousado e buscou desenvolver algoritmos baseados em aprendizagem de máquina capazes de avaliar cenários de negociação de um portfólio. O modelo construído apresentava uma combinação dos métodos SVM, floresta aleatória e redes neurais multi-camadas com decisão por voto majoritário. Segundo o autor, o método também obteve lucro quanto testado na bolsa da Turquia.

Fugindo da análise técnica, [Minev et al., 2012] mediu o impacto das notícias sobre os retornos anormais de investimentos em ações, tanto do ponto de vista das ciências econômicas quando da ciência da computação. As funções estatísticas elaboradas mediram as volatilidades incomuns de preços e mostraram suas relações com as notícias. Por sua vez, [Gu et al., 2015] e [Wu et al., 2014] propuseram a construção de modelos hábeis em identificar construções e conceitos úteis para serem combinados a análise de dados de modo a potencializar seus resultados e obter lucro. Com ênfase na análise de sentimento aplicada a opiniões de outros investidores, o trabalho de [Wu et al., 2014] integra à análise de sentimento aprendizagem de máquina com base na máquina de vetores de suporte e auto-regressão. Informações foram recolhidas do site *Sina Finance*[19] que contém opiniões de muitos investidores. Os resultados empíricos sugerem a existência de correlações entre as tendências de preços de ações e a análise de sentimento realizada nas mensagens publicadas nos fóruns de discussão.

Por fim, o trabalho de [Vega, 2010] confirmou hipóteses de especialistas de Wall Street que afirmam haver benefícios ao investir no mercado de ações em torno dos dias santos judaicos. Isto é, comprando no Yom Kippur (Dia do Perdão) e vendendo em Rosh Hashaná (o ano novo judaico) era possível obter lucro. De acordo com a CBS News, dados que

[17] Enquanto o mercado está aberto a negociações.

[18] A cada variação de preço, momento a momento.

[19] http://finance.sina.com.cn/

remontam a 1950 mostram que Setembro é geralmente um mês ruim para o índice S&P 500, enquanto Outubro tende a ser historicamente rentável. A seção 5.1.3 apresenta uma análise semelhante realizada ao longo dos meses para a bolsa de valores brasileira.

A Tabela 3.1 compara os trabalhos comentados nessa seção com o nosso trabalho. Aqui, todas as abordagens de previsão levaram em consideração, em grande medida, técnicas de aprendizagem de máquina baseadas em árvores de decisão treinadas com informações quantitativas extraídas de notícias econômicas que foram publicadas em jornais de alta circulação no Brasil. Além de buscar compreender as correlações entre as notícias publicadas pelos jornais e o mercado de ações brasileiro, este trabalho investigou modelos de previsão que apresentaram bons resultados quando aplicados tanto para previsão do índice Bovespa quando para a previsão de setores da BM&FBOVESPA, apresentado-se assim potencialmente úteis para auxiliar investidores a se posicionarem de forma estratégica diante do mercado.

		Mercado			**Técnicas de Aprenziagem Encontradas**				
Trabalhos	**Análise Técnica**	*Ásia*	*América*	**Lucro**	*SVM*	*Regressão*	*Floresta Aleatória*	*Árvore Decisão*	*Redes Neurais*
[Zhao and Wang, 2015]	√	√		√	√	√			
[Li, 2015]	√	√		√		√		√	
[Rao et al., 2015]	√	√			√				√
[Li et al., 2014]	√	√		√	√				√
[Asad, 2015]	√	√		√	√		√		√
[Minev et al., 2012]		√	√						√
[Gu et al., 2015]		√			√	√			
[Wu et al., 2014]		√	√		√	√			
Trabalho Proposto			√	√			√	√	

Tabela 3.1: Comparativo entre os trabalhos com foco em aprendizagem para o mercado acionário e o trabalho sendo proposto.

3.3 Informações de Notícias, Redes Sociais e o Mercado Financeiro

Um princípio básico da economia é que os preços dos ativos mudam em resposta a novas informações. De forma geral, os desafios da previsão do mercado acionário baseado em

notícias e redes sociais estão em determinar quais são as fontes de maior correlação com cada mercado e como processá-las de modo a extrair de cada uma delas sinais que permitam aos investidores diminuírem o risco associados a seus investimentos. A seguir são apresentados trabalhos relacionados que alcançam êxito ao incorporar a análise de sentimento e utilização de redes sociais em modelos de predição de mercados acionários.

3.3.1 Evidências de Relação entre Notícias e o Mercado Financeiro

Processando uma base de notícias econômicas desde 1950 via análise textual, [Boudoukh et al., 2013] mostraram que a chance de se obter bons negócios no mercado americano em dias onde há notícias econômicas é 120% maior em relação a dias sem notícias. Os autores também mostraram que dias com poucas notícias são prováveis de haver reversões de mercado, enquanto dias com notícias tendem a continuidade do cenário, ou seja, de os preços continuarem como estão, sem alterações. Finalmente, reforçam que tais resultados são fortalecidos ao medir-se a polaridade das notícias. Conclusões semelhantes sobre a importância das notícias e suas relações com o mercado acionário também foram obtidas em outro cenário completamente diferente da bolsa americana, como na bolsa turca, e a bolsa de Tóquio, por exemplo. Ao processar as frequências de termos extraídos de notícias econômicas de jornais de alta circulação da Turquia e em seguida correlacioná-las de forma temporal, associando-as a 10 indicadores técnicos diferentes, tais como, médias móveis, bandas de *Bollinger*, taxa de variação de preços, índice de força relativa, entre outros, [Seker et al., 2014] encontrou forte correlação inversa entre a frequência de termos e a variação do índice turco, ou seja, quando a frequência de determinados termos aumentam ou diminuem nas notícias, o índice turco reflete diretamente essa variação. Via análise empírica, [Jayawardena et al., 2015] também confirmou o potencial da utilização de notícias na previsão de índices e sua relação com a bolsa de valores de Tóquio.

Atualmente, por ser natural que muitos investidores tenham acesso a uma riqueza de informações por meio de uma variedade de canais de notícias, modelos como os propostos por [Ma and Liang, 2015] e [Minev, 2013] vão além de confirmar relações entre notícias e as reações do mercado de ações. A partir de informações extraídas de notícias textuais online, os protótipos produzidos servem de apoio a decisões de investidores, reconhecendo oportunidades de risco e investimento, auxiliando-os em suas decisões de comprar e vender

ações durante a gestão de suas carteiras. Outros como o proposto por [Gallegos and Hau, 2015] buscam simular e melhorar a previsão de preços de ações, mas de apenas empresas específicas, por meio de uma abordagem de aprendizagem supervisionada textual de artigos publicados sobre essas ações.

3.3.2 Análise de Sentimento e Mercado Financeiro

Um outro caminho explorado por muitos pesquisadores é a previsão de tendências do mercado acionário baseado no humor do mercado que, em geral, é obtido por meio da análise de sentimento de notícias e postagens coletadas das redes sociais [Sukprasert et al., 2015].

Baseados no sentimento do investidor, [Wu and Olson, 2015] buscam explicar o comportamento do mercado acionário, suas tendências, percepções de riscos e perdas, e características que afetam a estratégia de investimento baseado em comportamentos em curso. Diferente de abordagens anteriores, em que o objetivo era medir sentimentos globais, [Nguyen et al., 2015] mediu apenas sentimentos sobre temas específicos ligados a 18 empresas. Apesar de restrita, essa perspectiva em delimitar a análise de sentimentos para um grupo de temas e empresas rendeu boa capacidade preditiva. De fato, ainda que o sentimento geral sobre um determinado mercado seja ruim, é provável que para algumas ações ou produtos, haja um sentimento contrário. Nesse sentido, os temas escolhidos pelo autor seriam suficientes para classificar corretamente o sentimento para as empresas selecionadas sem sofrer o ruído da classificação geral. O problema desse modelo é que por ser restrito, nem sempre há informações suficientes para alimentar o modelo de forma a gerar previsões consistentes.

3.3.3 Redes Sociais e o Mercado Financeiro

Além do processamento de notícias via análise semântica, léxica ou contagem de frequência de termos, há uma linha de análise do mercado financeiro que busca realizar previsões baseadas no processamento de informações obtidas por meio das redes sociais em especial o *Twitter* [Buscaldi and Hernandez-Farias, 2015].

Por meio do processamento das redes sociais, é possível explorar as opiniões postadas por pessoas e coletar, assim, conhecimento comum sobre acontecimentos recentes. Neste sentido, muitos trabalhos vêm se utilizando dessas características das redes sociais para pre-

ver preços futuros de ações [Skuza and Romanowski, 2015] e índices [Piñeiro-Chousa et al., 2015]. Em seu trabalho [Kadambari et al., 2015] concluiu que o número de seguidores do perfil de um usuário é o parâmetro que estabelece maior correlação com os resultados da previsão, mais até que os textos compartilhados.

Especificamente para o mercado chinês, o *Twitter* dá lugar ao *Sina Weibo*[20] onde [Gong and Sriboonchitta, 2016] descobriram que apesar deste mercado viver os pontos fracos de um mercado de transição com comportamentos especulativos seguido de forte intervenção do governo, as medições de comentários públicos coletados do *Sina Weibo* estão diretamente correlacionadas com o valor do índice Shanghai ao longo tempo. Também foi verificado que mensagens do governo em redes sociais ajudam a melhorar significativamente as performances do mercado.

Por fim, é importante comentar que a coleta de informações das redes sociais é bastante restrita. As APIs fornecidas pelas empresas, em geral, acessam unicamente conteúdo público que, em geral, é bastante limitado.

3.4 Ferramentas de Previsão do Mercado Financeiro baseadas em Notícias

Nesta seção são listados alguns trabalhos onde o foco principal foi a construção de ferramentas de previsão do mercado financeiro com base em notícias e informações oriundas de redes sociais.

- **StockTwits**: O trabalho de [Al Nasseri et al., 2015] descreve a utilização do *StockTwits*[21] proposto por [Oliveira et al., 2013] de forma a conceber um novo sistema inteligente comercial de apoio capaz de realizar predição de sentimento por meio da combinação de técnicas de mineração de texto e algoritmo de árvore de decisão. Tal sistema extrai termos que expressam sentimento específicos relacionados a (vender, comprar ou manter) a partir de mensagens de *micro-blogging* relacionados às ações. A abordagem foi utilizada para prever valores do índice *Dow Jones* e apresentou melhores resultados que o modelo aleatório.

[20] http://weibo.com/login.php

[21] http://stocktwits.com/

- **TheStockSonar**: The Stock Sonar[22] integra dicionário de sentimentos, padrões de composição de frase e eventos de predicados semânticos para compreender as notícias emitidas diariamente sobre o mercado de ações americano. Segundo [Feldman et al., 2011] o projeto é capaz de distinguir notícias que tratam de acordos, ações judiciais, novos produtos, e outros eventos relativos ao preço, ponderando informações relevantes à medida que atualiza a previsão de preço das ações segundo as notícias. Neste trabalho Feldman também estende a visão clássica de análise de sentimento comum da literatura associando sentimentos tanto a informações objetivas quanto a subjetivas. A imagem 3.1 apresenta um exemplo de funcionamento da ferramenta *the Stock Sonar*.

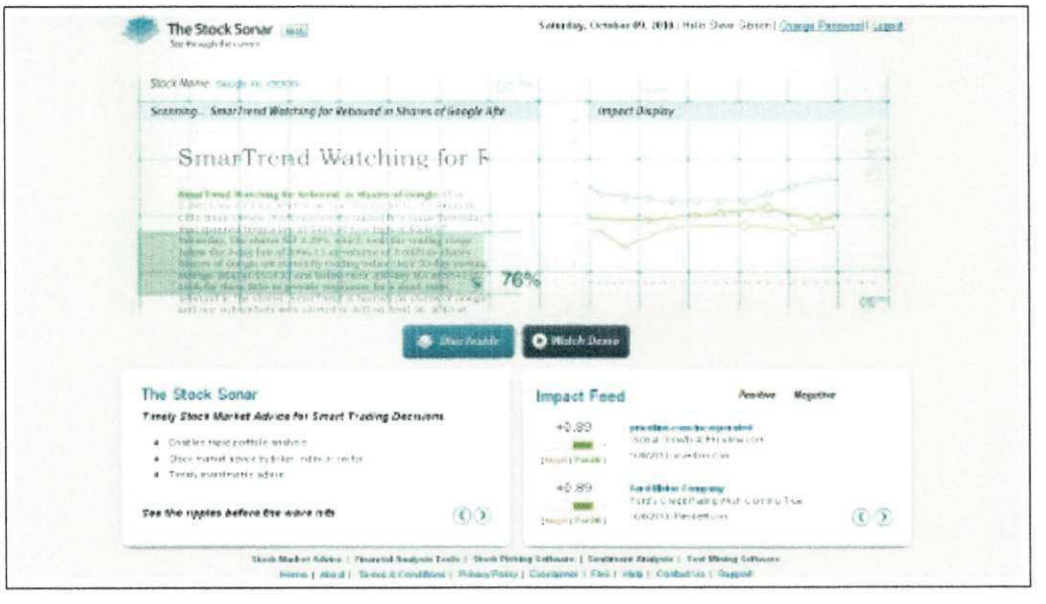

Figura 3.1: Ilustração da ferramenta the Stock Sonar.

A grande maioria dos trabalhos explora apenas uma única fonte de informação, seja ela notícia, comentários ou postagens em redes sociais. Aqui é discutida a utilização de mais redes sociais e suas correlações com o mercado acionário. Por fim, o ambiente no qual o trabalho se aplica é a bolsa de valores brasileira, ao seus setores e ao seu índice mais relevante, o IBOVE. A Tabela 3.2 apresenta uma comparação entre este trabalho e os trabalhos citados anteriormente que trataram de análise de sentimento e redes sociais na dinâmica do mercado de ações. Nosso trabalho se propôs a investigar elementos pouco explorados da

[22] www.thestocksonar.com

dinâmica do mercado de ações como as relações que as manifestações de usuários das redes sociais Facebook, Google Plus e LinkedIn mantém com o mercado acionário brasileiro. Nenhum dos trabalhos revisados explorou essas redes sociais e seus impactos no mercado acionário nacional ou internacional. Apesar de viver um mau momento, é importante mencionar que a BM&FBOVESPA ainda figura entre das bolsas mais influentes do mundo, porém, em comparação com outras bolsas de valores, ainda possui sua capacidade preditiva pouco explorada.

3.5 Considerações Finais

Recentemente muitos trabalhos vêm apresentando evidências contrárias a hipótese de eficiência de mercado. Em diferentes cenários financeiros, pesquisadores mostram que os inúmeros avanços tecnológicos não previstos ao momento da formulação desta hipótese – como técnicas de aprendizagem de máquina, jornais online, mídias sociais, e até o próprio avanço dos computadores – são capazes de explicar, até certo ponto, o comportamento de bolsas ao redor do mundo.

O levantamento da literatura confirma tanto o aspecto inovador, quanto a contribuição que este trabalho traz à área de análise de mercados. Desde uma perspectiva técnica, este trabalho explora o potencial não apenas de *micro-blogs* como o *Twitter*, mas de diversas redes sociais tais como *LinkedIn*, *Facebook* e *Google Plus* associando-os à análise de sentimento de notícias de modo a lançar luz sobre o impacto que essas mídias digitais tem no mercado acionário brasileiro. Os *insights* obtidos foram utilizados em um modelo de aprendizagem de máquina de modo a prever os próximos movimentos tanto do índice Bovespa quanto de cada um dos setores da BM&FBOVESPA. Como será visto no decorrer do documento, os resultados obtidos forneceram evidências sobre a vantagem estratégica da utilização decorrente de informações de notícias e redes socais na predição de eventos do mercado nacional.

				Análises		
Trabalhos	**Notícias**	**Comentários**	**Twitter**	*Sentimento*	*Técnica*	**Previsão**
[Boudoukh et al., 2013]	√			√		√
[Seker et al., 2014]	√				√	
[Jayawardena et al., 2015]	√				√	√
[Ma and Liang, 2015]	√					√
[Wu and Olson, 2015]		√		√		√
[Buscaldi and Hernandez-Farias, 2015]			√	√		√
[Kadambari et al., 2015]		√	√	√		
[Gong and Sriboonchitta, 2016]		√		√	√	√
[Sukprasert et al., 2015]	√			√		
[Piñeiro-Chousa et al., 2015]		√	√	√		
[Skuza and Romanowski, 2015]		√	√	√		√
[Al Nasseri et al., 2015]			√		√	√
[Gallegos and Hau, 2015]	√				√	√
[Feldman et al., 2011]	√				√	√
Trabalho Proposto	√	√	√	√		√

Tabela 3.2: Comparativo entre o trabalho sendo proposto e outros trabalhos que utilizaram informações de notícias, redes sociais e análise de sentimento em suas análises ou previsões de mercados acionários.

Capítulo 4

Coleta e Preparação de Dados

A seguir são detalhados os processos de coleta de notícias e extração de métricas para a construção da base de dados utilizada por todos os experimentos.

4.1 Coleta de Notícias

Os dados dessa pesquisa foram coletados dos cadernos econômicos dos jornais de domínio público G1, Folha de São Paulo, Estadão e de veículos direcionados aos investidores do mercado de ações tais como Reuters e Infomoney. Apesar de coletadas, as notícias dos veículos Reuters e Infomoney não foram utilizados em nenhuma análise deste trabalho devido a razões éticas de utilização das fontes e por constituírem um tipo diferente de notícias, mais detalhadas, direcionadas a um público especializado, diferente das notícias públicas pelos outros jornais considerados.

Para cada uma das notícias coletadas, buscou-se extrair a maior quantidade de atributos possíveis. Tanto relacionados ao estilo (e.g., jornalista, jornal, horário) quanto ao conteúdo (e.g., título, subtítulo). Especificamente, os atributos extraídos das notícias coletadas foram: *timestamp* da publicação da notícia, *título da notícia*, *subtítulo da notícia*, *jornalista que escreveu a notícia*, o *jornal que a publicou*, a *quantidade de repercussão no Facebook*, *Twitter*, *LinkedIn*, *GooglePlus*, a *quantidade de comentários* na página da notícia e sua *polaridade*. Para cada atributo quantitativo descrito anteriormente, foram também consideradas as médias, medianas, desvios-padrão e variâncias. Por exemplo, para o atributo *quantidade de repercussão no Facebook*, também foram calculadas a média da quantidade de repercus-

são via Facebook, mediana da quantidade de repercussão, desvio-padrão e variância dessa quantidade.

A seguir serão detalhadas os recursos utilizados e o processo de extração de notícias em cada veículo.

4.1.1 Recursos de Hardware

Para o desenvolvimento dessa pesquisa, foram utilizados dois computadores com as seguintes configurações:

1. **Maquina real:** 279,3 GB de HD, 64-bit, 6 CPU X5650, 2.67GHz e 3.8 GB de RAM.
2. **Máquina virtual:** 8G de RAM, 90G de HD e 4 processadores virtuais.

4.1.2 G1

O jornal G1 está disponível no *link* **www.globo.com**. As notícias de seu caderno de economia podem ser obtidas por meio do *link* **http://g1.globo.com/economia/noticia/plantao.html#1**. Incrementando o número que sucede o # é possível ter acesso a um novo conjunto de notícias (até o limite de notícias disponíveis pelo jornal).

A Figura 4.1 apresenta em detalhes o formato explorado para obter as notícias do caderno de economia do jornal G1.

4.1.3 Folha de São Paulo

O jornal online Folha de São Paulo está diponível no *link* **http://www.folha.uol.com.br/**. Diferente do G1, o caderno de economia da Folha de São Paulo apresenta-se bastante amplo e dividido em inúmeras seções relativas a economia: índices, Brasil, salões de automóveis, mercados, entre outros. Apenas notícias da seção **mercados** foram coletadas.

O mecanismo de busca[23] presente no jornal foi utilizado para realizar requisições sucessivas de coleta. Faz-se necessário, pela forma como está organizada a busca, que o usuário digite palavras de interesse para que sejam retornadas matérias relacionadas. Nesse ponto, e

[23] http://search.folha.com.br/search?

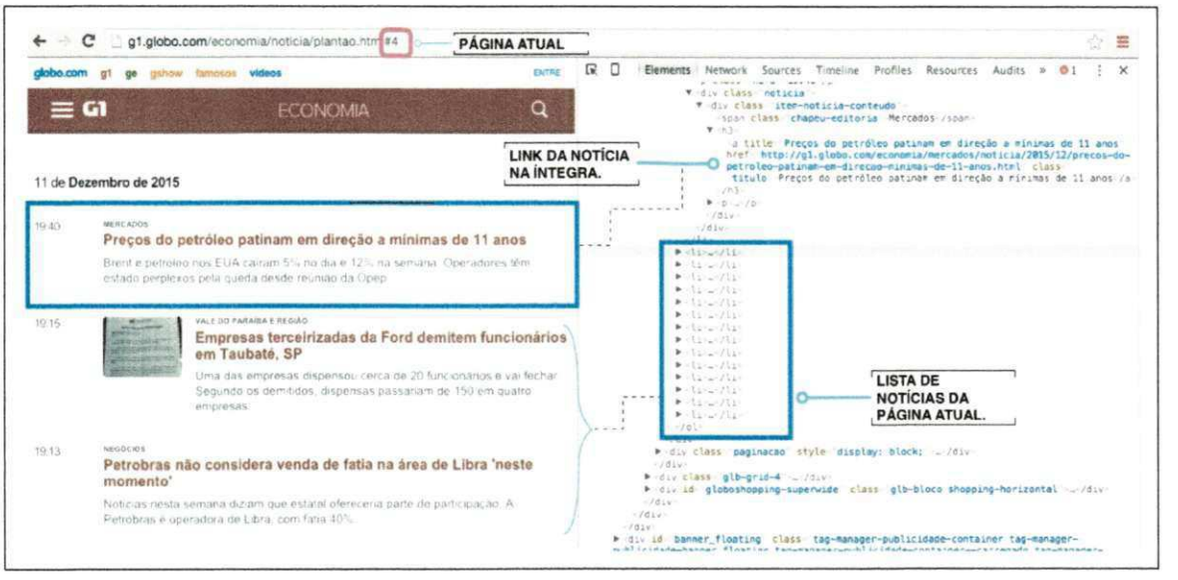

Figura 4.1: Estrutura explorada para criar o *script* coletor de notícias do jornal G1.

após várias tentativas com *palavras-chave* do contexto econômico, "mercado"foi a que apresentou o maior número de notícias retornadas, e, assim sendo, foi a semente de coleta para as notícias do jornal Folha de São Paulo.

A Figura 4.2 apresenta em detalhes a utilização do sistema de busca presente no jornal Folha de São Paulo e explica as decisões tomadas para construção da coleta de notícias.

4.1.4 Estadão

O jornal online Estadão diponível no *link* **http://www.estadao.com.br**, também teve as páginas coletadas por meio de seu mecanismo de busca[24]. Diferente de outros jornais, as informações no momento da publicação não encontravam-se presentes na estrutura *HTML* da página, sendo necessário o processamento textual dessa estrutura para coletar essas informações. A Figura 4.3 apresenta a estrutura de busca explorada para a coleta de notícias do jornal Estadão e enfatiza o problemática de obtenção do *timestamp*.

[24]http://busca.estadao.com.br/?editoria[]=Economia&pagina=

Figura 4.2: Estrutura explorada para criar o *script* coletor de notícias do jornal Folha de São Paulo.

Figura 4.3: Estrutura explorada para criar o *script* coletor de notícias do jornal Estadão.

4.1.5 Resumo Geral

Conforme apresentado pela Tabela 4.1, foram coletadas **471.430** notícias de domínio público considerando os jornais G1, Folha de São Paulo e Estadão entre 2000 e Março de 2015.

Jornal	Quantidade
G1	185.733
Folha de São Paulo	113.671
Estadão	172.026
Total	471.430

Tabela 4.1: Número total de notícias coletadas para cada um dos jornais de domínio público.

4.2 Repercussão

Para cada notícia coletada foi realizado o cálculo de repercussão (engajamento) que consiste em contar quantos leitores compartilharam o *link* dessa notícia em alguma das redes sociais analisadas – Facebook, Twitter, LinkedIn, GooglePlus, assim como, comentários recebidos na própria página.

Para obter essa informação de forma precisa, buscou-se utilizar as *APIs* oficiais de cada rede social. Algumas redes sociais como o Facebook e o Twitter produziram várias versões de suas *APIs* ao longo dos anos. Para encontrar o número de repercussão de cada notícia, este trabalho considerou as versões das *APIs* de cada rede social utilizadas pelos jornais online e contidas no interior do código HTML de cada notícia.

De fato, medir a repercussão real de cada notícia implicaria em realizar muito mais do que isso. Para tal, seria necessário saber não apenas a quantidade de compartilhamentos, mas além disso, a quantidade de vezes que a página da notícia foi lida, a quantidade de comentários que a notícia obteve quando compartilhada nas redes sociais, e assim por diante. Infelizmente, muitas destas informações não estão disponíveis. Apenas os próprios jornais online são capazes de saber, por exemplo, quantas vezes uma notícia foi solicitada. O objetivo dessa discussão é esclarecer que a repercussão sendo medida é uma simplificação necessária da realidade, porém, bastante útil para uma compreensão geral do cenário de interesse.

4.3 Tradução de Notícias

Cada uma das notícias coletadas foi traduzida de forma automática para o idioma Inglês para que fosse possível utilizar algoritmos estado da arte de análise de polaridade. A racionalidade por trás disso está no fato de que grande parte dos métodos são baseados em léxicos e esses léxicos (baseado em dicionário, i.e., o conjunto de vocábulos de uma língua) acabam sendo preservados na tradução automática [Araújo et al., 2016].

Empresas como Microsoft e Google possuem os serviços de tradução de texto Bing Tradutor e Google Translate, respectivamente. Ambas cobram pela utilização desse serviços e apenas uma quantidade muito pequena de traduções são possíveis de serem realizadas de forma gratuita.

A Microsoft disponibiliza por meio da *Azure Marketplace* o Microsoft Translator[25] onde é possível traduzir até 4.000.000 caracteres/mês ao custo de €29,89. Já a Google fornece o serviço de cloud para tradução de texto via Translate API[26] onde são cobrados $20 a cada 1.000.000 caracteres traduzidos com um limite de 2.000.000 de caracteres/dia.

Com uma base de notícias contendo 985.865.517 caracteres, e utilizando as versões básicas de cada plano seriam necessários:

- Microsoft Translator: $\sim$ 247 meses e €7.400.
- Translate API: $\sim$ 16 meses e $9.859.

Apesar de solicitado formalmente a utilização dos serviços de tradução de ambas as empresas para os fins científicos, nenhum pedido foi aceito.

Por fim, foi desenvolvido um *script* que utiliza a ferramenta do Google Tradutor[27] online para realizar as traduções necessárias. Basicamente o algoritmo divide o texto da notícia em frases e para cada frase é criada uma requisição de tradução. Após todas as partes serem traduzidas, são unidas em um único texto e acrescentadas a base de dados.

Com os recursos apresentados na seção 4.1.1 funcionando 24h por dia a tarefa pôde ser concluída em duas semanas.

[25] https://datamarket.azure.com/dataset/bing/microsofttranslator

[26] https://cloud.google.com/translate/v2/pricing

[27] https://translate.google.com.br/?hl=pt-BR

4.4 Armazenamento de Notícias

Todas as notícias coletadas foram armazenadas em um banco de dados *NoSQL* **MongoDB**[28]. Essa escolha é justificada pelas seguintes características desse tipo de banco de dados:

- Maior velocidade de inserção de dados em comparação com bancos de dados relacionais (onde são verificadas inúmeras condições de integridade).
- Esquema flexível, sendo possível acrescentar e remover atributos sem a necessidade de reestruturar o esquema de dados.
- Fácil replicação em diferentes ambientes.
- Possibilidade de escrever consultas complexas em JavaScript (mesma linguagem da programação Web) eliminando, assim, a necessidade de aprender uma nova linguagem/paradigma apenas para manipular os dados inseridos no bando de dados.

4.5 Validação da Base de Dados

Para verificar a integridade das informações presentes na base de dados foram realizadas as seguintes atividades:

1. Verificou-se a correspondência entre a quantidade de notícias retornadas via ferramentas de consulta do jornal e a quantidade de notícias presentes na base de dados.
2. Notícias que possuíam algum atributo vazio foram novamente coletadas e reinseridas.
3. Notícias que possuíam o mesmo *timestamp* e título foram retiradas por serem consideradas duplicadas.
4. Para cada jornal foram selecionadas 50 notícias aleatoriamente e verificado se todos os atributos estavam conforme esperado.

[28]https://www.mongodb.org/

4.6 Desafios, Incoerências e Limitações

A seguir são apresentados desafios, incoerências e limitações encontradas por esse trabalho na atividade de coleta e preparação dos dados:

1. É um grande desafio automatizar o processo de coleta de páginas de notícias online por longos períodos. Nomes de variáveis, disposição de conteúdos e até a própria estrutura da linguagem HTML mudam ao longo do tempo e isso implica em realizar muitos ajustes até o final do processo.

2. O cálculo de repercussão de notícias é variável ao longo do tempo. As notícias continuam sendo acessadas, comentadas e compartilhadas. Por isso, acredita-se que quanto mais antiga a notícia menor será o interesse nela, entretanto, a lógica se inverte para as notícias mais recentes que foram coletadas. É provável acontecer que após a coleta da repercussão de uma notícia seus valores de repercussão estejam diferentes.

3. Para todos os alvos considerados neste trabalho o estado de consolidação foi considerado como sendo qualquer variação igual ou inferior a 0.5%. Esse fator abre duas discussões principais. A primeira delas é que qualquer tendência que se estabeleça de forma gradativa em valores iguais ou inferiores a 0.5 não será percebida pelo método e essas oportunidades serão perdidas. A segunda é que ainda que 0.5% seja um valor aceitável para a variável alvo preço, conforme verificou-se empiricamente, para as outras variáveis alvo como volume financeiro, por exemplo, essa porcentagem por ser insuficiente para estabelecer um limiar para consolidação. É importante que no futuro sejam realizados estudos que apontem limiares ideias para todas as variáveis alvo.

4. Durante o processo de coleta de repercussões de notícias, o limite de submissões às APIs do Twitter e Facebook eram rapidamente atingidos. Nesses casos, novas tentativas eram realizadas em momento futuro até o retorno ser satisfatório.

5. Algumas notícias possuem associadas ao seu conteúdo campanhas publicitárias de alguma natureza. Quando essas informações encontravam-se inseridas dentro do conteúdo da mensagem não foi possível retirá-la e ela é considerada junto ao conteúdo da notícia.

6. Há um pequeno número de notícias do mesmo jornal que apresentam a mesma informação. Apesar de terem sido removidas notícias de mesmo *timestamp* e título, com o passar do tempo foi percebido que algumas notícias foram atualizadas em momento posterior (diferente timestamp) e sofreram ajuste no título, ainda assim informando sobre o mesmo acontecimento.

7. É fato de que, por vezes, notícias de outros cadernos, como por exemplo, o caderno de política, impactam diretamente na economia, entretanto, esta pesquisa limitou-se a coletar notícias apenas dos cadernos de economia.

8. O jornal G1 disponibiliza notícias apenas a partir de 2010 apesar de estar em circulação desde 2006.

9. Entre os anos [2001,2006] há apenas publicações do jornal Estadão. Por algum motivo, a Folha de São Paulo não liberou notícias desse período como evidenciado no Apêndice AB.1.

10. Foram encontradas notícias publicadas pelo jornal Estadão referindo-se a acontecimentos contemporâneos porém com data de 1969. Acredita-se que seja devido a algum modelo de funcionamento interno para publicação de notícias.

4.7 Reprodutibilidade

Todo o código utilizado por este trabalho pode ser acessado por meio do *link* https://github.com/zegildo/PhD.

4.8 Considerações Finais

A coleta de notícias é uma atividade desafiadora e complexa. Cada um dos jornais online analisados neste trabalho possuem arquitetura e conjuntos de tecnologias diferentes para os quais houve a necessidade de construir mecanismos de coleta específicos para cada um deles. A própria tecnologia evolui ao longo do tempo e as soluções apresentadas precisam adequar-se a esta realidade. Montar a base de dados de notícias, verificar sua integridade, e torná-la apta a ser explorada foi uma das atividades mais onerosas deste trabalho.

Capítulo 5

Análise Descritiva

Este capítulo apresenta a análise descritiva que foi realizada com os dados das notícias previamente coletados.

Esta análise buscou compreender o *modi operandi* referente as publicações por jornal e repercussão dessas notícias por parte de seus leitores de forma a utilizar os sinais encontrados como atributos informativos na construção de modelos preditivos.

A seguir cada seção detalhará um aspecto diferente da análise.

5.1 Análise das Quantidades

Nesta seção é avaliada a regularidade de publicações de notícias econômicas por cada jornal em diferentes granularidades de tempo. O objetivo dessa investigação foi o de conhecer algumas características de publicações diárias dos jornais, analisar os limites entre quantidades normais e eventuais *outliers*, investigar se há diferenças significativas na quantidade de publicação dos jornais em algum ano, mês ou dia específico, discutir seus eventuais porquês e apresentar hipóteses.

5.1.1 Ano

Nesta seção é analisada a quantidade de publicações de notícias econômicas nos jornais ao longo dos anos, seu índice de tendência central e sua variabilidade.

A hipótese inicial era de que a variação entre a quantidade de notícias publicadas por ano

entre os jornais era praticamente equivalente devido ao objetivo de informarem seus leitores sobre fatos e eventos econômicos, relevantes no contexto nacional.

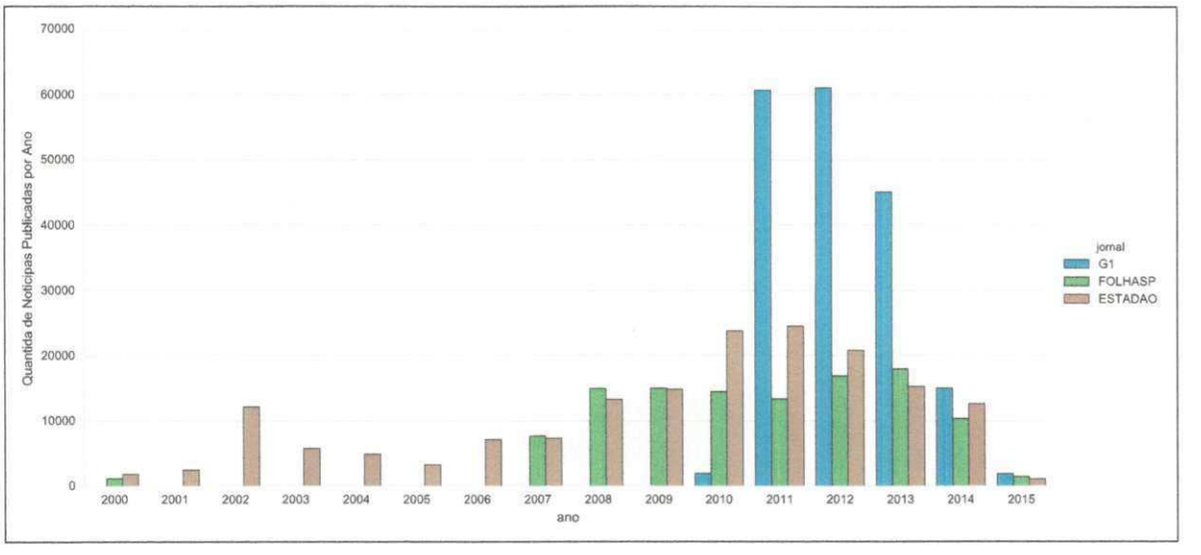

Figura 5.1: Quantidade de notícias publicadas por cada jornal entre 2000 e 2015.

A Figura 5.1 mostra a quantidade de notícias por ano publicadas por jornal e evidencia o que foi discutido na seção 4.6 do capítulo anterior. Como nem todos os jornais possuem notícias ao logo de todos os anos, restringimos as análises aos períodos de Janeiro de 2010 à Março de 2015 como evidenciado pela Figura 5.2.

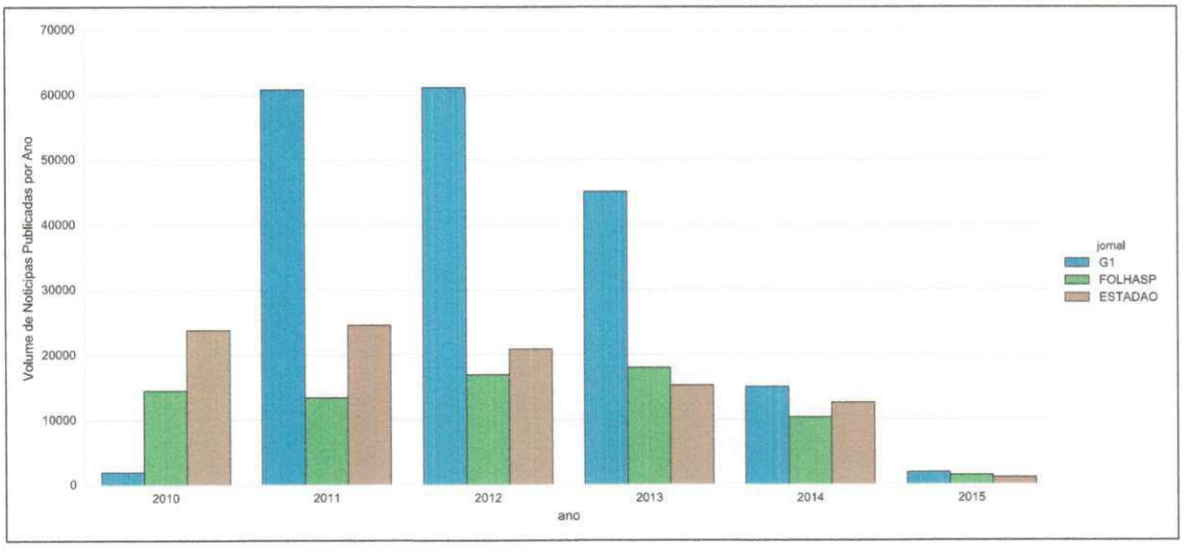

Figura 5.2: Quantidade de notícias publicadas apenas entre 2010 e 2015.

Variabilidade da Quantidade de Publicações por Dia

A variabilidade entre a quantidade de publicações diárias dos jornais pode ser verificada por meio de dois gráficos *bloxplot* apresentados pelas Figuras 5.3 e 5.4. É possível perceber por meio da Figura 5.3 que a variabilidade da quantidade de publicações do jornal G1 é bastante superior aos demais jornais. Além disso, é possível ainda verificar uma notória descontinuidade a partir do valor 100 bastante atípica em relação a outros jornais. Por fim, verifica-se que a tendência central de todos os jornais está aparentemente próxima umas das outras. A Figura 5.4 detalha a análise anterior segmentando a análise ano a ano e a partir dela uma nova percepção é observada. É possível perceber que o jornal G1 apresenta bastante variabilidade no número de publicações econômicas diárias ao longo dos anos. Pelo índice de tendência central, percebe-se que em média o jornal G1, mesmo não sendo considerado uma mídia especializada em assuntos econômicos, vêm publicando diariamente uma quantidade de notícias econômicas que supera em dobro a soma do que é publicado pela Folha de São Paulo e Estadão.

Por fim, também é possível perceber que as variações do índice de tendência central do jornal G1 vêm sendo fielmente acompanhados pelo jornal Estadão ao longo dos anos.

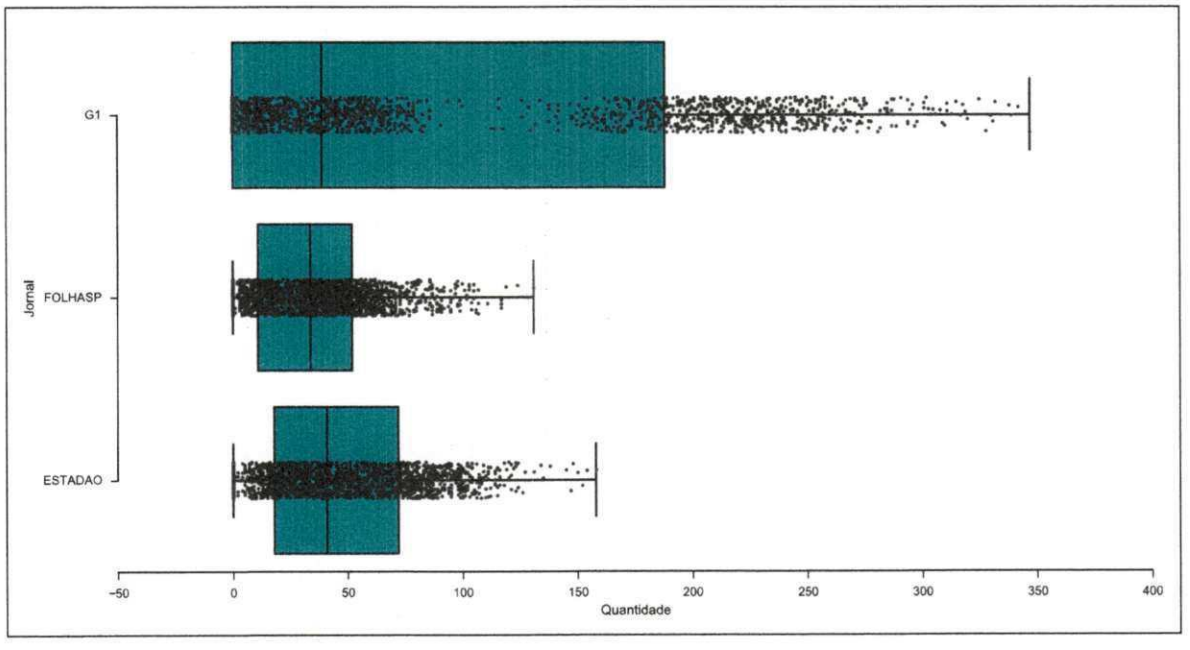

Figura 5.3: *BloxPlot* da quantidade de notícias publicadas para cada jornal diariamente. O eixo x representa o número de publicações diárias por cada jornal.

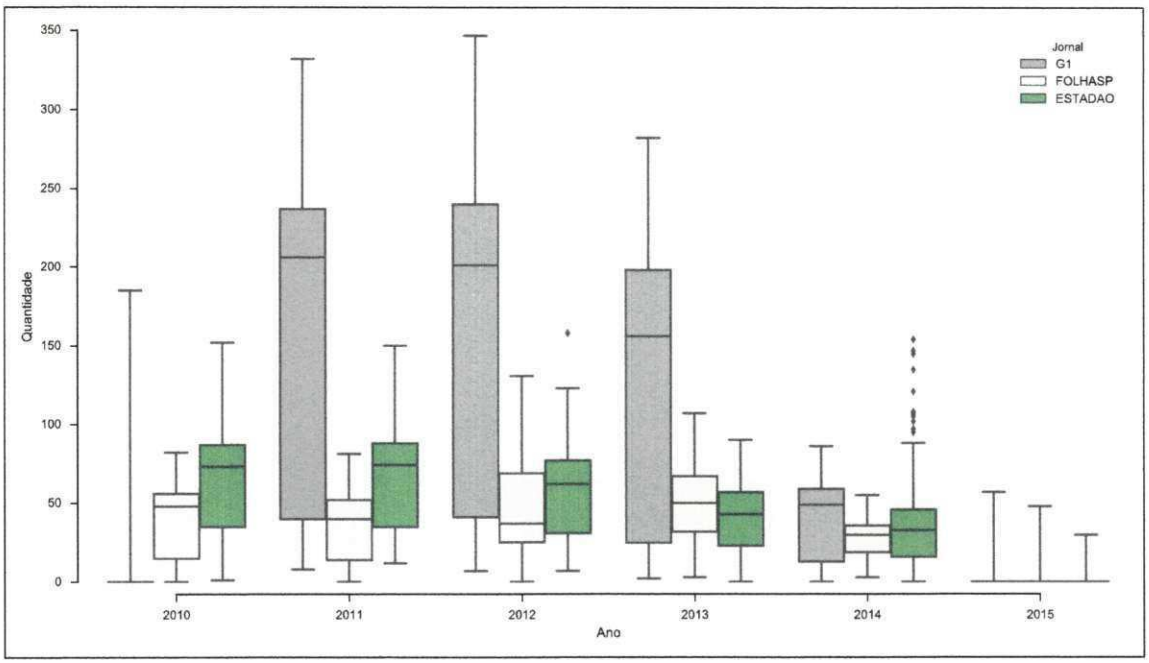

Figura 5.4: *BloxPlot* da quantidade de notícias publicadas de todos os jornais entre 2010 e 2015.

5.1.2 Densidade de Probabilidade

A função densidade de probabilidade é utilizada aqui para informar o comportamento da quantidade de publicações econômicas diárias para cada jornal analisado.

Folha de São Paulo

A Figura 5.5 apresenta a densidade de probabilidade para a quantidade de publicações econômicas diárias do jornal Folha de São Paulo. É possível constatar que a variabilidade das publicações diárias encontram-se entre 0 e 120. Nota-se que a partir de 50 publicações há uma diminuição gradativa desta probabilidade à medida que a quantidade de publicações diárias aumenta.

Dito de outra forma, a maioria dos dias possuem poucos eventos econômicos a serem reportados. Alguns eventos econômicos mais complexos carecem de mais publicações, seja para apresentar diferentes pontos de vista ou maior riqueza de detalhes o que justificaria dias que variam quase que uniformemente até as 50 publicações diárias. A partir disso, apresentam-se cenários econômicos atípicos a ponto de necessitarem de mais de 50 notícias

diárias para informar os fatos, como por exemplo, a paralisação do mercado Chinês após uma queda de mais de 7% e a derrocada das ações da OGX.

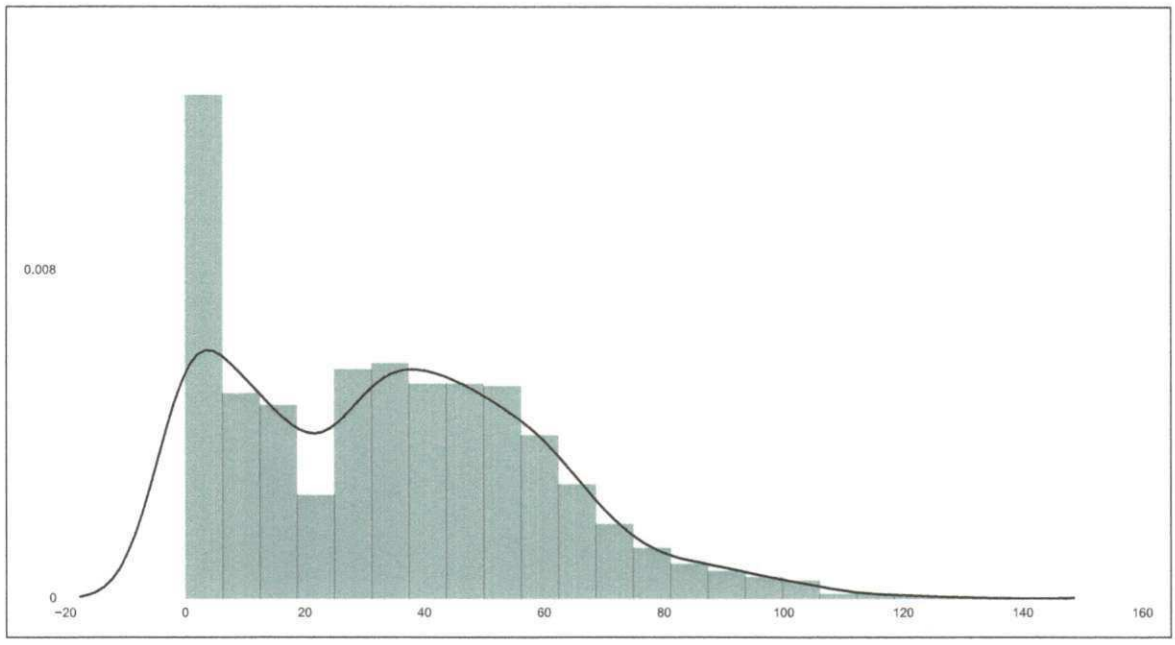

Figura 5.5: Densidade de probabilidade da quantidade de publicações econômicas para o jornal Folha de São Paulo.

Estadão

A Figura 5.6 apresenta a densidade de probabilidade para a quantidade de publicações econômicas do jornal Estadão. Do ponto de vista da amplitude da quantidade é possível observar uma variação de 0 a 150 o que é bem próximo ao observado pela Folha de São Paulo. A partir de 75 publicações diárias se inicia uma tendência de queda, demonstrando a raridade de dias com um número de publicações diárias superior a esse limiar.

G1

A Figura 5.7 apresenta a distribuição bimodal verificada para o jornal G1. Percebe-se por meio desta figura uma incoerência quanto a normalidade da quantidade de publicações verificada pelos outros jornais. É possível verificar no gráfico que é mais provável que um determinado dia tenha entre 200 e 250 notícias publicadas que entre 50 e 150 notícias. Outro fato curioso é a presença de dias com mais de 300 publicações econômicas. É bastante

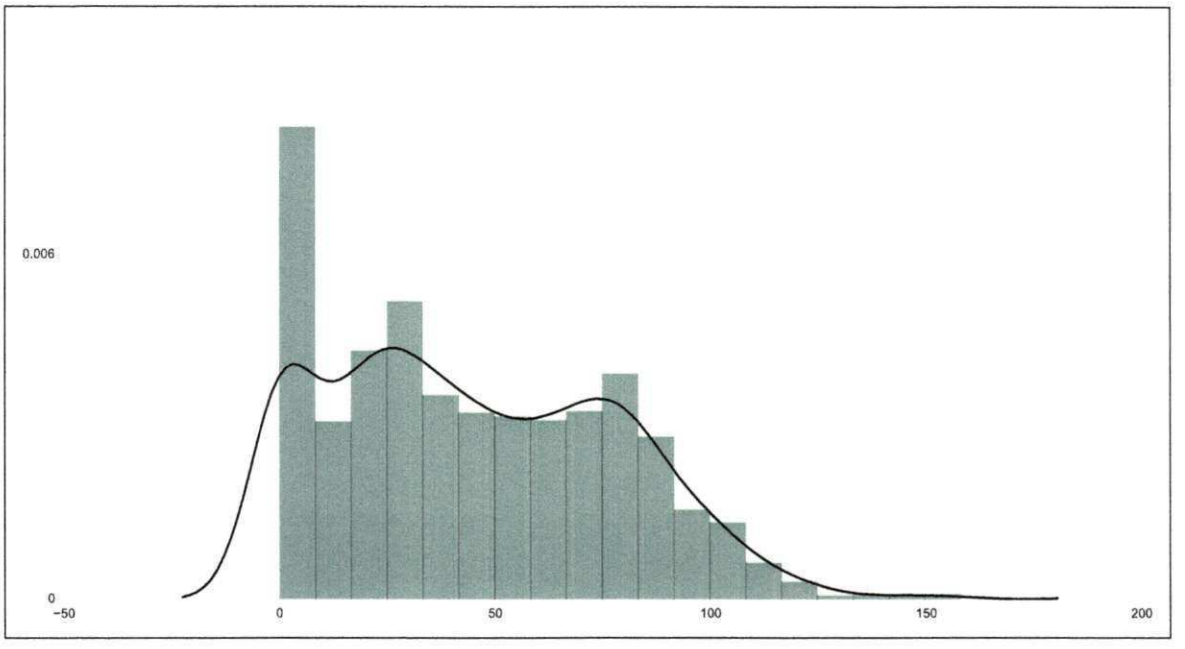

Figura 5.6: Densidade de probabilidade da quantidade de publicações econômicas para o jornal Estadão.

improvável que leitores sejam capazes de acompanhar 300 publicações em 8 horas de funcionamento do mercado acionário.

5.1.3 Mês

A análise por mês buscou verificar a regularidade da quantidade de publicações dos jornais ao longo dos meses para saber se algum mês apresenta maior tendência de publicações em relação à outros meses como [Vega, 2010] evidenciou para a bolsa americana.

A Figura 5.8 apresenta um agregado da quantidade de notícias publicadas entre os meses do ano para cada jornal. É possível observar que o mês de agosto apresenta maior quantidade de publicações econômicas para todos os jornais. Via de regra, agosto é o mês em que o governo apresenta ao Congresso Nacional projeções de orçamento para o próximo ano, previsões de crescimento econômico, inflação, e salário mínimo. Assim sendo, é provável que esse fato explique esse aumento. Após o mês de agosto registra-se uma queda contínua na quantidade de publicações econômicas em todos os jornais contrariando a hipótese de que o mês de outubro, devido as eleições, apresentava tal comportamento. O Apêndice C.1 apresenta detalhes da variabilidade para cada um dos jornais separadamente.

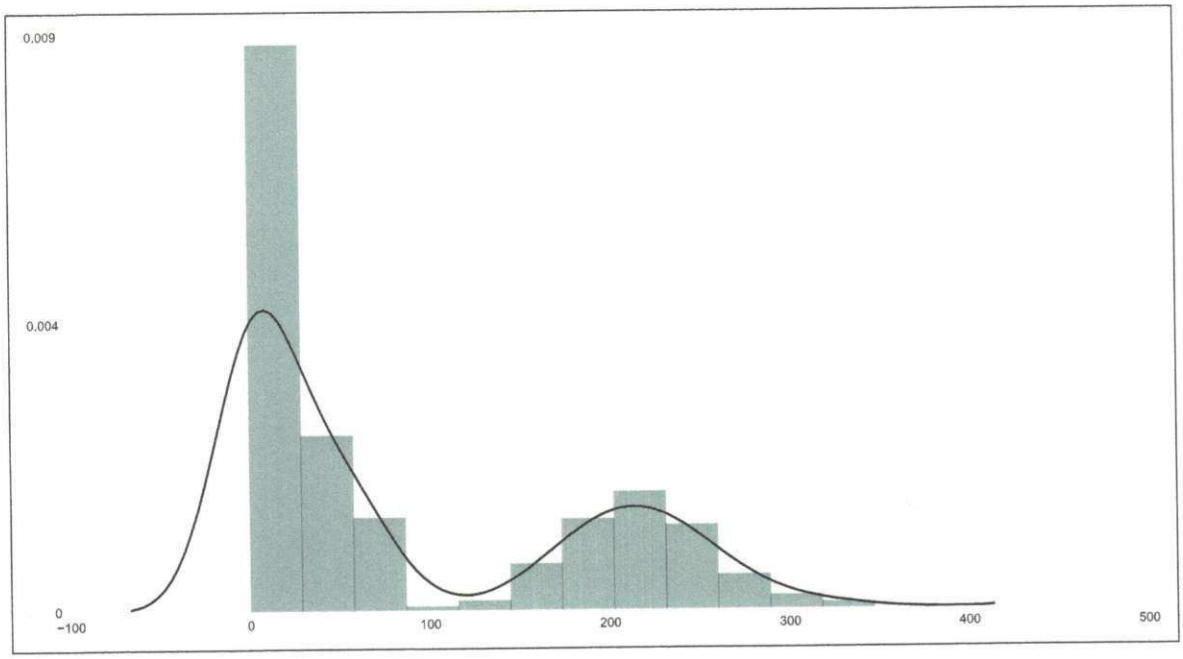

Figura 5.7: Densidade de probabilidade da quantidade de publicações econômicas para o jornal G1.

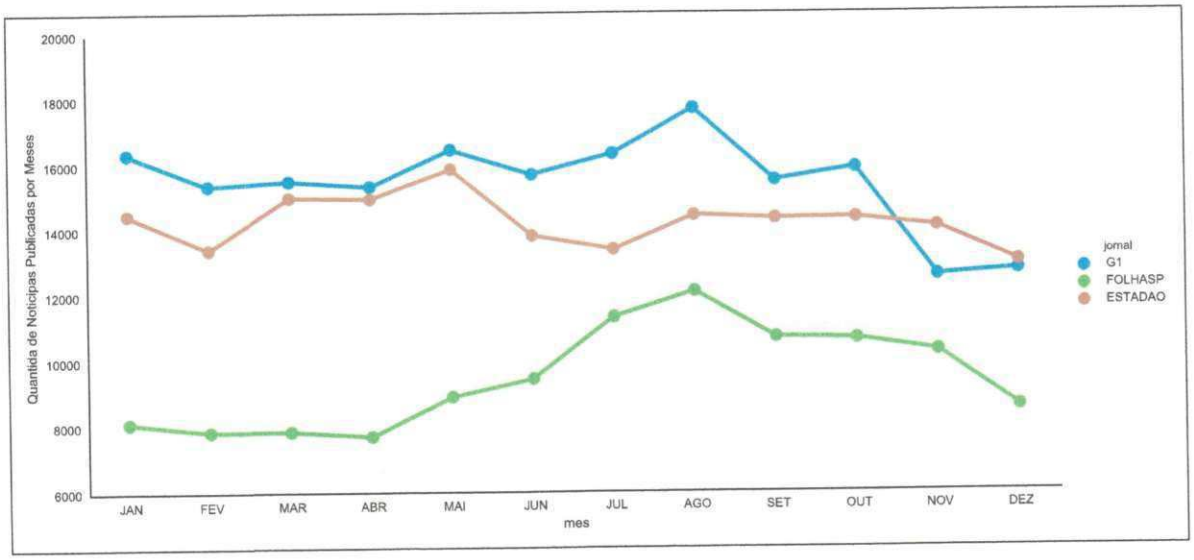

Figura 5.8: Quantidade de notícias publicadas por mês para todos os jornais.

5.1.4 Dia

A Figura 5.9 apresenta a variabilidade do número de publicações agrupadas por dia do mês. A hipótese inicial era de que a variação entre o número de publicações durante os dias do mês era imperceptível. Entretanto, é possível observar evidências contrárias para os dias 14 e 27. É provável que o dia 14 seja explicado, provavelmente, devido ao fechamento dos contratos de índices ocorrerem nas quartas-feiras mais próximas ao dia 15 de meses pares.

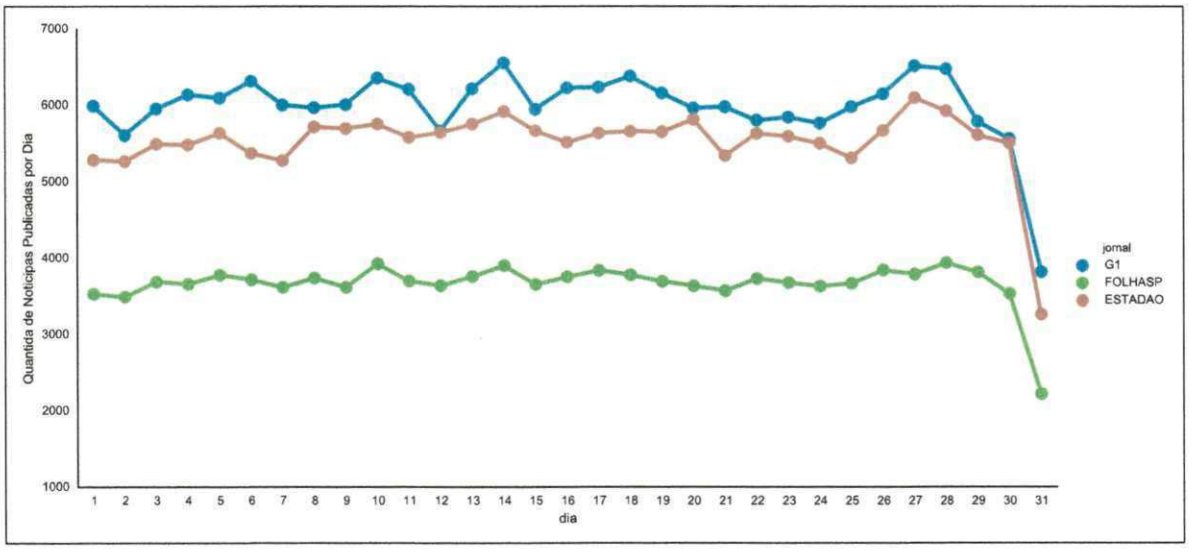

Figura 5.9: Quantidade de notícias publicadas por dia para todos os jornal.

5.1.5 Dia da Semana

A Figura 5.9 apresenta a variabilidade do número de publicações agregadas por dias da semana para todos os jornais. A hipótese inicial era de que nos finais de semana haveria poucas publicações e que para os demais dias da semana o número de publicações seria uniforme. É possível observar que parte da hipótese foi observada, de fato. Sábados e Domingos possuem um número de publicações menor que os demais dias da semana, entretanto, as quartas-feiras apresentam uma sutil superioridade em relação aos demais dias. Provavelmente explicada pela regra de fechamento dos índices. O Apêndice C.2 apresenta detalhes do comportamento de cada um dos jornais separadamente.

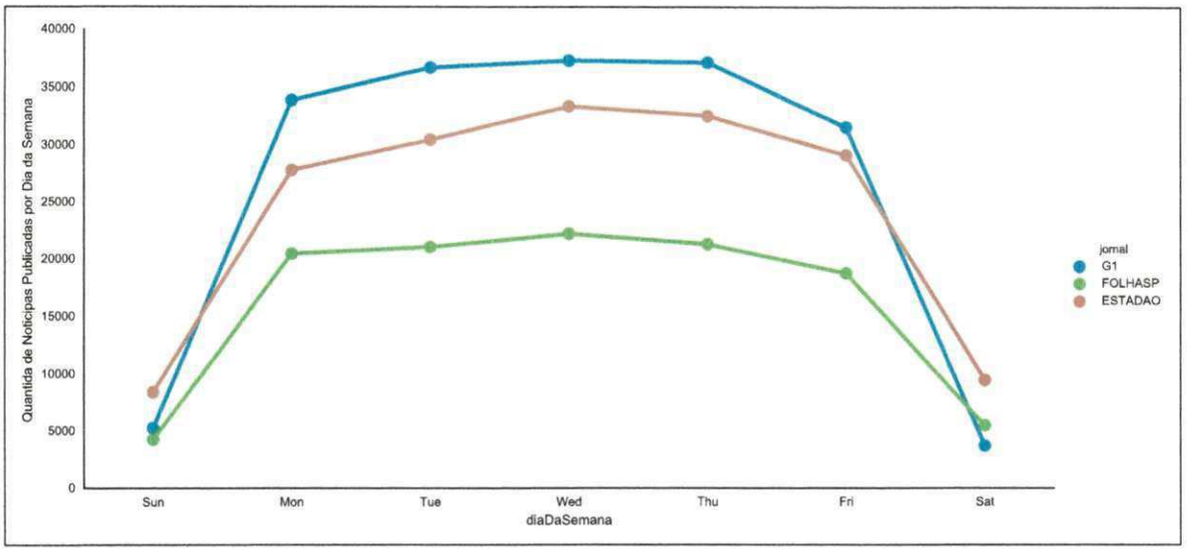

Figura 5.10: Quantidade de notícias publicadas por dia da semana para todos os jornal.

5.1.6 Discussão

A seguir são apresentadas discussões para para a análise do número de publicações:

1. A hipótese de que o número de publicações diárias por jornal é semelhante não é sustentada pelas evidências. O comportamento bimodal apresentando pelo jornal G1 em todas as análises de diferentes granularidades apresenta uma característica de normalidade nos extremos, ou seja, dias com poucas notícias econômicas e dias com abundante número de publicações são igualmente prováveis. Há dias com mais de 300 publicações, o que notoriamente torna difícil o acompanhamento por seus leitores. O fato é que muitas notícias que enfatizam um determinado ponto de vista sobre algum tema relevante pode influenciar a opinião das pessoas e, de certa forma, influenciá-las em suas decisões.

2. A hipótese de que a variação entre a quantidade de publicações econômicas entre jornais é semelhante não se sustenta. Em número de publicação e variabilidade, o jornal Folha de São Paulo e Estadão apresentam-se semelhantes. O jornal G1 apresenta comportamento particular em ambos os critérios.

3. O jornal Estadão apresentou menor variabilidade do número de publicações diárias e tendência central em todas as análises realizadas. Sendo assim, o jornal apresentou-se

como o menos vulnerável a *outliers*[29] entre todos os analisados.

4. Durante os dias da semana, a quarta-feira é provavelmente o dia onde haverá mais notícias econômicas. Por fim, as quartas e quintas-feira são dias mais suscetíveis a *outliers*.

5. Há dias onde o número de publicações econômicas ultrapassa 600 publicações para os jornais analisados. Para um investidor que deseja respaldar sua compra ou venda de ações na análise de notícias econômicas terá de realizar um trabalho bastante duro ao ler, interpretar e decidir. O desenvolvimento de um sistema de software que permitisse a automatização de análises fundamentalistas por si só já seria uma contribuição relevante aos investidores de varejo.

6. Do ponto de vista do investidor, tendo a necessidade de escolher um mês para analisar com mais atenção as notícias, esse mês seria Agosto. Entre os dias do mês, os dias 14 e 27. Já entre os dias da semana, as quartas-feira. Já entre os jornais analisados, a Folha de São Paulo ou Estadão por sua maior regularidade nas publicações.

5.2 Análise de Repercussão

Nesta seção são apresentadas análises de repercussão de notícias dos jornais G1, Folha de São Paulo e Estadão referente a quantidade de compartilhamentos nas redes sociais: Facebook, Twitter, LinkedIn e GooglePlus e a quantidade de manifestações de seus leitores via comentários na própria página da notícia.

5.2.1 Comentários

A seguir são apresentadas análises sobre a repercussão de notícias dos diferentes jornais em diferentes granularidades de tempo via comentários da página do jornal. O objetivo é conhecer o interesse dos leitores em se manifestarem escrevendo comentários.

[29]Para esta análise um *outlier* é um dia em que a quantidade de notícias excede em muito o valor médio esperado.

Ano

A Figura 5.11 apresenta o agregado da quantidade de comentários que as notícias dos jornais Folha de São Paulo, Estadão e G1 receberam ao longo dos anos. Não há registros de comentários coletados para o jornal G1. A coleta de notícias não apresentava acesso aos comentários dos leitores, o que não nos permite concluir que não haviam comentários ou se eles simplesmente não são retornados para notícias antigas. É possível verificar por meio do gráfico que os leitores do jornal Folha de São Paulo utilizam massivamente este recurso. O Apêndice C.3.1 apresenta evidência de que o comportamento apresentando pelos jornais não é devido a presença de nenhum *outlier*.

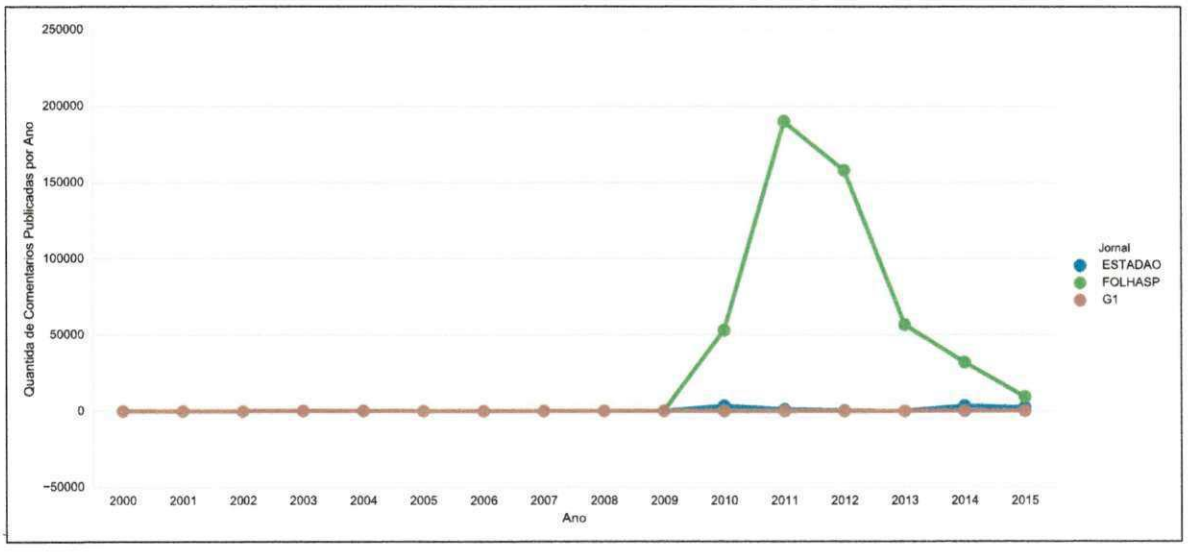

Figura 5.11: Quantidade de comentários recebidos por notícias publicadas ao longo dos anos.

Mês

Na análise de granularidade por mês, o objetivo foi verificar se havia algum mês em especial onde o número de comentários ganhasse algum destaque e, assim, tentar buscar respostas para o fato. A Figura 5.12 mostra a quantidade de comentários em notícias econômicas recebidas pelos jornais ao longo dos meses. É possível verificar duas características principais: a primeira delas é de que a quantidade de comentários é crescente de março a setembro, a segunda é que aparentemente setembro é o mês onde ocorre maior número de comentários. O Apêndice C.3.1 apresenta o comportamento de cada jornal separadamente.

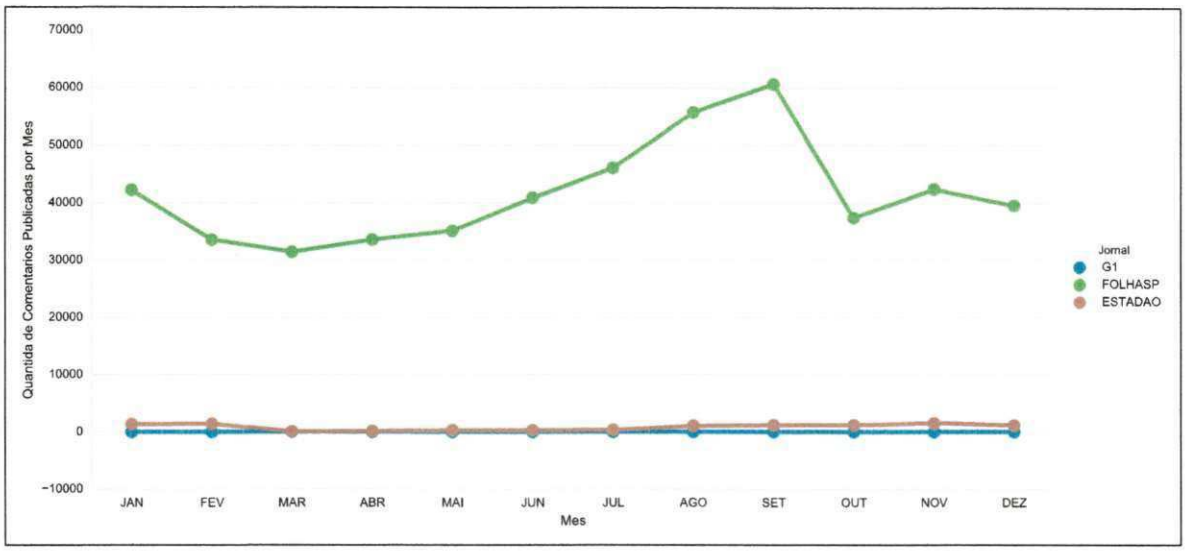

Figura 5.12: Quantidade de comentários recebidos por notícias publicadas ao longo dos meses.

Dia

A Figura 5.13 apresenta a quantidade de comentários publicados por dia do mês para todos os jornais. É possível verificar que há algumas variações bruscas ao longos dos dias. O dia 14 por exemplo, é compreensível ser o dia com maior número de publicações como já discutido. O Apêndice C.3.1 apresenta a variabilidade individual de cada jornal e mostra que os resultados apresentados não sofrem influência de nenhum *outlier*.

Dia da Semana

A Figura 5.14 apresenta o comportamento da publicação de comentários ao longo dos dias da semana. Há uma notória predileção pelos leitores pela quarta-feira. Isto é compreensível dado que nesse dia também foi verificado o maior número de publicações econômicas. O Apêndice C.3.1 apresenta os resultados individuais para cada jornal considerado.

5.2.2 Twitter

O Twitter[30] é uma rede social e microblog onde seus usuários possuem um limite de 140 caracteres para postar mensagens. Por vezes, notícias são compartilhadas entre seus seguidores

[30] www.twitter.com

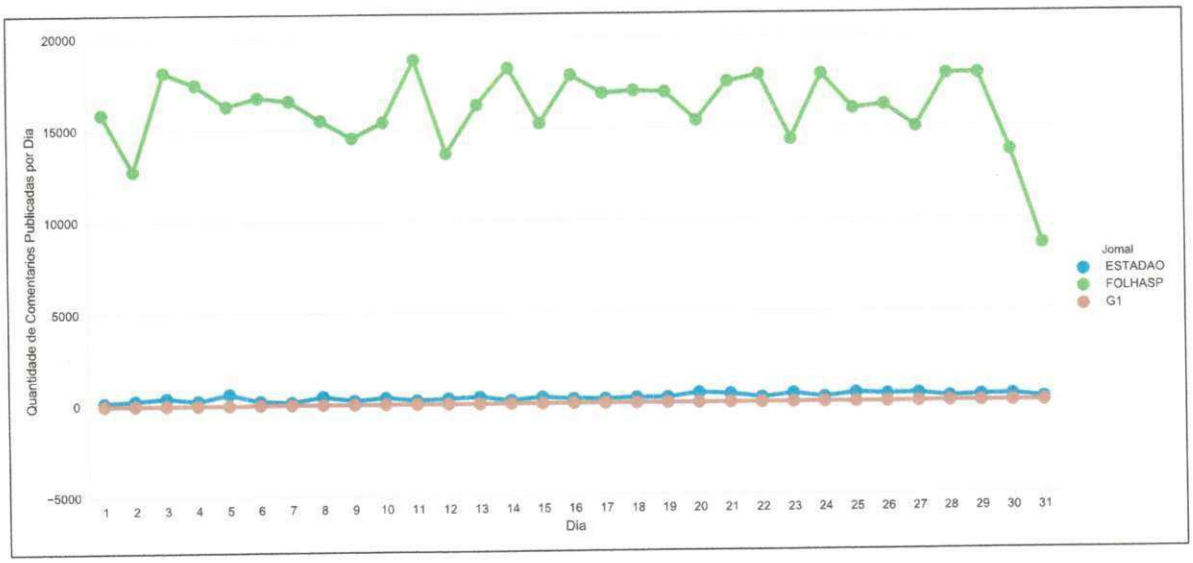

Figura 5.13: Quantidade de comentários de notícias econômicas por dia do mês.

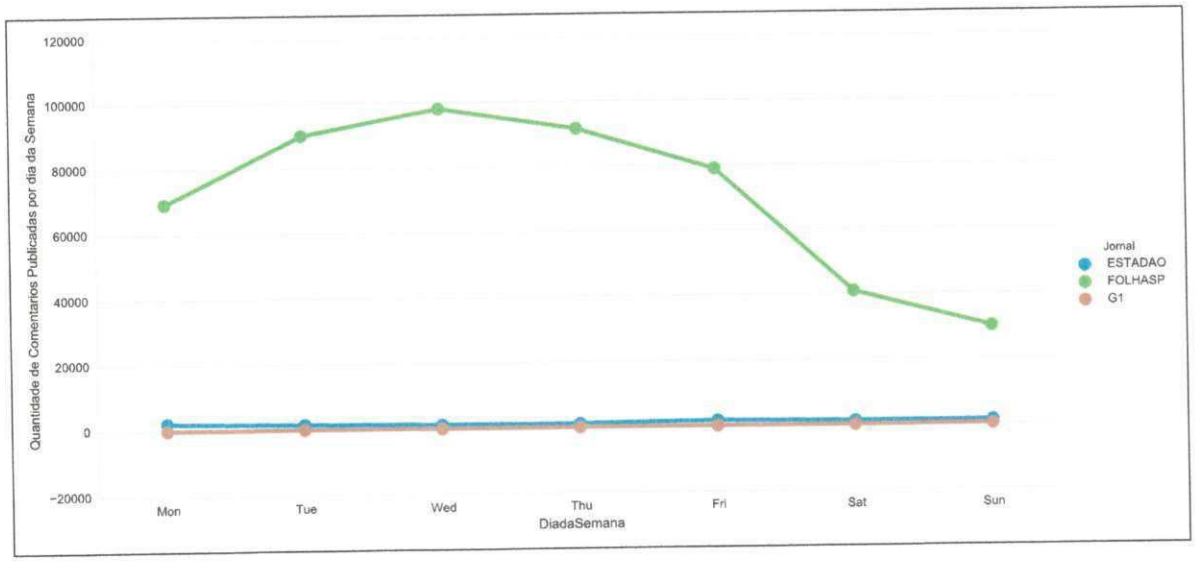

Figura 5.14: Quantidade de comentários de notícias econômicas por dia da semana.

e viralizam gerando assim grande quantidade de compartilhamentos. Nesta análise, busca-se compreender a utilização do Twitter para compartilhamento de notícias dos jornais analisados em diferentes granularidades de tempo. Nenhum dos cálculos incorporou *retweets*, ou seja, foram usados apenas os *tweets* diretamente postados.

Ano

A Figura 5.15 apresenta a quantidade de postagens de notícias econômicas via Twitter ano a ano para os jornais analisados. O G1 é o jornal que possui mais notícias repercutidas nesta mídia ultrapassando a dezena de milhão. Os jornais Folha de São Paulo e Estadão possuem entre si um comportamento semelhante ao longo dos anos. O Apêndice C.4 apresenta os resultados individuais para cada jornal considerado e expõe que algumas tendências são oriundas de *outliers*.

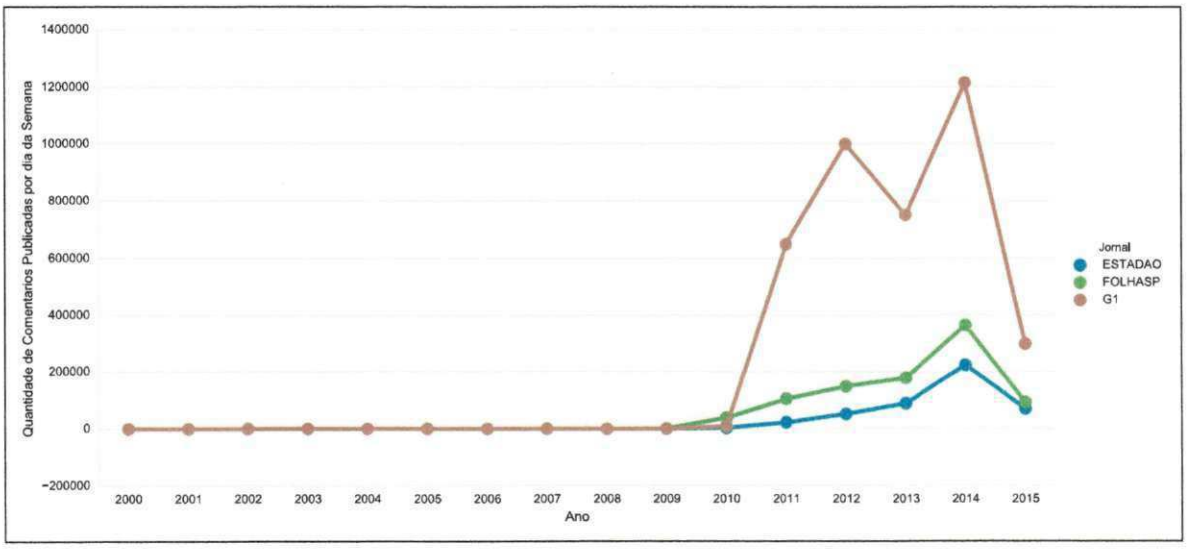

Figura 5.15: Quantidade de compartilhamentos de notícias econômicas via Twitter ano a ano.

Mês

A Figura 5.16 apresenta a quantidade de *tweets* que as notícias econômicas dos jornais analisados receberam ao longo dos meses. Percebe-se que há um declive da participação dessa mídia no primeiro semestre do ano, para todos os jornais de forma mais acentuada para G1 e

Estadão e mais suave para a Folha de São Paulo. O Apêndice C.4.1 apresenta os resultados individuais para cada jornal considerados.

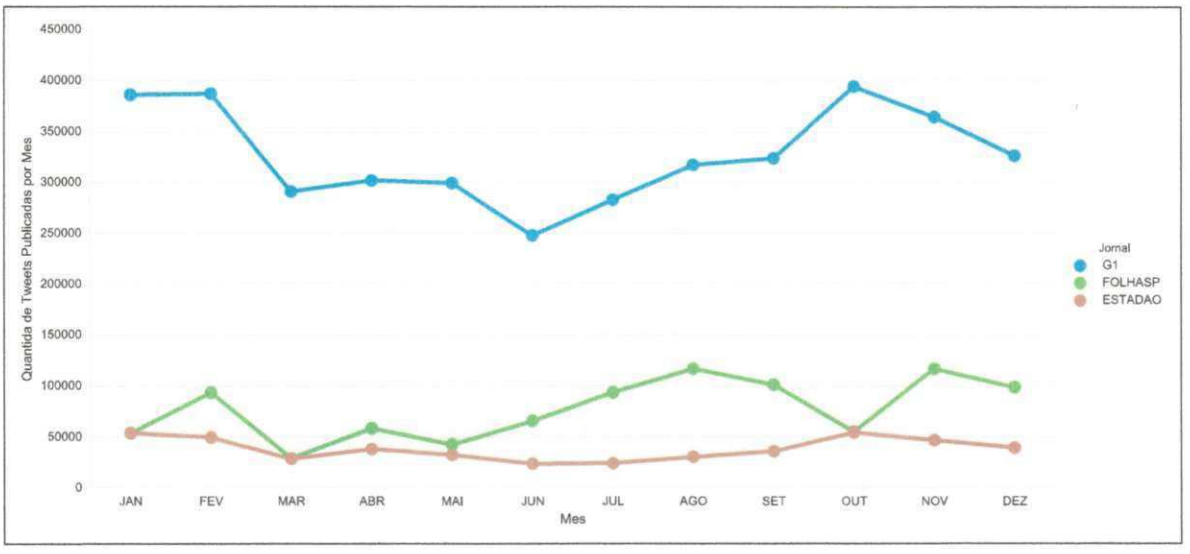

Figura 5.16: Quantidade de compartilhamentos de notícias econômicas dos jornais analisados via Twitter mês a mês.

Dia

A Figura 5.17 apresenta o número de *tweets* que notícias econômicas receberam agrupadas por dia do mês para os jornais analisados. Em média, a terceira semana do mês apresenta um sutil aumento em relação aos outros dias do mês.

Dia da Semana

A Figura 5.18 apresenta o número de *tweets* que notícias econômicas receberam agrupadas por dia da semana. Há uma regularidade no compartilhamento de notícias econômicas equitativo de Segunda à Sexta e uma diminuição nos finais de semana. O Apêndice C.4.2 apresenta os resultados individuais para cada jornal.

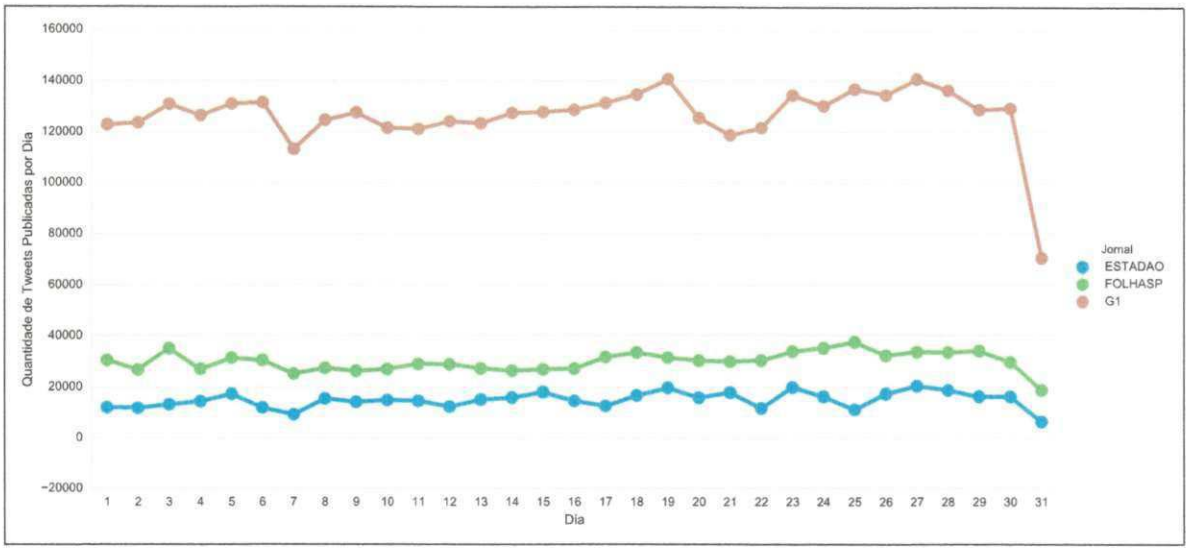

Figura 5.17: Quantidade de compartilhamentos de notícias econômicas dos jornais analisados durante os dias do mês.

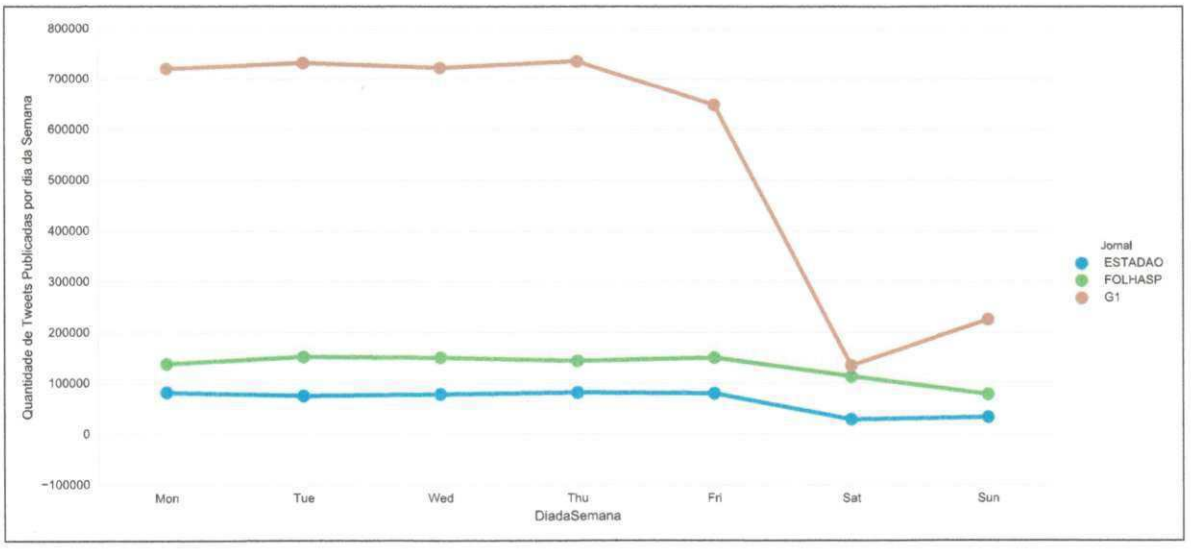

Figura 5.18: Quantidade de compartilhamentos de notícias econômicas dos jornais analisados durante os dias da semana.

5.2.3 Facebook

O Facebook[31] é uma rede social lançada em 2004 com a finalidade conectar pessoas entre si para então compartilharem informações de interesse mútuo.

Esta seção analisa o número de vezes que notícias econômicas publicadas nos jornais analisados foram compartilhadas via Facebook. É necessário reiterar que não foi possível contabilizar nesta análise, comentários que as notícias receberam quanto compartilhadas, número de re-compartilhamentos, curtidas ou outras manifestações de engajamento.

Ano

A Figura 5.19 apresenta o número de compartilhamentos das notícias econômicas dos jornais analisados ao longo dos anos. Percebe-se que há uma tendência crescente bastante íngrime da utilização do Facebook na repercussão desse tipo de notícia com um crescimento médio de ~3.5 vezes ao ano. O Apêndice C.5.1 apresenta os resultados individuais para cada jornal.

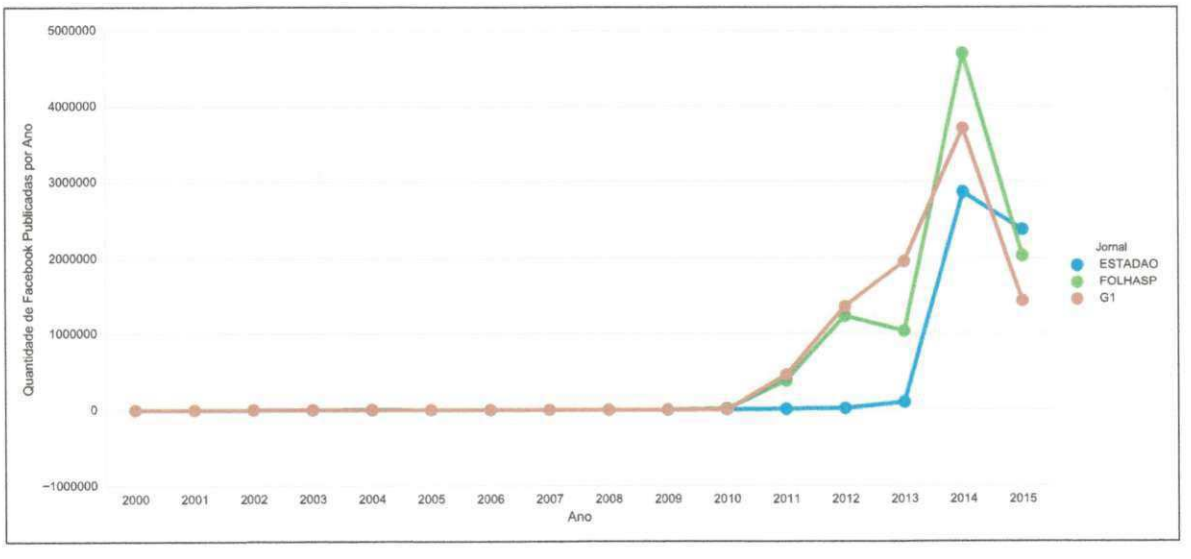

Figura 5.19: Quantidade de publicações de notícias econômicas dos jornais analisados via Facebook ano a ano.

[31] www.facebook.com

Mês

A Figura 5.20 apresenta a quantidade de compartilhamentos de notícias econômicas via Facebook ao longo dos meses do ano. Há dois fatos que merecem atenção nesta análise: o primeiro deles é o crescimento marcado de junho à outubro e pico em Fevereiro. O Apêndice C.5.2 apresenta os resultados individuais para cada jornal.

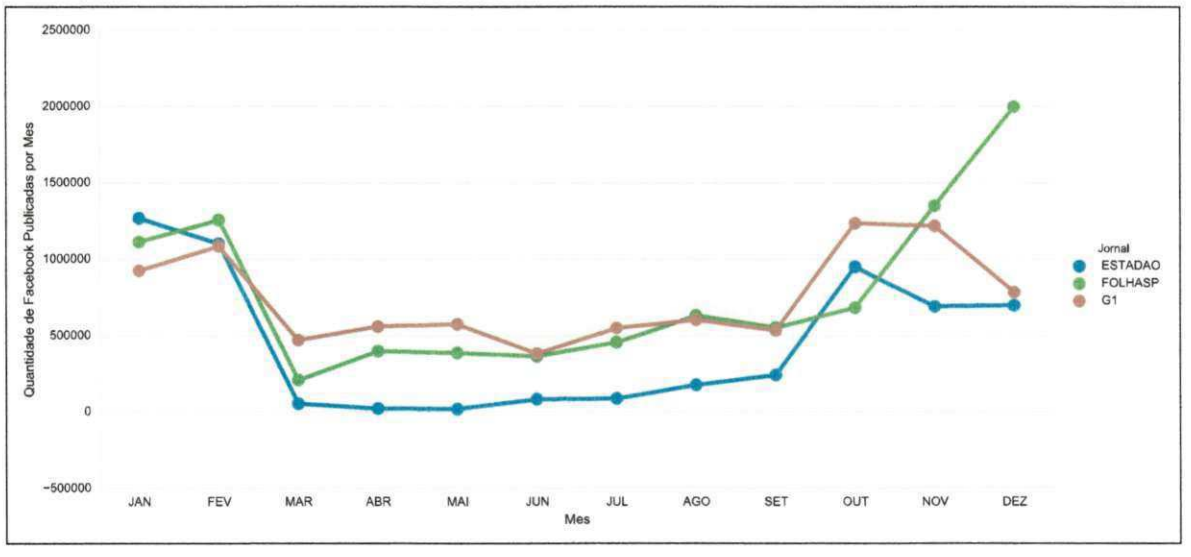

Figura 5.20: Quantidade de publicações de notícias econômicas dos jornais analisados via Facebook mês a mês.

Dia

A Figura 5.21 mostra que a quantidade de compartilhamento de notícias econômicas via Facebook apresenta bastante irregularidade ao longo dos dias do mês, diferente do comportamento das outras mídias vistas até agora para esse tipo de análise. As Figuras 5.22, 5.23 e 5.24 apresentam a variabilidade do número de compartilhamento de notícias ao longo dos dias do mês para os jornais Estadão, Folha de São Paulo e G1 respectivamente. Percebe-se que apesar do índice de tendência central dos dias estarem na base das observações alguns dias apresentam diferenças consideráveis em relação a outros dias como é o caso dos dias 14 e 21 para o jornal G1. Por fim, também é possível evidenciar que o Facebook apresenta a maior variabilidade do número de compartilhamentos entre todas as mídias, saindo de dias (1,2) com alguns milhares de compartilhamentos a dias com alguns milhões (13,29).

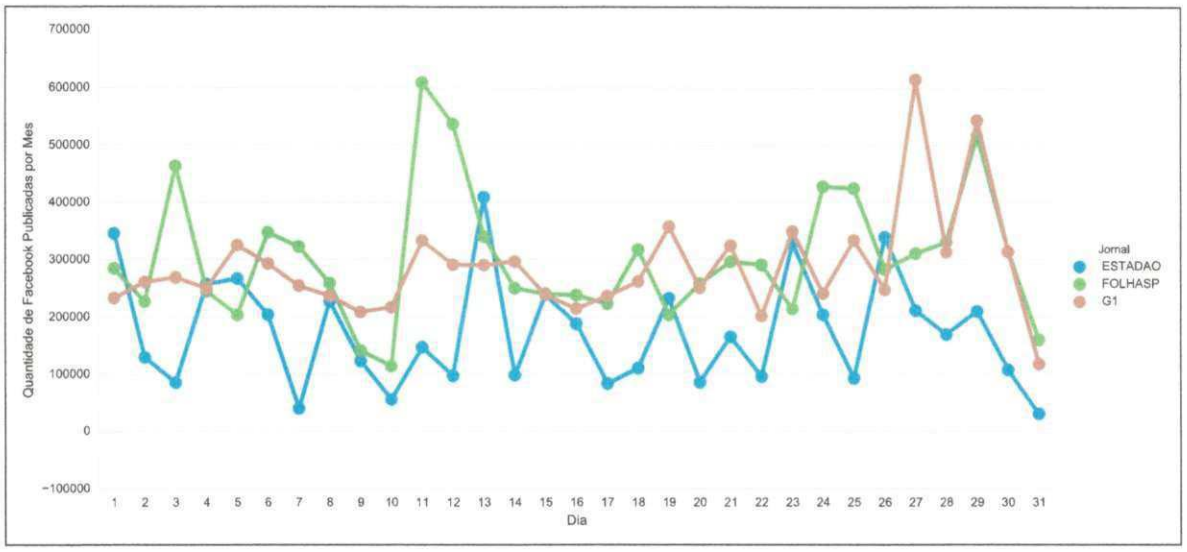

Figura 5.21: Quantidade de publicações de notícias econômicas dos jornais analisados via Facebook ao longo dos dias do mês.

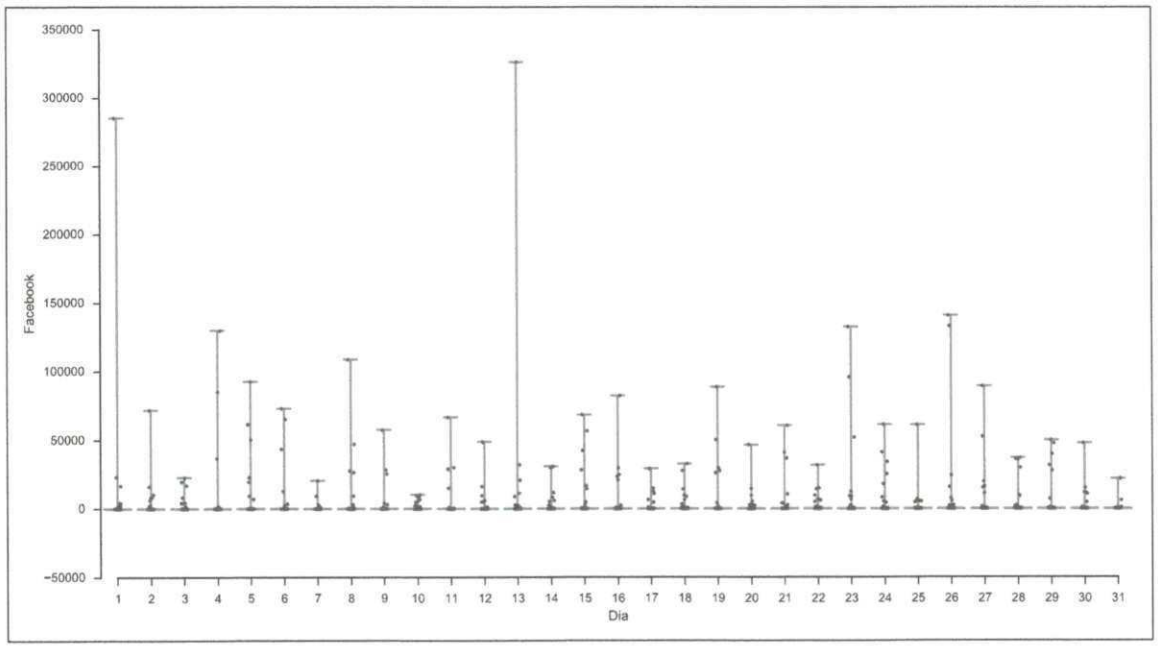

Figura 5.22: Quantidade de publicações de notícias econômicas do jornal Estadão via Facebook ao longo dos dias do mês.

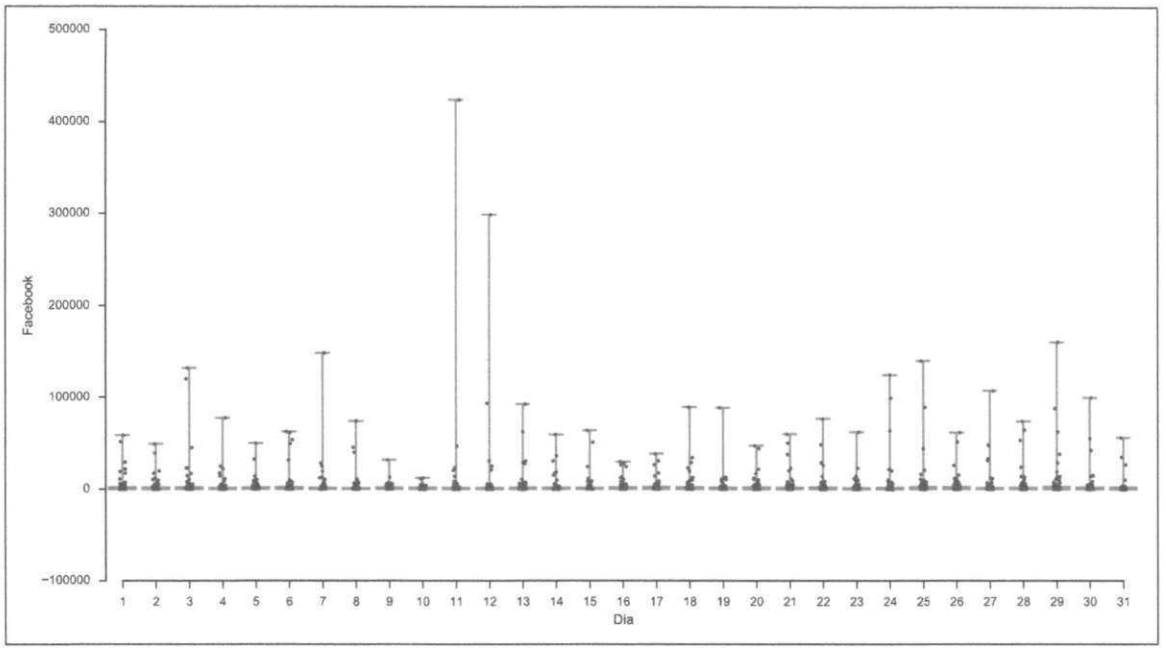

Figura 5.23: Quantidade de publicações de notícias econômicas do jornal Folha de São Paulo via Facebook ao longo dos dias do mês.

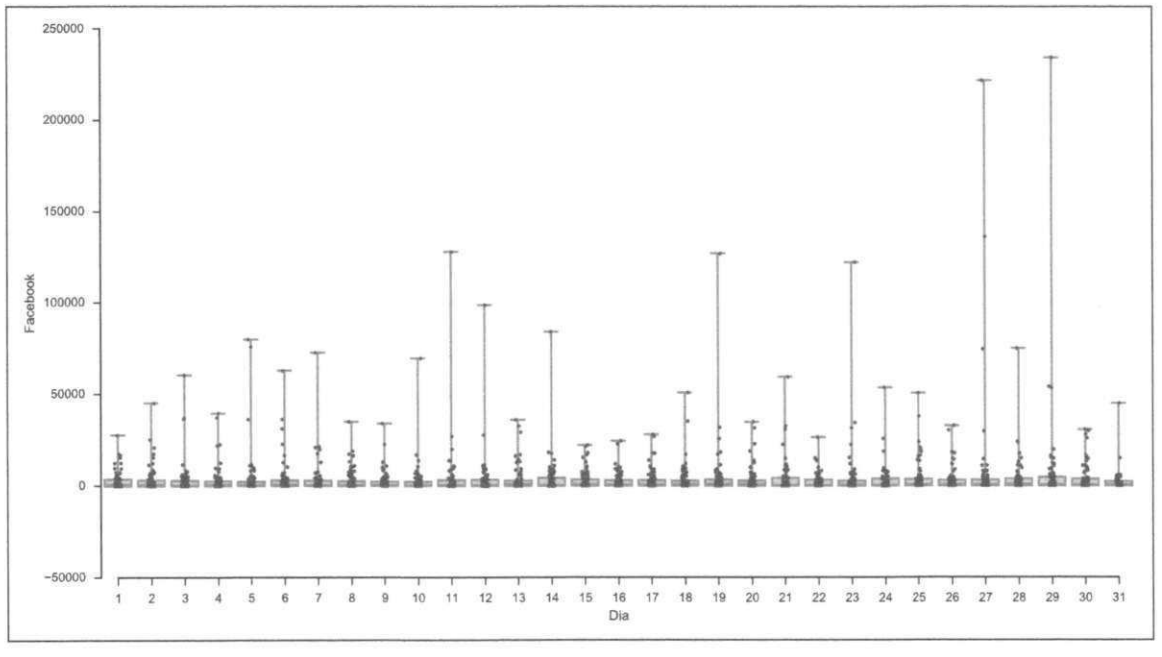

Figura 5.24: Quantidade de publicações de notícias econômicas do jornal G1 via Facebook ao longo dos dias do mês.

Dia da Semana

A Figura 5.25 apresenta o número de compartilhamentos de notícias econômicas via Facebook ao longo dos dia da semana. Percebe-se que a partir da quinta-feira há uma diminuição dos compartilhamentos para todos os jornais analisados. O Apêndice C.5.3 apresenta os resultados individuais para cada jornal.

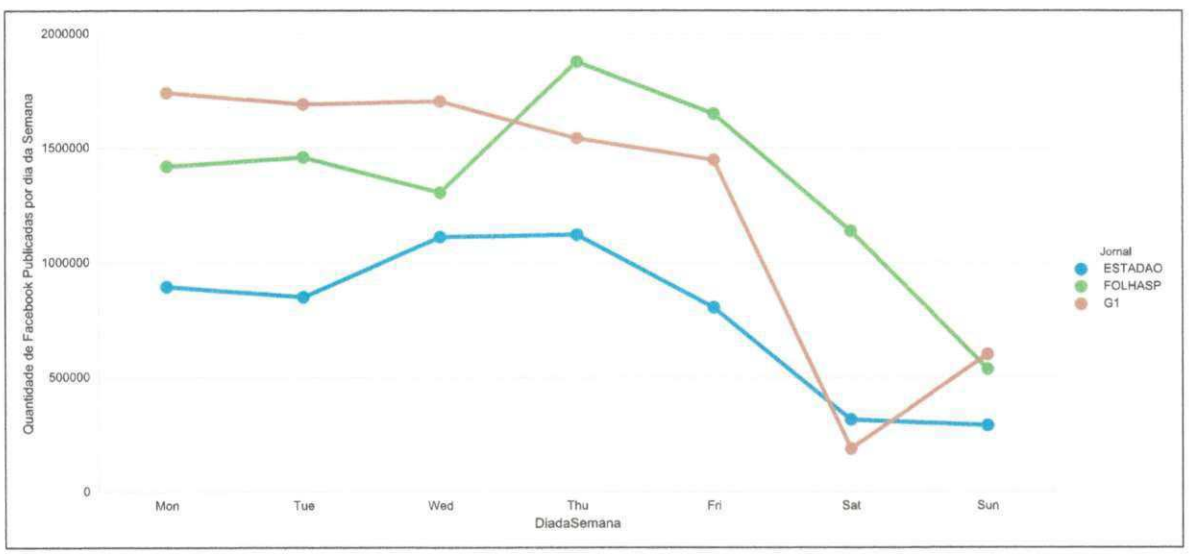

Figura 5.25: Quantidade de publicações de notícias econômicas dos jornais analisados via Facebook ao longo dos dias do mês.

5.2.4 LinkedIn

O LinkedIn[32] é uma rede social lançada em 2003 para a interação entre profissionais e empresas. Inicialmente a hipótese era de que o compartilhamento de notícias econômicas no LinkedIn era irrelevante dado que o objetivo desta rede social visa o compartilhamento de informações pertinentes apenas ao entorno profissional do usuário, ao invés de todo seu espectro de interesse como no caso do Facebook.

Após a coleta percebeu-se o inverso, que em grande medida o LinkedIn apresenta milhares de compartilhamentos superando outras redes sociais e os próprios comentários dos leitores da página.

[32] www.linkedin.com

Ano

A Figura 5.26 apresenta o número de notícias econômicas compartilhadas via LinkedIn ano a ano. Percebe-se que com exceção do jornal G1 os demais jornais demonstram tendência de crescimento. A Folha de São Paulo, por exemplo, com apenas 2 meses coletados de 2015 superou todo o ano de 2014. O Apêndice C.6.1 apresenta os resultados individuais para cada jornal.

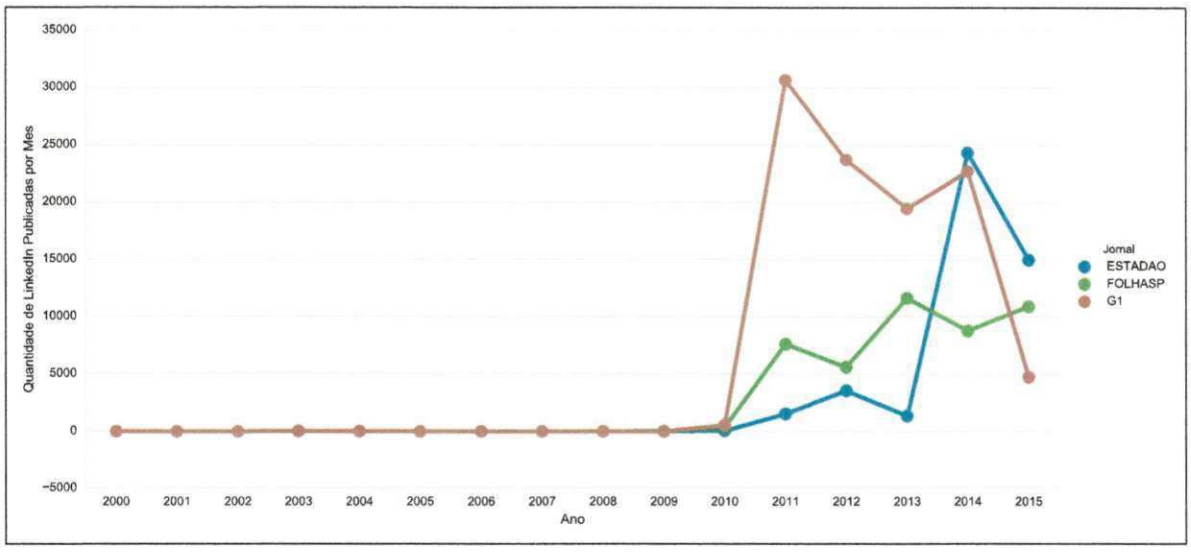

Figura 5.26: Quantidade de publicações de notícias econômicas dos jornais analisados via LinkedIn ano a ano.

Mês

A Figura 5.27 apresenta o número de compartilhamentos de notícias econômicas via LinkedIn para todos os jornais analisados mês a mês. Também para o LinkedIn percebe-se sutil tendência crescente no segundo semestre. O mês de Maio apresenta a maior variabilidade de repercussões via LinkedIn. O Apêndice C.6.2 apresenta os resultados individuais para cada jornal.

Dia

A Figura 5.28 apresenta o número de notícias econômicas compartilhadas via LinkedIn por dia do mês. Percebe-se que a aparente variação ao longo dos dias é explicada pela menor

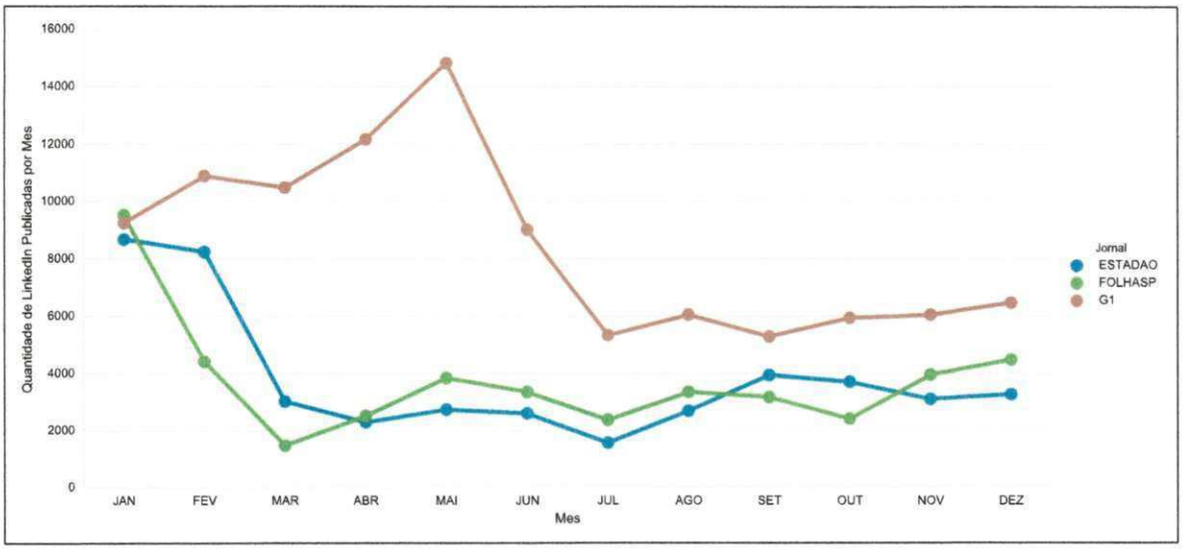

Figura 5.27: Quantidade de publicações de notícias econômicas dos jornais analisados via LinkedIn mês a mês.

amplitude da escala e de alguns eventuais dias com muitas notícias, representando índices de compartilhamento bastante regular. O Apêndice C.6.3 apresenta os resultados individuais para cada um dos jornais analisados.

Dia da Semana

A Figura 5.29 apresenta o número de notícias econômicas dos jornais analisados compartilhadas via LinkedIn durante os dias da semana. Pela primeira vez é possível perceber que para a maioria dos jornais o número de repercussão de notícias durante o final de semana é equivalente ou superior aos dias durante a semana. O Apêndice C.6.4 apresenta os resultados individuais para cada um dos jornais analisados.

5.2.5 Google Plus

O Google Plus[33] é uma rede social e de serviços lançada em 2011 e mantida pela Google. Segundo a declaração do CEO do Google no último Social Media Week[34] o Google Plus tem por objetivo ser um elo comum de engajamento entre usuários e não um competidor direto ao Facebook. Nessa perspectiva foi analisado também o número de notícias econômi-

[33] https://plus.google.com/

[34] http://socialmediaweek.org/saopaulo/

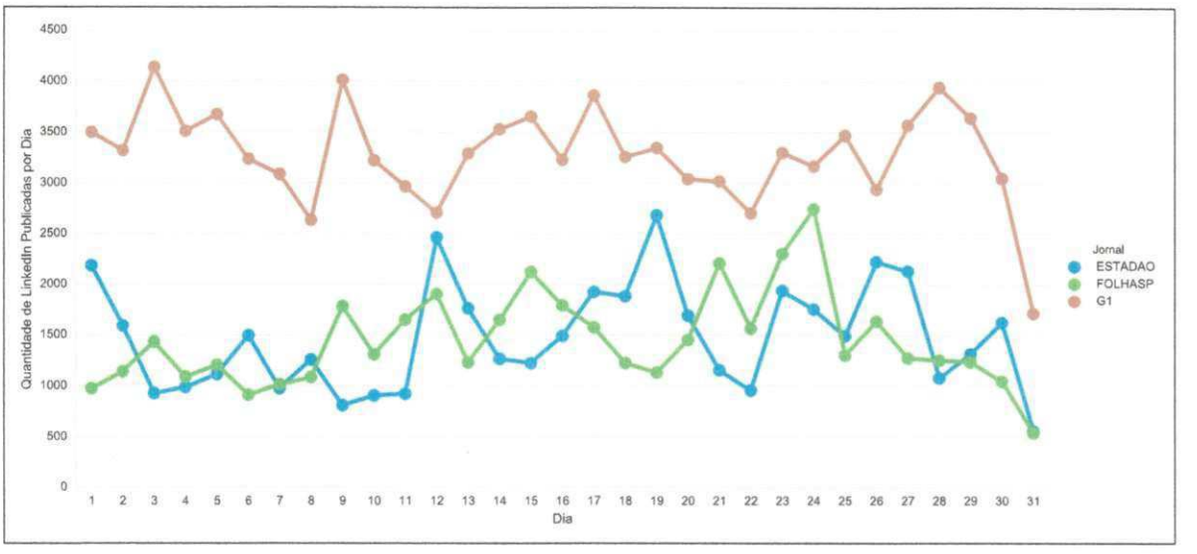

Figura 5.28: Quantidade de publicações de notícias econômicas dos jornais analisados via LinkedIn ao longo dos dias do mês.

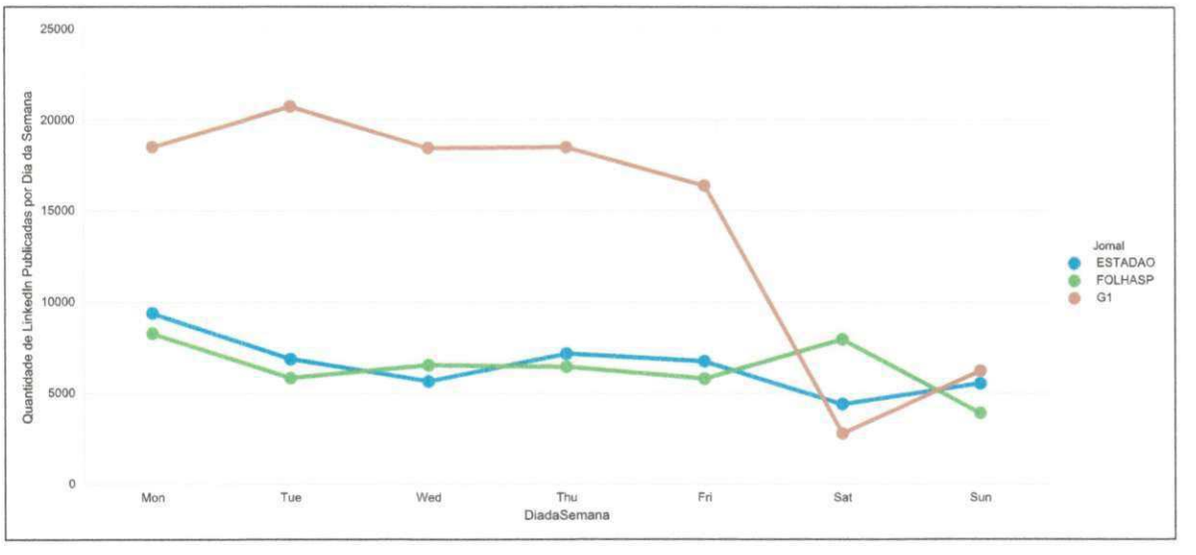

Figura 5.29: Quantidade de publicações de notícias econômicas dos jornais analisados via LinkedIn ao longo dos dias do mês.

cas publicadas pelos jornais Estadão, Folha de São Paulo e G1 via Google Plus em diversas granularidades de tempo. Por surpresa, apenas o jornal Estadão possui um número de compartilhamento de suas notícias econômicas relevante para a análise, provavelmente devido a exposição em cada notícia do ícone do Google Plus, diferente dos outros jornais.

Ano

A Figura 5.30 apresenta o número de compartilhamentos de notícias dos jornais analisados via Google Plus ano a ano. É notória a superioridade do jornal Estadão nesta mídia e aparente inatividade dos outros jornais em relação a esta mídia.

É interessante que o Google Plus, apesar de apresentar menor aceitação em relação a outras redes sociais, é mais utilizado pelos leitores do Estadão para compartilhar notícias econômicas que o Twitter, LinkedIn e comentários da própria página dos jornais. A Figura 5.31 apresenta a variabilidade das notícias econômicas publicadas pelo jornal Estadão e compartilhadas via Google Plus ano a ano. É possível perceber pelo aumento gradual da mediana que há uma crescente utilização dessa mídia ao longo dos anos pelos leitores do Estadão, diferente das outras mídias.

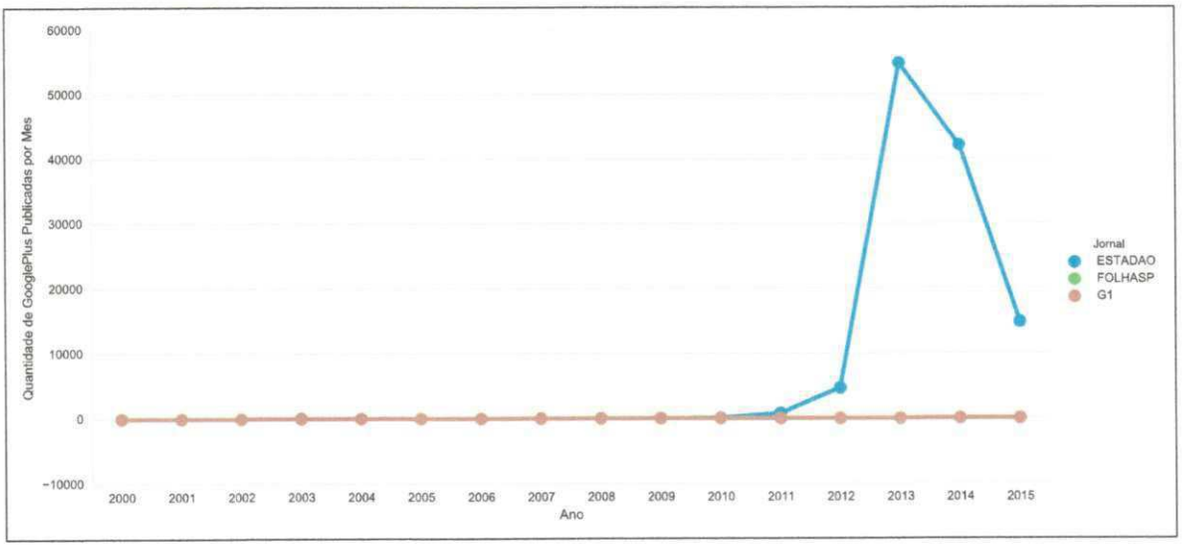

Figura 5.30: Quantidade de publicações de notícias econômicas dos jornais analisados via Google Plus ano a ano.

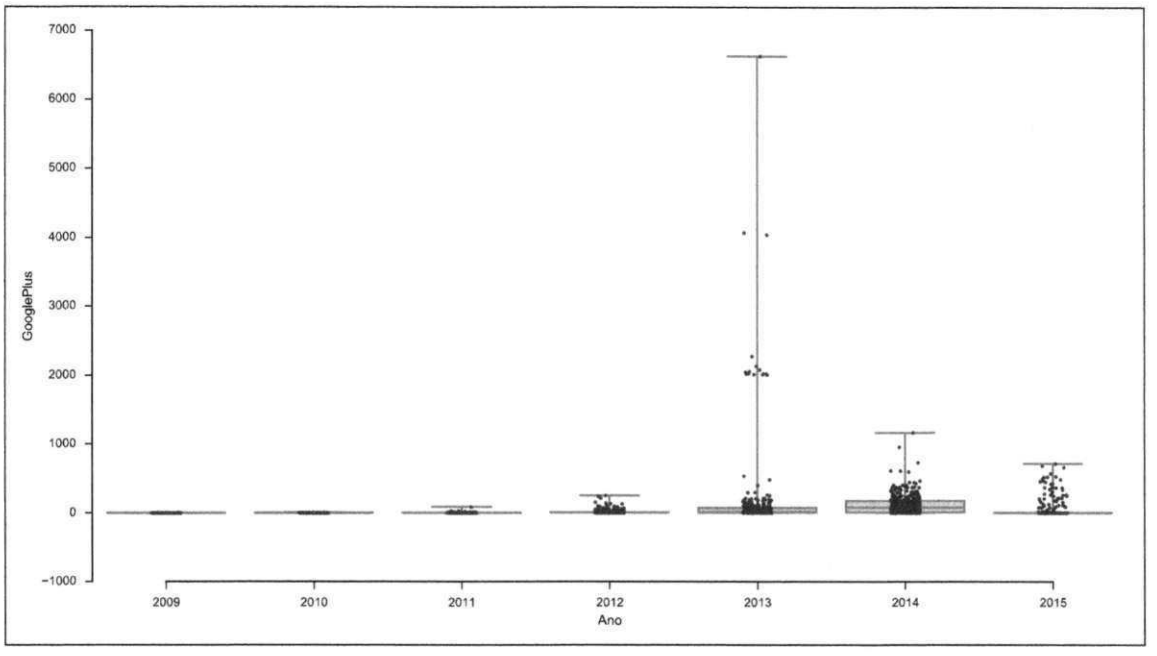

Figura 5.31: Quantidade de publicações de notícias econômicas do jornal Estadão via Google Plus ano a ano.

Mês

A Figura 5.32 apresenta exatamente as mesmas características já evidenciadas por outras mídias, ou seja, o mês de Maio acentuado, e uma tendência crescente de compartilhamentos no segundo semestre do ano. A Figura 5.33 mostra a variabilidade do número de notícias econômicas publicadas pelo jornal Estadão via Google Plus em cada mês. O mês de Maio, apesar de apresentar um pico bastante íngreme, é explicado por pontuais valores acima da média mostrando que o pico foi causado por eventos pontuais. Por sua vez, as inúmeras publicações de Fevereiro e o crescimento gradual de junho à dezembro é de fato verificado pois não se atribui a *outliers* a explicação por tal crescimento.

Dia

A Figura 5.34 apresenta o número de notícias econômicas publicadas por todos os jornais analisados que foram compartilhadas via Google Plus durante os dias do mês. Da mesma forma, a Figura 5.35 apresenta a variabilidade para o número de publicações. O número de publicações por dia do mês apresenta-se bastante regular. Os picos e valores são totalmente explicados por valores pontuais esporádicos.

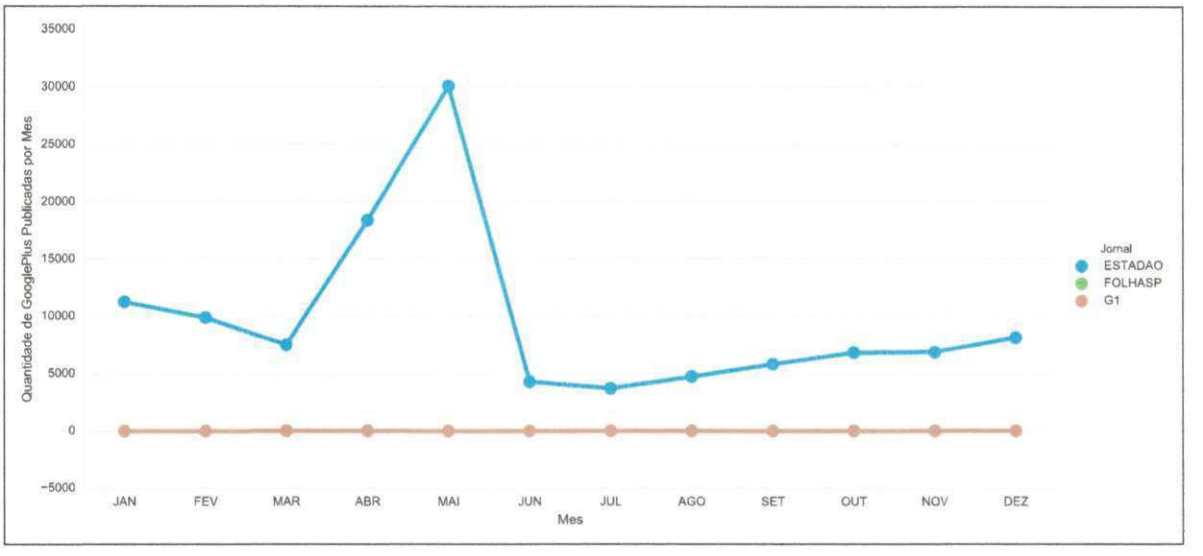

Figura 5.32: Quantidade de publicações de notícias econômicas dos jornais analisados via Google Plus mês a mês.

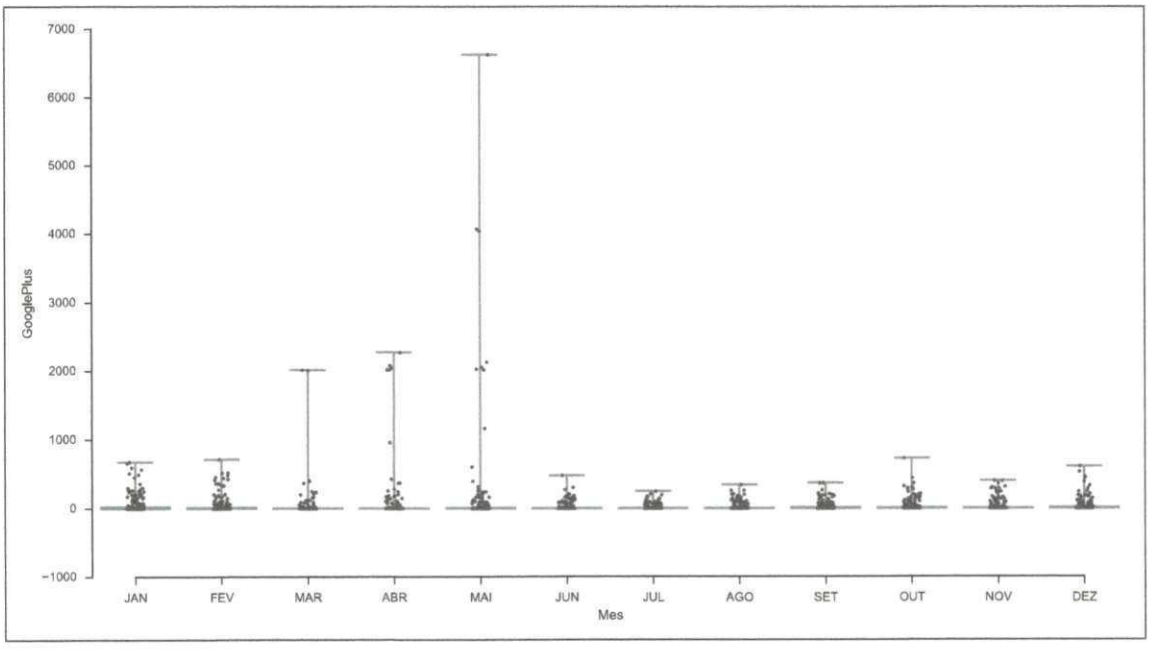

Figura 5.33: Quantidade de publicações de notícias econômicas do jornal Estadão via Google Plus mês a mês.

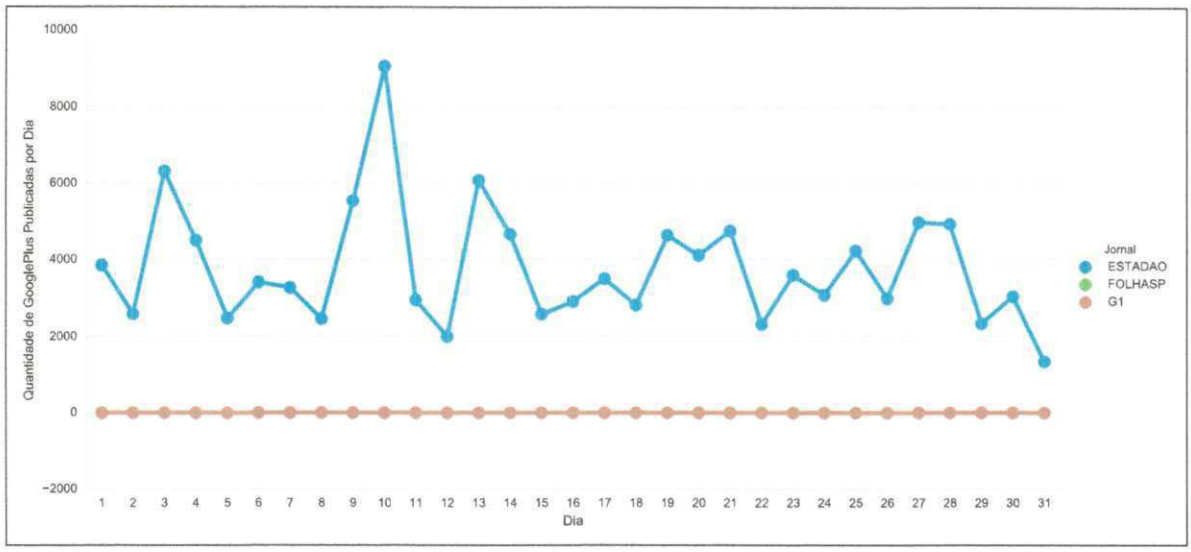

Figura 5.34: Quantidade de publicações de notícias econômicas dos jornais analisados via Google Plus ao longo dos dias do mês.

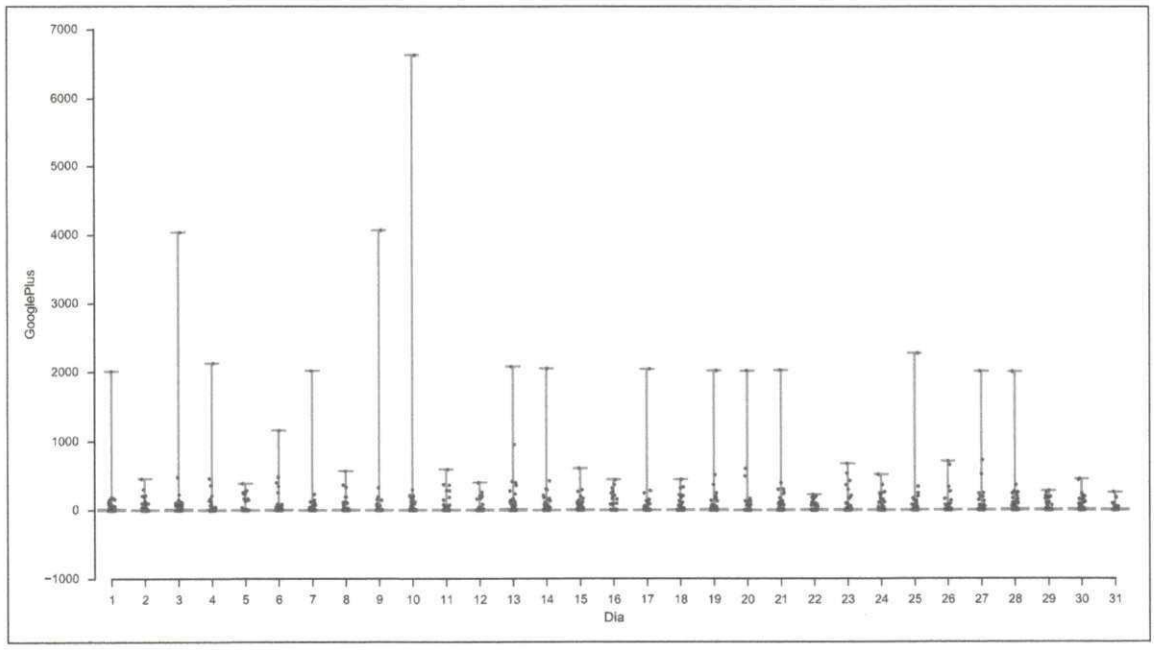

Figura 5.35: Quantidade de publicações de notícias econômicas do jornal Estadão via Google Plus ao longo dos dias do mês.

Dia da Semana

A Figura 5.36 apresenta o número de notícias compartilhadas via Google Plus pelos jornais analisados ano a ano. A aparente preferência pela sexta-feira é vista com o auxílio do gráfico de variabilidade na Figura 5.37 como fruto de eventos esporádicos. Por fim, com o auxílio das duas imagens é possível concluir que há pouca atividade durante os finais de semana assim como verificado pelas outras mídias.

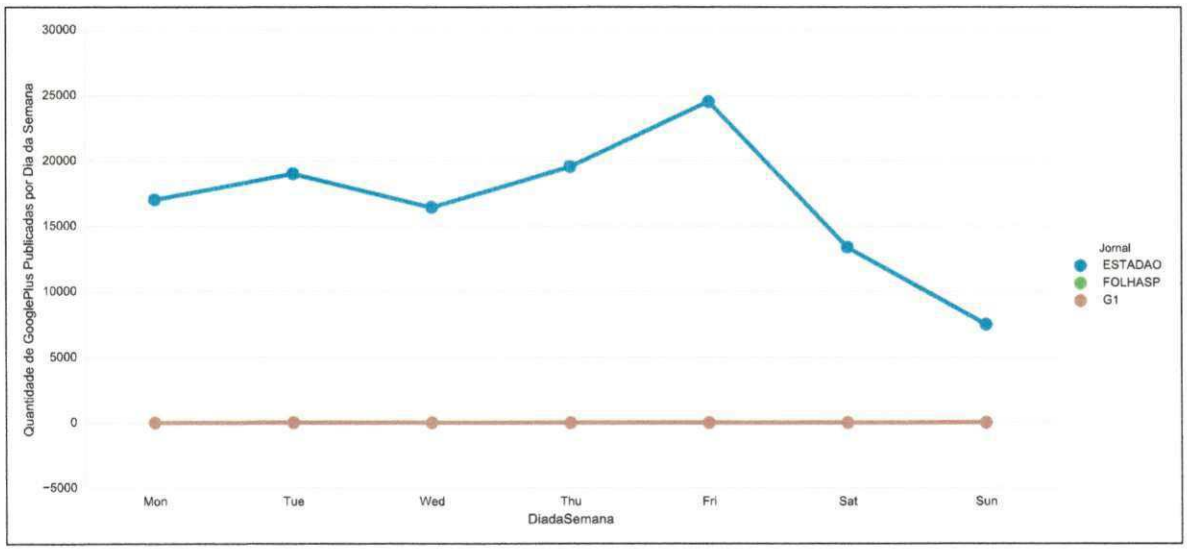

Figura 5.36: Quantidade de publicações de notícias econômicas dos jornais analisados via Google Plus ao longo dos dias da semana.

5.3 Repercussão vs Quantidade de Notícias por Jornal

Nesta seção é discutida a relação entre o número de notícias econômicas publicadas e a repercussão que essas notícias ganharam ao longo dos anos.

A Figura 5.38 apresenta a quantidade de repercussão total que as notícias publicadas por cada jornal receberam ao longo dos anos analisados, ou seja, o agregado das repercussões obtidas em todas as mídias ao longo dos anos. Há uma tendência crescente da repercussão de notícias econômicas publicadas em todos os jornais. E, apesar da coleta contemplar apenas 2 meses de 2015[35], já percebe-se um volume de repercussão superior a muitos anos anteriores.

[35] A atividade de coleta foi finalizada 3 de Março de 2015.

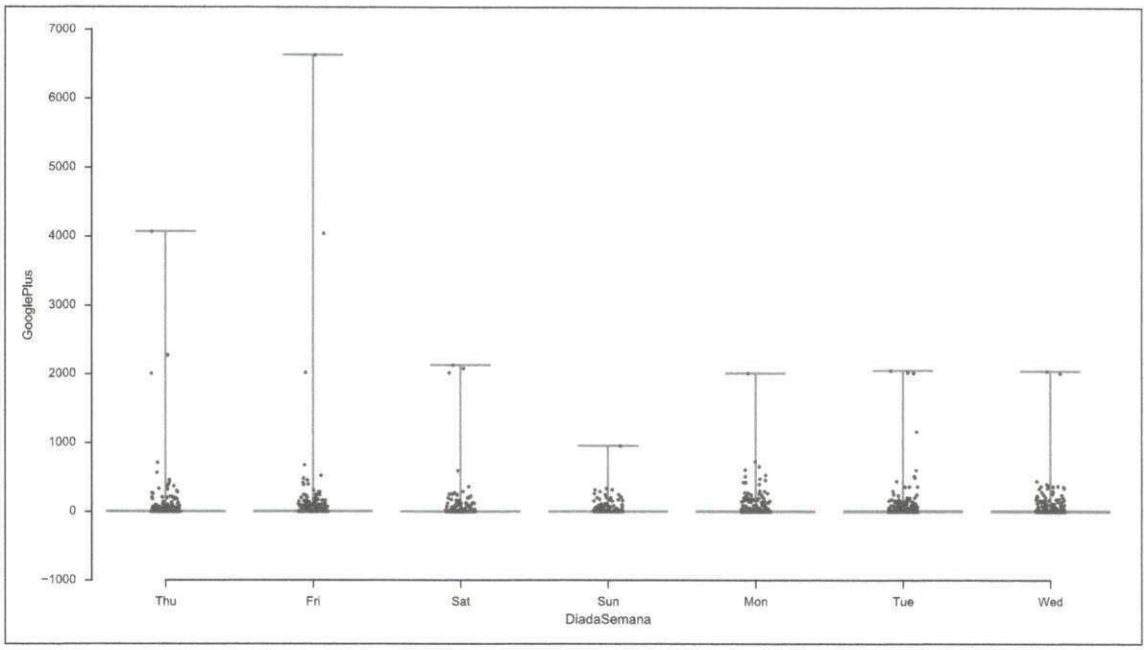

Figura 5.37: Quantidade de publicações de notícias econômicas do jornal Estadão via Google Plus ao longo dos dias da semana.

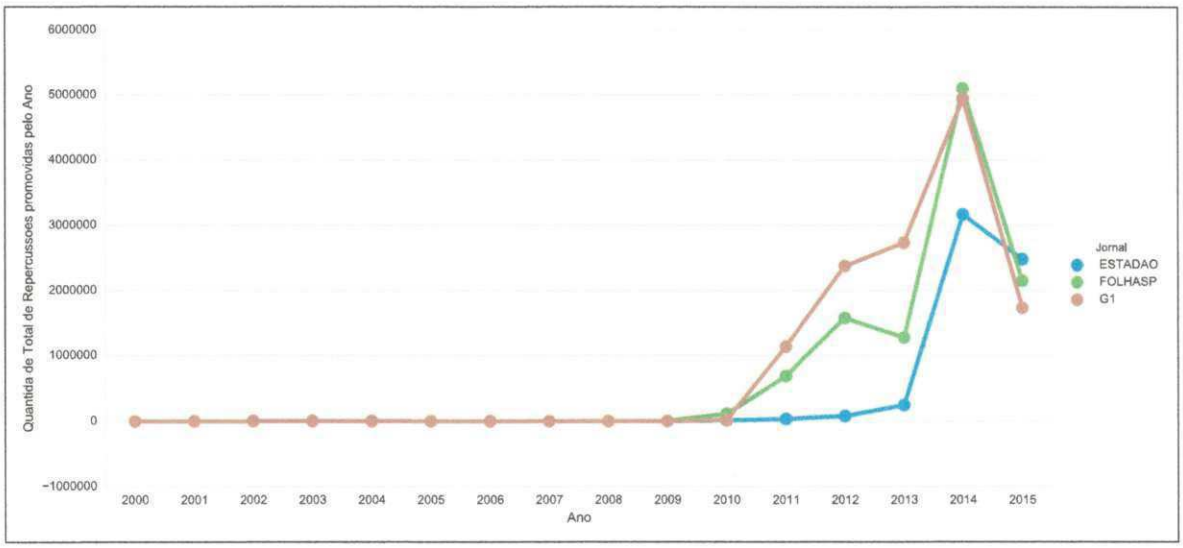

Figura 5.38: Quantidade total de compartilhamento de notícias econômicas dos jornais analisados ano a ano.

5.3.1 G1

A Figura 5.39 apresenta o número de notícias econômicas publicadas pelo jornal G1 ao longo dos anos versus a quantidade de compartilhamentos que as notícias do jornal recebeu nas redes sociais ao longo dos anos. E percebe-se:

- Entre 2010 e 2013 é possível perceber que para um aumento exponencial na quantidade de notícias há um aumento linear para o número repercussões.
- A partir de 2012 percebe-se uma diminuição do número de notícias econômicas publicadas.
- A diminuição da quantidade de notícias publicadas é seguida por uma explosão de repercussão para cada notícia. Especula-se que, com o passar do tempo, os jornais passaram a compreender que tipo de notícias geram altas repercussões e adotaram estratégias quanto ao estilo e a forma das notícias para que elas fossem produzidas de forma a serem propagadas. Associada a isso também está a difusão do acesso a Internet que faz com que a notícia online ganhe gradativamente mais leitores.

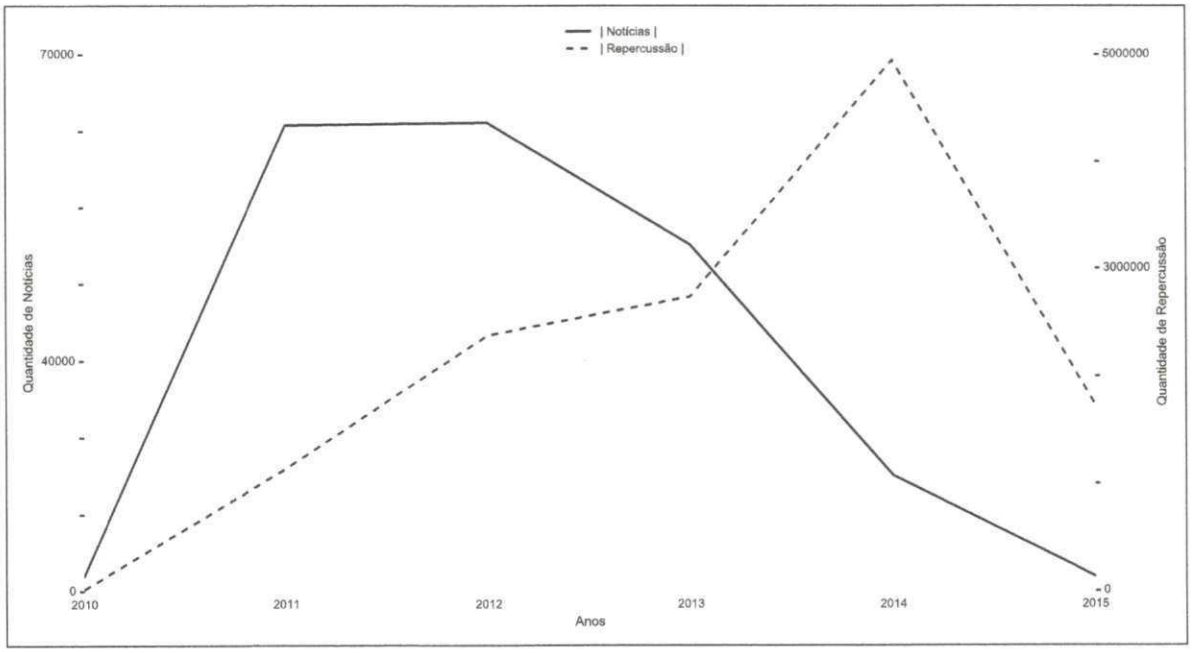

Figura 5.39: Notícias econômicas publicadas pelo jornal G1 ao longo dos anos versus número de compartilhamentos recebidos.

5.3.2 Folha

A Figura 5.40 apresenta o número de notícias econômicas publicadas pelo jornal Folha de São Paulo ao longo dos anos versus a quantidade de compartilhamentos que as notícias do jornal receberam nas redes sociais longo dos anos. Por meio desta Figura, percebe-se que entre 2009 e 2014 a quantidade de publicações tendem a refletir uma correlação inversa com a quantidade de repercussões. Ou seja, sempre que uma delas cresce a outra decresce proporcionalmente.

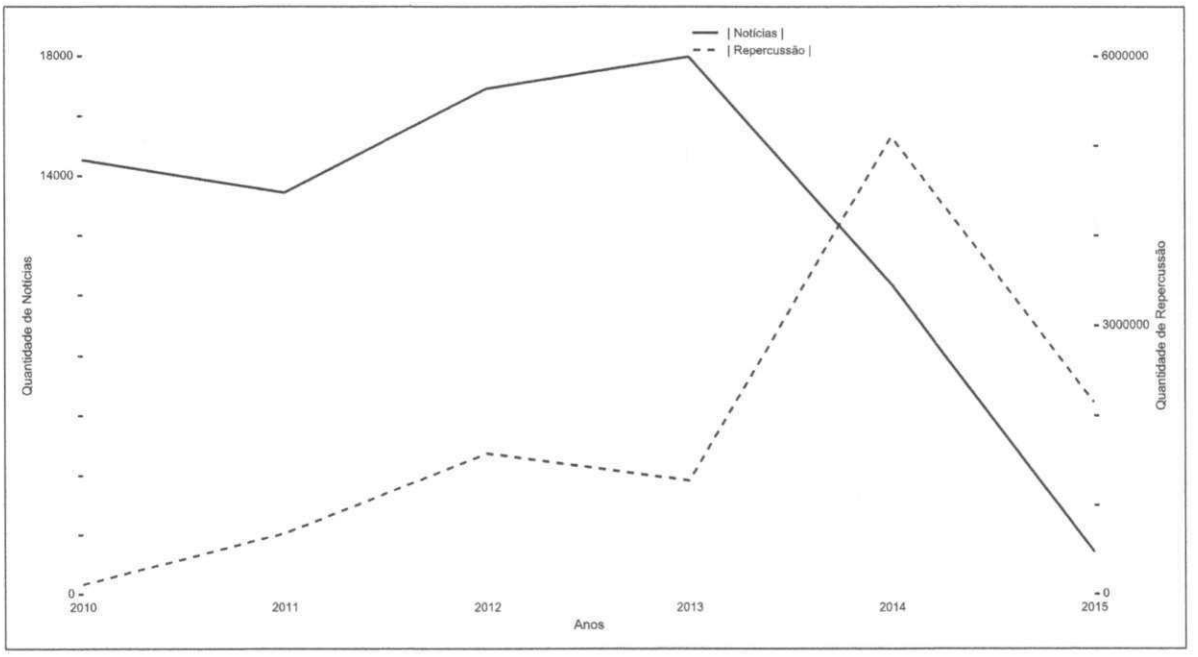

Figura 5.40: Notícias econômicas publicadas pelo jornal Folha de São Paulo ao longo dos anos versus número de compartilhamentos recebidos.

5.3.3 Estadão

A Figura 5.41 apresenta o número de notícias econômicas publicadas pelo jornal Estadão ao longo dos anos versus a quantidade de compartilhamentos que as notícias do jornal receberam nas redes sociais ao longo dos anos. E percebe-se que entre 2011 e 2014 a quantidade do número de notícias decresce enquanto a quantidade de repercussões aumenta.

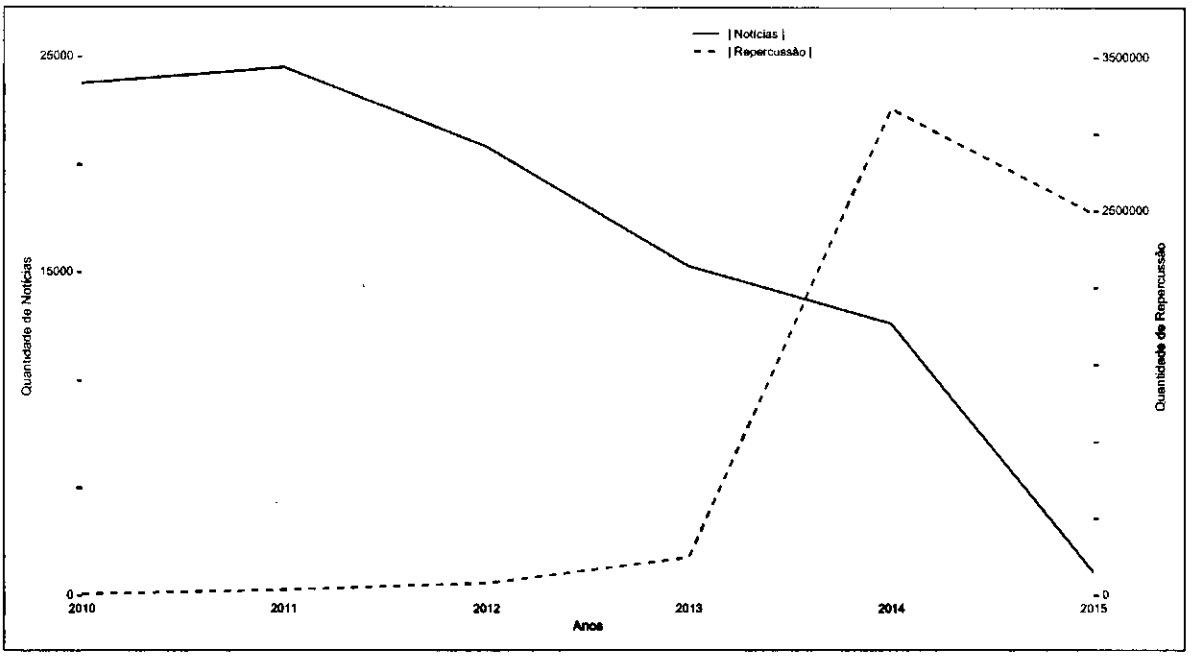

Figura 5.41: Notícias econômicas publicadas pelo Estadão ao longo dos anos pelo número de compartilhamentos recebidos.

5.4 Considerações Finais

Apresentam-se a seguir as conclusões para as análises realizadas neste capítulo:

1. O jornal G1 apresenta comportamento atípico em relação aos demais jornais analisados. O comportamento bimodal do número de publicações diárias evidencia que o jornal publica normalmente quantidade extremas de notícias (ou poucas ou muitas notícias). É provável que tal comportamento seja responsável pela ampla divulgação nas redes sociais.

2. A Folha de São Paulo é o jornal que, apesar de publicar menos que o G1, possui mais repercussões em todas as redes sociais. Em geral o leitor do jornal Folha de São Paulo é bastante eclético quanto a forma de compartilhar as notícias econômicas do jornal.

3. Os leitores do Estadão possuem preferência por Facebook e Google Plus, sendo tímidas suas participações nas outras mídias.

4. Agosto é o mês com o maior número publicações econômicas provavelmente porque é o mês em que o Palácio do Planalto apresenta o orçamento para o próximo ano e as pre-

visões de crescimento do PIB, inflação, selic e demais taxas. Apesar disso, fevereiro é o mês com a maior mediana para compartilhamentos e comentários de notícias econômicas evidenciado em todas as análises. Provavelmente por ser o mês onde as pessoas discutem o planejamento financeiro ao longo do ano, leem e compartilham notícias econômicas. Maio é o mês com maior número de repercussão de notícias econômicas, provavelmente o primeiro reflexo após o 1° trimestre da economia nacional e onde os leitores reavaliam suas premissas de fevereiro.

5. Os dias 13,14,27 e 28 são ligeiramente mais prováveis para haver compartilhamento de notícias econômicas em relação aos demais dias.

6. Apesar de supostamente o final de semana proporcionar mais tempo livre para os leitores poderem se dedicar as mídias sociais o maior número de repercussões ocorreram de Segunda à Sexta. Além disso, há vertiginosa queda do número de compartilhamento de notícias econômicas durante o final de semana – sábado e domingo – em relação aos outros dias da semana. Por fim, durante os dias da semana a quarta-feira apresentou-se como potencial dia para haver bastante notícias econômicas e compartilhamentos.

7. Há um comportamento crescente do número de repercussões de notícias econômicas durante o segundo semestre de todos os anos, evidenciado em todas as análises realizadas.

8. LinkedIn é a única rede social analisada que o compartilhamento de notícias econômicas durante o final de semana é equivalente ou superior aos dias da semana.

9. O Google Plus é a mídia digital mais utilizada pelos leitores do Estadão (após o Facebook) superando as redes sociais Twitter, LinkedIn e comentários na página da notícia.

10. Para todos os jornais analisados, o Facebook é a mídia por onde mais são repercutidas notícias econômicas e o que apresenta a maior variabilidade da quantidade de compartilhamentos. Também é a que mais vem crescendo, apresentando uma taxa de crescimento de $\sim$3.5 vezes ao ano.

11. A maioria das redes sociais são mais utilizada que a própria área de comentários disponibilizada pelo jornal. O que indica que o leitor muitas vezes quer compartilhar a

notícia em vez de opinar sobre seu conteúdo.

12. Períodos com mais publicações geram mais comentários mas não necessariamente mais repercussões.

13. O compartilhamento de notícias econômicas via Twitter vem crescendo gradualmente ao longo dos anos. Principalmente entre os meses de março e setembro.

14. Para todos os jornais percebe-se que o número de notícias publicadas vem diminuindo ao longo dos anos enquanto que o número de compartilhamento de suas notícias nas mídias sociais vem aumentando. Tal comportamento foi evidenciado em todos os jornais analisados e pode significar que a mídia estrutura a forma e conteúdo da notícia de modo obter maior repercussão.

Capítulo 6

Análise de Polaridade e valores Extremos

Neste capítulo, serão apresentadas as análises de polaridade e de valores extremos ou *outliers* que foram utilizadas para todas as notícias coletadas.

Inicialmente é descrita a metodologia de classificação de polaridade adotada. Em seguida, é realizada análise das quantidades de notícias positivas, negativas e neutras em diferentes granularidades de tempo. O comportamento dessas quantidades é verificado entre os jornais G1, Estadão e Folha de São Paulo. Por fim, apresenta-se uma discussão dos valores encontrados e suas contradições em relação a fatos da economia.

A parte final deste capítulo apresenta a análise dos títulos de notícias que levaram a comportamentos extremos por parte dos leitores, seja compartilhamento em redes sociais ou pelo número de comentários recebidos. Nesta análise buscou-se compreender que palavras contidas nos títulos das notícias estão mais relacionadas com comportamentos atípicos, superiores a tendência central.

6.1 Polaridade

Entende-se por polaridade a qualidade que permite distinguir a orientação de determinada informação entre positiva ou negativa [Feldman, 2013]. Enquanto algumas análises de sentimento preocupam-se em detectar os sentimentos associados a um determinado texto como por exemplo: amor, ódio, felicidade, angústia, entre outros, a análise de sentimento deste trabalho atem-se a classificar um texto como sendo positivo ou negativo (ou neutro quando não é possível detectar a polaridade).

Em certa medida, a classificação da polaridade de um texto é subjetiva e dependente de contexto. Por exemplo, para o seguinte título de notícia: *"Dólar sobe 1% e fecha a R$3,943, de olho na China e no cenário político."*. Ao ser lida por alguém que possui dólar este título pode naturalmente ser considerado positivo sem constituir incoerência alguma. Por outro lado, em um contexto econômico geral, a subida do dólar é acompanhada de aumento no preço do pão, combustível, passagens de ônibus, remédios e todos os elementos importados, sendo também possível a classificação da notícia como negativa. Para um outro exemplo, tem-se o seguinte título: *"Contas do governo têm pior resultado para todos os meses em 19 anos"*. Para este caso, a notícia já afirma algo negativo sobre o sujeito principal e a classificação torna-se óbvia.

Nesta perspectiva, entende-se que a classificação ideal é aquela que leva em consideração o que se diz sobre o sujeito principal da notícia ou o que a notícia reflete em contexto econômico geral.

6.2 Metodologia

A metodologia seguida para a classificação da polaridade de notícias foi a seguinte:

- Inicialmente, foram selecionadas aleatoriamente 200 notícias da base de notícias coletadas.
- As notícias foram então agrupadas em grupos de 10 e 15 notícias para que pudessem ser classificadas.
- Foi solicitado que cada aluno do curso de Ciência da Computação que cursava a disciplina Análise de Dados II no segundo semestre de 2015 lesse e classificasse um grupo de notícias em relação ao sujeito principal da notícia ou em relação ao contexto econômico nacional.
- Notícias que foram classificadas de forma diferente por diferentes alunos foram descartadas conforme recomendado pela *análise delphi* [Linstone et al., 1975]. Ao final restaram 165 notícias igualmente classificadas. Em primeiro lugar a significância estatística depende da variância da amostra, isto é, $n = (\frac{Z_{\alpha/2}}{E})^2$, onde n é o tamanho da

amostra, E é a dimensão do erro tolerado e $Z_{\alpha/2}$ é o nível de confiança. Uma prova prática desse conceito pode ser encontrada nos exames de sangue onde a quantidade de sangue utilizado é muito menor que a quantidade total presente no corpo, porém a amostra tende a ter pouquíssima variabilidade. Fato este também verificado neste experimento. Outro ponto importante é que ao final, como será visto posteriormente, os resultados obtidos por meio da análise de polaridade não foram superiores a outras análises de correlação o que torna inválido os argumentos decorrentes desta análise de modo a justificar os resultados encontrados.

- As notícias classificadas serviram de treino e verificação de métodos de análise de polaridade de modo a selecionar o que apresentasse a melhor acurácia para então utilizá-lo em toda a base.

- Todas as notícias classificadas pelos alunos foram reclassificadas por 19 métodos de análise de polaridade. Os métodos e os resultados são apresentados na Tabela 6.1.

- Os métodos que apresentaram melhores resultados foram o SO-CAL [Voll and Taboada, 2007] e o Vader [Hutto and Gilbert, 2014], com 73% de acerto.

- Por fim, o método Vader foi escolhido baseado nos seguintes critérios:

  - **Tempo de Execução:** É necessário levar em consideração que o método escolhido deveria ser utilizado para classificar toda a base de notícias em tempo aceitável. Utilizando-se os recursos de hardware descritos na seção 4.1.1, o método SO-CAL processou as 165 notícias em $\sim 240s$. O método Vader em $\sim 42s$.

  - **Recência do Trabalho:** O método Vader foi publicado em 2014 enquanto o SO-CAL em 2007.

  - **Facilidade de Adaptação:** O método Vader está em diversas plataformas para ser facilmente utilizado e adaptado[36]. O método SO-CAL exige a instalação de inúmeros componentes secundários que deve ser feita caso a caso e são provenientes de URLs não oficiais.

[36]https://github.com/cjhutto/vaderSentiment

Observações

A implementação dos métodos testados, com exceção do Vader e do SO-CAL, foram obtidas da ferramenta iFeel[37] [Gonçalves et al., 2013a].

Como muitos dos métodos foram concebidos para o idioma Inglês, todas as notícias tiveram de ser traduzidas antes de submetidas conforme comentado na seção 4.3. Segundo os resultados apresentado por [Araújo et al., 2016] essa estratégia apresentou bons resultados para o idioma português. O fato é que muitos métodos são baseados em léxicos e esses léxicos mantem-se preservados após a tradução.

6.3 Análise Geral

A seguir será apresentada a análise do comportamento da polaridade das notícias econômicas publicadas nos jornais G1, Folha de São Paulo e Estadão em diferentes perspectivas e granularidades de tempo.

6.3.1 Ano

A Figura 6.1 mostra a quantidade de notícias econômicas classificadas como positiva, negativa e neutra para todos os jornais. A quantidade de notícias classificadas como positivas é de ~ 4.7 vezes superior ao de notícias negativas. Mesmo em anos onde a economia apresentou problemas e perda de crescimento, o número de notícias econômicas classificadas como positivas superou o número de notícias negativas.

6.3.2 Mês

A Figura 6.2 apresenta a quantidade de notícias classificadas como positivas, negativas e neutras publicadas pelos jornais analisados ao longo dos meses do ano para todos os anos. A quantidade de notícias negativas e neutras é praticamente a mesma para todos os meses analisados. Apesar disso, tanto o cenário econômico quanto sua perspectiva de futuro oscilaram em vários desses meses.

[37] http://blackbird.dcc.ufmg.br:1210/

Métodos	Acertos	Erros	Acurácia
SentiWordNet **[Esuli and Sebastiani, 2006]**	97	68	0.59
SenticNet **[Cambria et al., 2010]**	95	70	0.58
Emoticons **[Park et al., 2013]**	8	157	0.05
Panas-t **[Gonçalves et al., 2013b]**	22	143	0.13
Sasa **[Wang et al., 2012]**	26	139	0.16
Happiness Index **[Dodds and Danforth, 2010]**	97	68	0.59
Sentistrength **[Thelwall, 2013]**	22	143	0.13
Emolex **[Mohammad and Turney, 2013a]**	102	63	0.62
NRC Emotion **[Mohammad and Turney, 2013b]**	54	111	0.33
Opinion Lexicon **[Ding et al., 2008]**	99	66	0.60
Emoticon Distant Supervisor **[Nunes Ribeiro et al., 2015]**	93	72	0.56
SO-CAL [Taboada et al., 2008]	**121**	**44**	**0.73**
Pattern.En **[De Smedt and Daelemans, 2012]**	97	68	0.59
Umigon **[Levallois, 2013]**	75	90	0.45
AFINN **[Nielsen, 2011]**	113	52	0.68
OpinionFinder **[Wilson et al., 2005]**	59	106	0.36
Vader [Cambria et al., 2010]	**120**	**45**	**0.73**
Sentiment 140 **[Go et al., 2009]**	104	61	0.63
Combined Method **[Gonçalves et al., 2013a]**	101	64	0.61

Tabela 6.1: Lista dos métodos utilizados para análise de polaridade seguida do número de acertos, erros e a porcentagem de acerto obtido ao final da análise.

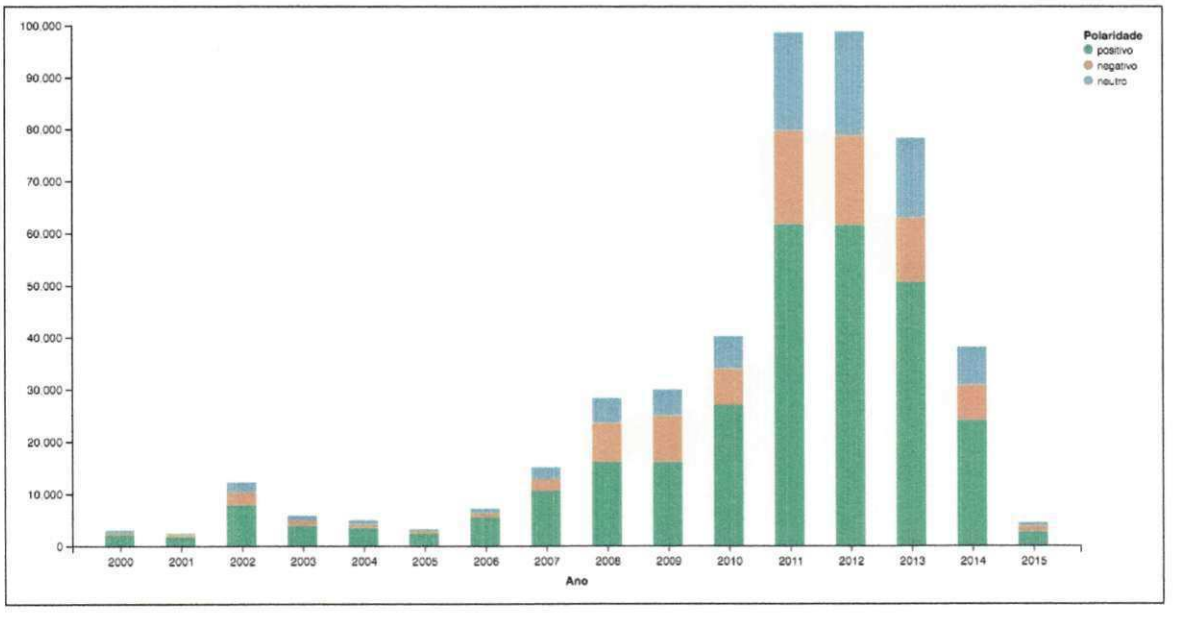

Figura 6.1: Quantidade de notícias positivas, negativas e neutras publicadas ao longo dos anos.

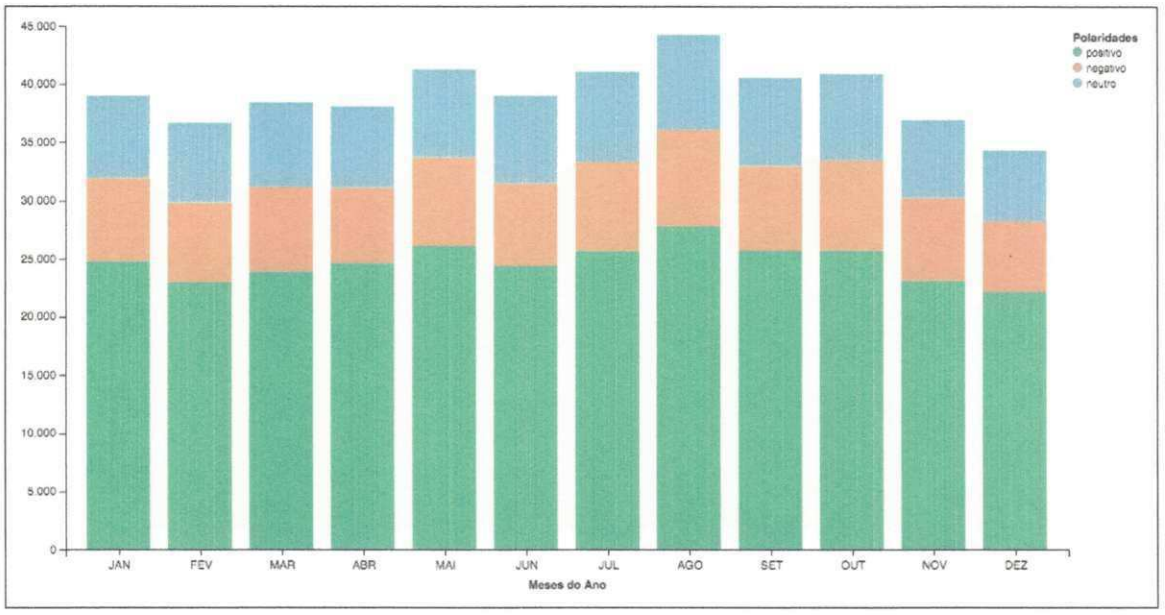

Figura 6.2: Quantidade de notícias positivas, negativas e neutras publicadas ao longo dos meses.

6.3.3 Dia

A Figura 6.3 apresenta a classificação da polaridade de notícias econômicas publicadas pelos jornais analisados para os dias do mês. Novamente as proporções entre notícias negativas e neutras são mantidas ao longo dos dias do mês. Há uma pequena predileção de notícias positivas para os dias 14, 15, 27 e 28. E uma ligeira diferença em notícias negativa para o dia 10.

6.3.4 Dia da Semana

A Figura 6.4 apresenta a classificação da polaridade de notícias econômicas publicadas pelos jornais analisados durante os dias da semana. Percebe-se que de forma geral também são mantidas as mesmas proporções entre as quantidades de notícias positivas, negativas e neutras. Os finais de semana – Sábados e Domingos – em geral possuem bem mais notícias positivas em relação às demais polaridades.

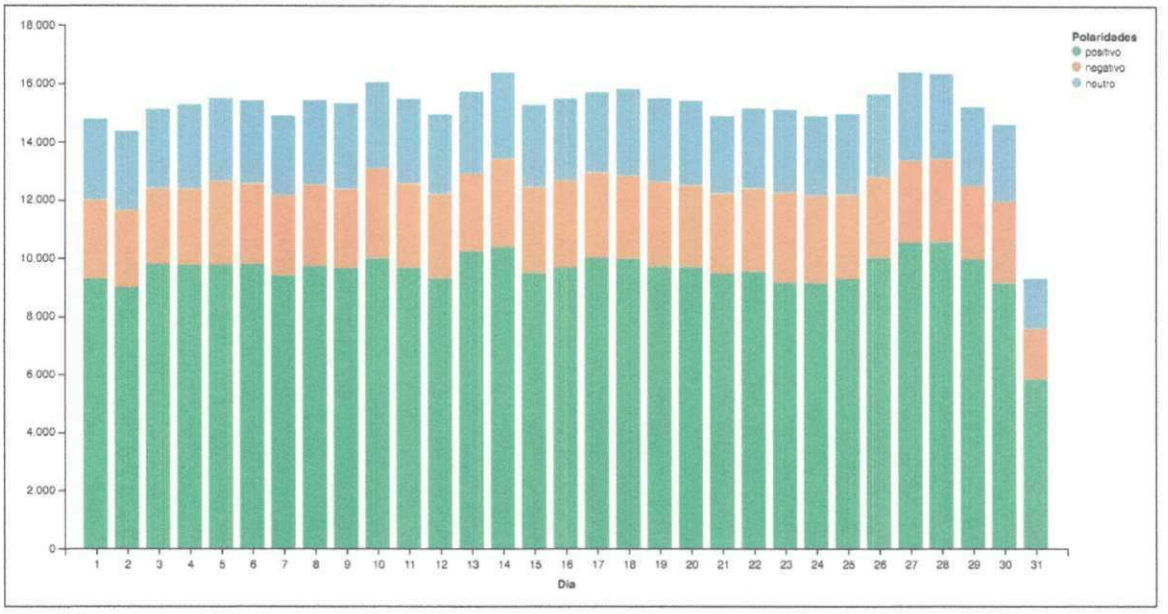

Figura 6.3: Quantidade de notícias positivas, negativas e neutras publicadas durante os dias do mês.

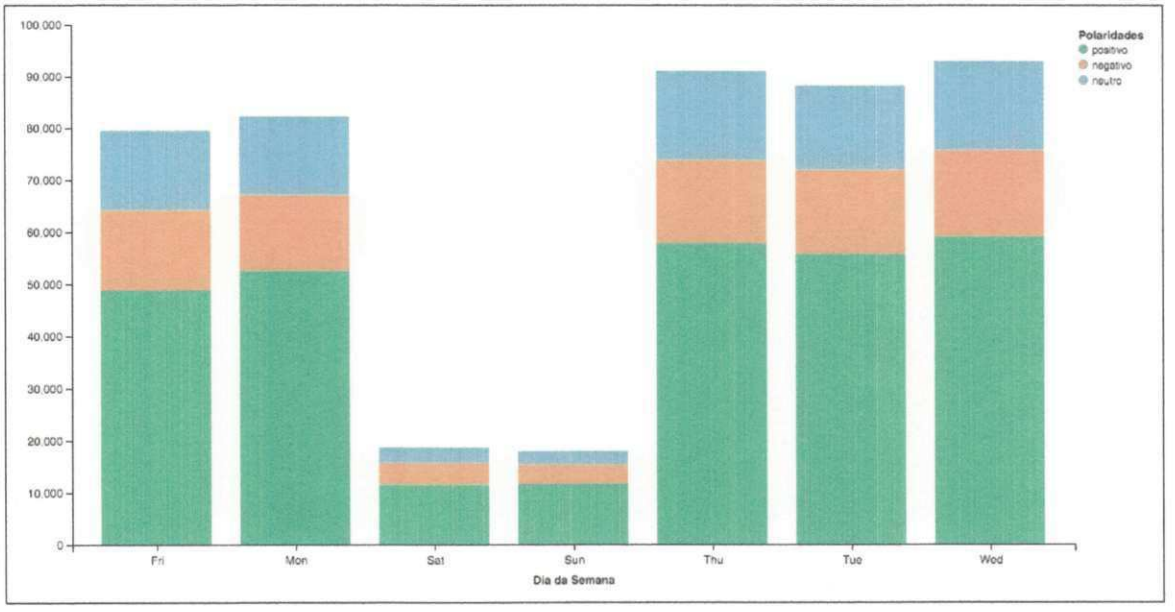

Figura 6.4: Quantidade de notícias positivas, negativas e neutras publicadas durante os meses do ano.

6.3.5 G1

A Figura 6.5 apresenta o comportamento das notícias econômicas publicadas pelo jornal G1 ao longo dos anos referente a quantidade de notícias classificadas de acordo com sua polaridade. Em média o G1 publica ~ 4 vezes mais notícias positivas que notícias negativas. Em anos de eleições presidenciais (2010 e 2014) essa proporção aumentou para a proporção para $5:1$.

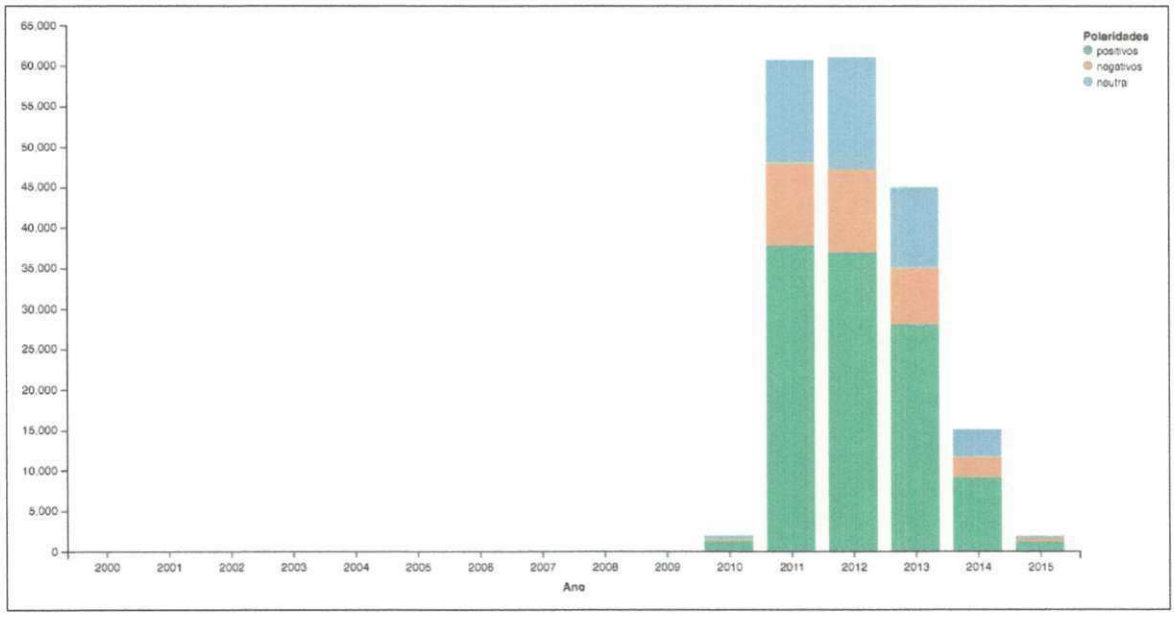

Figura 6.5: Quantidade de notícias positivas, negativas e neutras ao longo dos anos para o jornal G1.

6.3.6 Folha

A Figura 6.6 apresenta o comportamento das notícias econômicas publicadas pelo jornal Folha de São Paulo referente a quantidade de notícias classificadas de acordo com sua polaridade ao longo dos anos. Em média a Folha de São Paulo publica ~ 2.5 vezes mais notícias positivas que notícias negativas. Assim como no G1 em anos de eleições presidenciais essa proporção aumenta para $5:1$.

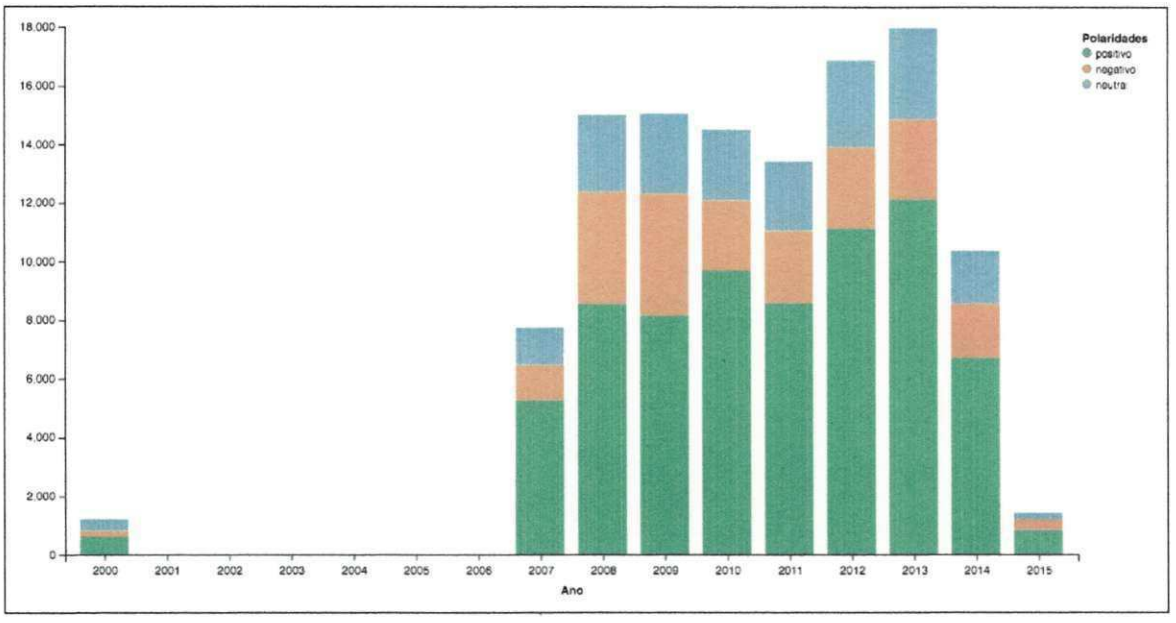

Figura 6.6: Quantidade de notícias positivas, negativas e neutras ao longo dos anos para o jornal Folha de São Paulo.

6.3.7 Estadão

A Figura 6.7 apresenta o comportamento das notícias econômicas publicadas pelo jornal Estadão referente a quantidade de notícias classificadas de acordo com sua polaridade ao longo dos anos. Em média o jornal Estadão publica ~ 3 vezes mais notícias positivas que negativas. Assim como visto em outros jornais em anos de eleições presidenciais essa proporção aumenta para $4 : 1$ – Nessa análise é possível verificar outros anos eleitorais como 2002, 2006, 2010 e 2014.

6.4 Análise de Repercussões Extremas

Para esta análise, foram selecionadas apenas notícias que tiveram uma repercussão superior ao terceiro quartil $(Q3 + 1,5 * IQR)$[38] em cada uma das mídias consideradas.

Para cada notícia, foram extraídas e contabilizadas as frequências de cada palavra presente no título e texto.

[38]IQR ou (interquartile range) é a diferença entre o quartil de maior valor (75%) e o de menor valor (25%). $IQR = Q_3 - Q_1$

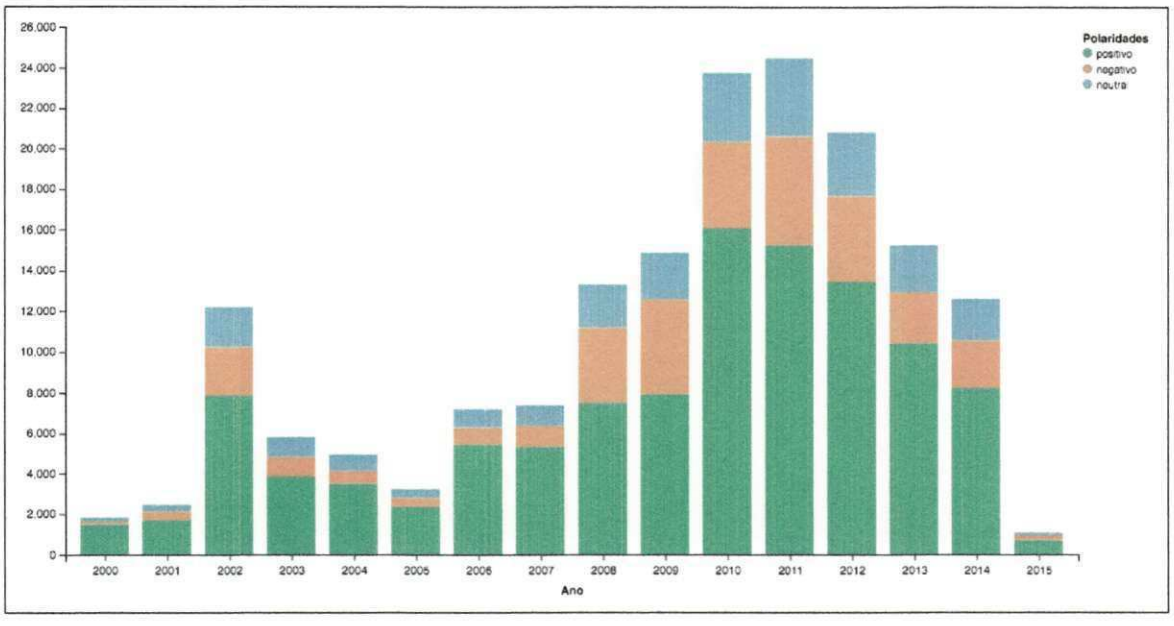

Figura 6.7: Quantidade de notícias positivas, negativas e neutras ao longo dos anos para o jornal Estadão.

Esta investigação permitiu verificar que palavras estão associadas a altas repercussões e que tipo de notícia os leitores de cadernos econômicos dos jornais analisados se sentem motivados a compartilhar de forma efusiva e qual ambiente escolhem para fazê-lo.

6.4.1 Análise de Títulos – TOP-15

A Tabela 6.2 apresenta uma lista de até 15 palavras, ordenadas de forma decrescente por frequência com que apareceram em notícias amplamente repercutidas por leitores de notícias econômicas. A análise por linha informa que palavras presentes nas notícias geraram grandes repercussões na mídia de referência. Por outro lado, a análise por coluna informa que palavras utilizadas pelo jornal de referência resultaram em grande repercussão de suas notícias nas mídias analisadas.

Mídias

Observando as linhas da Tabela 6.2 percebe-se que palavras de conotação negativas como *queda* e *cai* aparecem preferencialmente em títulos de notícias que foram amplamente compartilhadas via Twitter. De forma contrária, as notícias econômicas amplamente divulgadas

via LinkedIn, Facebook e Google Plus tem aparente predileção por títulos de conotação positiva ou neutra, como por exemplo *alta* e *sobe*.

É possível perceber que o LinkedIn é a mídia escolhida para compartilhamento de notícias de conteúdo técnico. Títulos contendo as palavras *inflação*, *juros*, *trimestre* e *lucro* foram amplamente compartilhados apenas nessa mídia. Sendo o LinkedIn uma rede social de objetivo profissional, essa análise é bastante pertinente. A notícia mais compartilhada pelo LinkedIn, por exemplo, trata do rebaixamento da nota da Vale por parte da agência de risco S&P [39].

Termos como *pão*, *IPI* e *PIB* geraram largas discussões de leitores em formato de comentários da página mas não compartilhamentos em mídias sociais.

Jornais

De forma geral, as notícias econômicas mais compartilhadas publicadas pelo jornal G1 enfatizam comportamentos ao invés de substantivos. Palavras como : *queda*, *alta*, *fecha*, *sobe* e *mais* em geral estão sempre associadas a medo e emergência confirmando o estilo jornalistico do G1 que vem sendo mostrado ao longo deste trabalho. Os leitores do jornal Folha de São Paulo, ao contrário dos leitores do jornal G1, enfatizam notícias sobre substantivos. Independente de boas ou más notícias, títulos que continham palavras com: *Brasil*, *Bovespa*, *governo* e *Petrobrás* foram bastante compartilhados[40]. Para o jornal Estadão o que chama atenção é a quantidade de nomes e eventos externos em evidência. Palavras como: *Levy*, *Dilma*, *Rodrigues*, *China* e *EUA* são sempre amplamente compartilhados. Aparentemente leitores do Estadão compartilham com ênfase notícias de fatores externos e de personalidades que podem influenciar na economia interna, demonstrando uma visão mais macroeconômica dos fatos.

6.4.2 Limitações

A análise de palavras possui algumas limitações, principalmente a interpretação semântica e as dependências de contexto. Por exemplo, a palavra *ações* que aparece na lista de comen-

[39]http://www1.folha.uol.com.br/mercado/2015/01/1579719-agencia-de-risco-sp-corta-nota-da-vale-por-queda-de-preco-do-minerio.shtml

[40]O nome "Eike"é imensamente compartilhado apenas pelo jornal Folha de São Paulo.

Jornais Métodos	G1	Folha de São Paulo	Estadão
Comentários	–	bovespa,brasil,alta, fecha,bolsas,mais,maior, pão,após,governo,queda, ações, vai,ipi,petróleo.	agência,mais,alta, brasil,pib,governo, china,tem,dólar, inflação,bolsas,eua, queda,economia.
Twitter	queda,petrobras,tem, alta,brasil,após, eua,dólar,bovespa, fecha,nesta,mais,sobe.	governo,caminhoneiros, petrobras,Brasil,eike, dólar,sobe,mais,fiscal, bolsa,após,vai,cai,país.	negócios,diz,Brasil,mais, tem,após,alta,china, eua,dólar,R$,BC,pib, inflação,dilma.
Facebook	alta,Brasil,que,após, eua,sobe,bovespa, milhões,fecha,dólar, petrobras,mais,ano, sobre.	petrobras,governo,por, caminhoneiros,mais, após,dólar,brasil,bolsa, sobe,tem,alta,veja, eike,sobre.	negócios,brasil,joão, dantas,não,villaverde, dólar,governo,rodrigues, adriana,fernandes,2015, mais,murilo,alves,
LinkedIn	lucro,alta,por,eua, dólar,Brasil,tem, mais,trimestre, sobe, inflação,fecha.	diz,petrobras,maior, energia,brasil,mais deve,tem,ano,bolsa, inflação,juros,china,vale.	economia,diz,estado, negócios,Brasil,agência, mais,alta,correspondente, tem,vai,levy,deve,R$.
Google Plus	–	–	estado,agência,alta, com,economia,diz, mais,inflação,tem,caixa, abril,dólar,bolsas,R$.

Tabela 6.2: Lista de palavras presentes nas notícias dos jornais que mais causaram repercussão as redes sociais analisadas

tários do jornal Folha de São Paulo não nos permite saber se as ações referem-se a *ações da bolsa* ou são *ações tomadas pelo governo*, por exemplo. O mesmo vale para a palavra *veja*, onde não é possível saber se a palavra faz referência a revista *Veja* ou o verbo ver no imperativo. Algumas vezes foi necessária a intervenção manual para ajustar a escrita de algumas palavras, como, por exemplo, Petrobrás com e sem acento.

Para uma análise mais detalhada é necessário repetir o processo e analisar a frequência não apenas de uma palavra, mas de grupos de *n-gramas*. Dessa forma seria possível ir além das palavras, e tirar conclusões sobre expressões. Também seria interessante relacionar os títulos de maior engajamento dos leitores com uma componente temporal e saber se ocorrem em momentos distintos (o que daria força a ideia da influência dos títulos) ou se ocorrem em momentos específicos do ano (o que reforçaria também uma dependência temporal ao invés de apenas semântica).

6.5 Considerações Finais

A seguir são apresentadas as conclusões da análise de polaridade para as notícias coletadas:

1. O método Vader apresentou melhor resultado e desempenho que outros métodos para a classificação de polaridade de notícias.

2. Mantém-se uma proporcionalidade de ~ 4.5 vezes o número de notícias econômicas classificadas como positivas em relação as classificadas como negativas em todas as granularidades de tempo consideradas. É possível especular que essa característica é uma forma de não desmotivar os leitores do caderno econômico quanto as perspectivas do país e do mercado.

3. Em anos de eleições presidenciais todos os jornais diminuem a quantidade de notícias econômicas negativas. G1 e Folha de São Paulo publicam aproximadamente 50% menos notícias negativas enquanto o Estadão 25%. Também foi verificado que todos os jornais também aumentam a quantidade de notícias positivas publicadas na proporção de $4:1$.

4. Leitores do jornal Estadão repercutem com ênfase informações sobre pessoas – Dilma, Levy, Rodrigo – e assuntos externos, principalmente referentes a EUA e China. Nesse

ponto de vista percebe-se que seus leitores motivam-se por notícias que se relacionam *"Quem"* e *"Onde"* impactam na economia.

5. Leitores do jornal Folha de São Paulo inclinam-se em compartilhar de forma efusiva notícias que comentem sobre *"O que"* afeta na economia. Títulos com palavras *bolsa*, *juros*, *governo* e *caminhoneiros* ganham muitos compartilhamentos.

6. Por fim, os leitores do G1 sentem-se motivados a fatos econômicos, ou seja, que enfatizem o *"Como"* determinado fato afeta a economia, seja alta, baixa, cai, sobe, entre outros.

7. Palavras como *Brasil*, *Petrobrás*, *dólar* e *inflação* são palavras-chave que geram uma mobilização maior que o normal em todos os jornais. Interessante que para os leitores de notícias econômicas e de sua perspectiva, notícias que comentam sobre o *dólar* geram mais repercussão que as que comentam sobre a própria moeda nacional o *real*.

Capítulo 7

Índice Bovespa

Este capítulo apresenta ainda uma análise de correlação realizada entre os atributos coletados e o índice Bovespa. O objetivo principal foi conhecer o grau de correlação de cada atributo em relação ao índice Bovespa (IBOVE), investigar seus porquês e, por fim, utilizá-los em modelos preditivos de modo a prever valores futuros. Ao invés de limitar-se à predição de empresas isoladas, ou a um pequeno grupo de empresas amplamente divulgadas pela mídia nacional, como por exemplo, Petrobrás, Vale e Banco do Brasil, buscou-se construir modelos de predição que fossem hábeis em prever o comportamento do mercado de forma geral. Neste sentido, o índice Bovespa foi selecionado por sua característica de sumarizar o desempenho das principais empresas do mercado nacional, conforme detalhado na seção 2.1.3.

7.1 Experimento

A seguir são apresentadas as condições de realização do experimento:

1. Os dados históricos do IBOVE foram obtidos da revista online Exame[41] por já estarem em formato estruturado e processado. Estes dados contém, além do preço da carteira, outros atributos como quantidade negociada e variação. Por outro lado, a série histórica do índice Bovespa, fornecida pelo site da BM&FBOVESPA[42], contém apenas o valor do índice.

[41] http://exame.abril.com.br/mercados/cotacoes-bovespa/indices/BVSP/historico

[42] http://www.bmfbovespa.com.br/indices/ResumoEvolucaoDiaria.aspx?Indice=IBOVESPA&idioma=pt-br

2. Dias para os quais não havia notícias e dias para os quais não havia negociação do mercado (Sábados, Domingos e Feriados) foram removidos. É possível questionar que as notícias publicadas desde o fechamento do último dia de bolsa possam influenciar o próximo dia de bolsa e dessa maneira a análise de correlação deveria buscar uma forma de agregar as notícias que são publicadas em dias onde houve notícias econômicas, porém, não houve abertura da bolsa. Especificamente nesse experimento essas considerações não foram implementadas. Foram correlacionados apenas dias onde coexistiam notícias e IBOVE.

3. A janela de tempo considerada neste experimento de análise de correlação é dia.

4. Apesar dos testes de correlação de **Kendall** (τ), **Spearman** (r_s) e **Pearson** (ρ) serem aplicados para dados contínuos [Chok, 2010] eles não podem ser utilizados indistintamente em relação a quaisquer bases de dados. O teste de correlação de Pearson mesmo sendo amplamente utilizado no mercado de ações, principalmente em trabalhos medindo a relação entre *commodities*, pressupõe que ambas as variáveis sejam normalmente distribuídas. Neste trabalho, a série temporal de valores do IBOVE não é normalmente distribuída e, sendo assim, *Pearson* não pôde ser utilizado. Apesar de não haver nenhuma premissa quanto a distribuição dos dados para o teste de *Spearman*, há a necessidade de ajustar o *ranking* sempre que há valores iguais em uma das variáveis. Em casos onde existe abundância de valores iguais este fato pode comprometer os resultados da correlação de forma significativa, mascarando-o. A sequencia de valores do IBOVE por si só apresenta muitos dias com valores iguais e, deste modo, o teste de *Spearman* não foi utilizado. Por fim, foi utilizado o teste de correlação de *Kendall*, por ser, assim como *Spearman*, não-paramétrico e não ter premissas quanto à distribuição dos dados. Tal teste possui sustentação algébrica para medir a força de dependência entre duas variáveis sem realizar ordenações e ajustes matemáticos para valores iguais. No teste de *Kendall*, verifica-se o número de pares concordantes e discordantes entre as variáveis, respeitando a disposição dos dados da série temporal. O teste de *Kendall* irá verificar se, ao variar o valor de uma das variáveis, a variável sendo comparada também irá variar nas mesmas proporções e sentido. Um fator pertinente que afeta a utilização do teste Kendall e impacta em seu coeficiente acontece quando

a quantidade de dados considerada é muito pequena, o que não ocorre com os dados sendo considerados.

7.2 Publicações, Compartilhamentos e o Índice Bovespa

Nesta seção são apresentadas as análises de correlação entre o índice Bovespa e os atributos relacionados as quantidades de publicações dos jornais, comentários das notícias, e seus compartilhamentos nas redes sociais. Todas as tabelas apresentadas expõem os resultados obtidos para o teste de normalidade, o grau de relacionamento linear entre a variável observada e o índice Bovespa (força da correlação) e o *p-value* encontrado quando submetidos ao teste de *Kendall*.

7.2.1 Quantidade de publicações de Notícias e Comentários e Índice Bovespa

A tabela 7.1 mostra os resultados do teste de correlação para os atributos de quantidade de publicações diárias e quantidade de comentários de notícias tanto para o agregado dos jornais quanto para cada jornal individualmente.

Há evidências de que existe uma correlação de 37% entre a variação da quantidade de publicações do jornal Estadão ao longo dos dias e o preço da carteira hipotética representada pelo índice Bovespa. As Figuras 7.1 e 7.2 apresentam respectivamente as séries temporais do índice Bovespa com quantidade de publicações diárias do jornal Estadão e o gráfico de dispersão ou *scatter plot*. Ambas as figuras apresentam, de forma gráfica, a relação linear existente entre o índice Bovespa e a quantidade de publicação de notícias econômicas do Jornal Estadão.

7.2.2 Compartilhamentos de Notícias Econômicas nas Redes Sociais e o Índice Bovespa

A Tabela 7.2 apresenta os valores que foram obtidos após a realização da análise de correlação entre o índice Bovespa e a quantidade de compartilhamentos que as notícias econômicas publicadas pelos jornais tiveram nas redes sociais.

Atributos	Análise	Normalidade	Correlação	p-value
Publicações Diárias	**Jornais**	*não*	0.33	3.9e-129
	G1	*não*	-0.03	0.089
	Folha	*não*	0.12	3.9e-11
	Estadão	***sim***	**0.37**	**2.7e-93**
Quantidade de Comentários	**Jornais**	*não*	0.17	3.8e-21
	G1	–	–	–
	Folha	*não*	0.17	2.4e-21
	Estadão	*não*	0.02	0.18

Tabela 7.1: Valores obtidos pela análise de correlação entre o índice Bovespa e os atributos de quantidade de publicações diária e comentários.

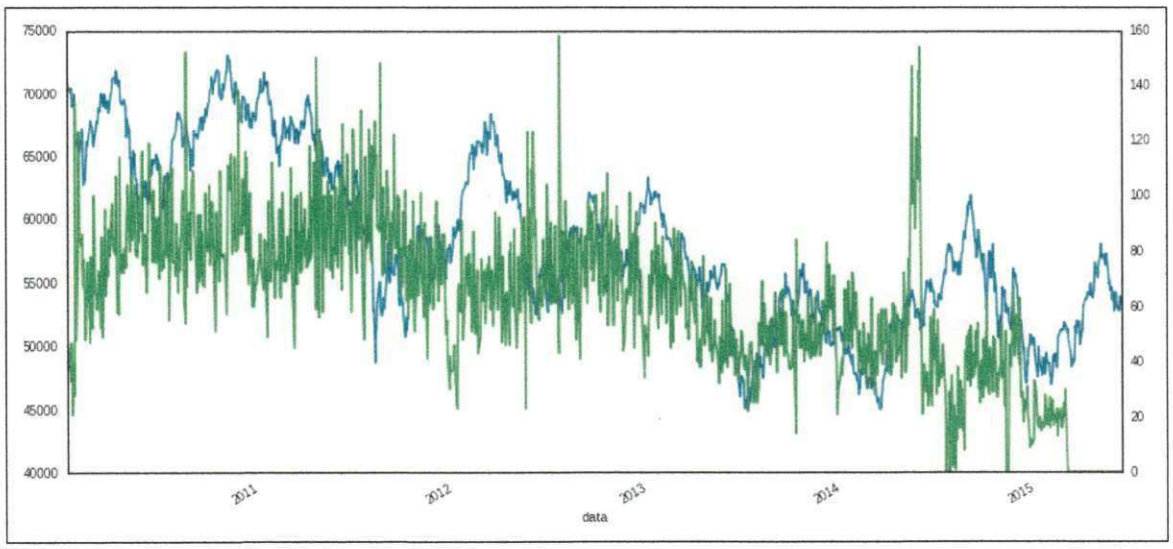

Figura 7.1: Sobreposição das séries temporais do IBOVE e da quantidade de publicações do jornal Estadão.

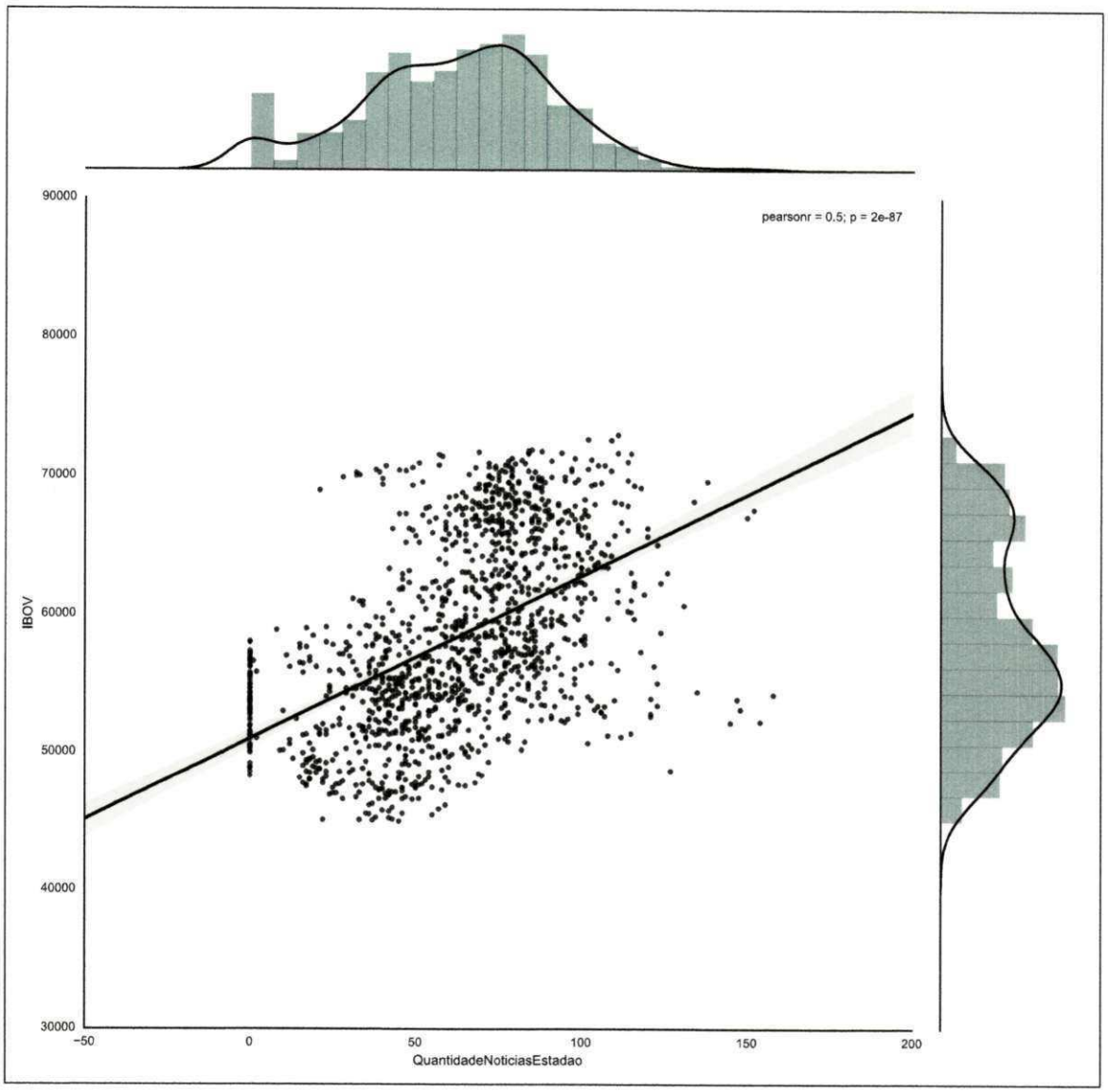

Figura 7.2: Gráfico de dispersão entre IBOVE e quantidade de publicações do jornal Estadão.

O jornal Estadão apresentou os maiores valores para o grau de correlação entre o índice Bovespa e a quantidade de compartilhamento de notícias via Twitter, LinkedIn e Google Plus. Também apresentou o maior grau de correlação de toda a análise com 46% para as notícias que são compartilhadas via Google Plus. As Figuras 7.3 e 7.4 mostram, respectivamente, o gráfico das séries temporais do índice Bovespa sobreposto pelos números de compartilhamentos de notícias do jornal Estadão via Google Plus e o gráfico de dispersão para os mesmos atributos. Ambos os gráficos refletem de forma detalhada a correlação de -46% que foi evidenciada na Tabela 7.2.

Atributos	Análise	Normalidade	Correlação	p-value
Twitter	**Jornais**	*não*	0.16	2.6e-35
	G1	*não*	-0.23	8.8e-37
	Folha	*não*	-0.25	8.5e-46
	Estadão	***não***	**-0.42**	**4.5e-123**
Facebook	**Jornais**	*não*	0.14	1.1e-25
	G1	***não***	**-0.42**	**4.2e-120**
	Folha	*não*	-0.39	2.9e-100
	Estadão	*não*	-0.41	2.1e-115
LinkedIn	**Jornais**	*não*	0.12	2.1e-19
	G1	*não*	-0.16	8.6e-21
	Folha	*não*	-0.24	4.3e-42
	Estadão	*não*	-0.34	3.5e-79
Google Plus	**Jornais**	–	–	–
	G1	–	–	–
	Folha	–	–	–
	Estadão	***não***	**-0.46**	**1.7e-146**
Repercussão Total	**Jornais**	*não*	0.17	7.8e-38

Tabela 7.2: Valores obtidos pela análise de correlação entre a quantidade de compartilhamentos das notícias econômicas ao longo do tempo e a variação do índice Bovespa.

7.2.3 Classificação da Polaridade das Notícias e o Índice Bovespa

A Tabela 7.3 apresenta a análise de correlação entre o índice Bovespa e a polaridade das notícias publicas diariamente pelos jornais G1, Folha de São Paulo e Estadão. Nesta análise foram correlacionados o número de notícias positivas, notícias negativas, e o saldo que é

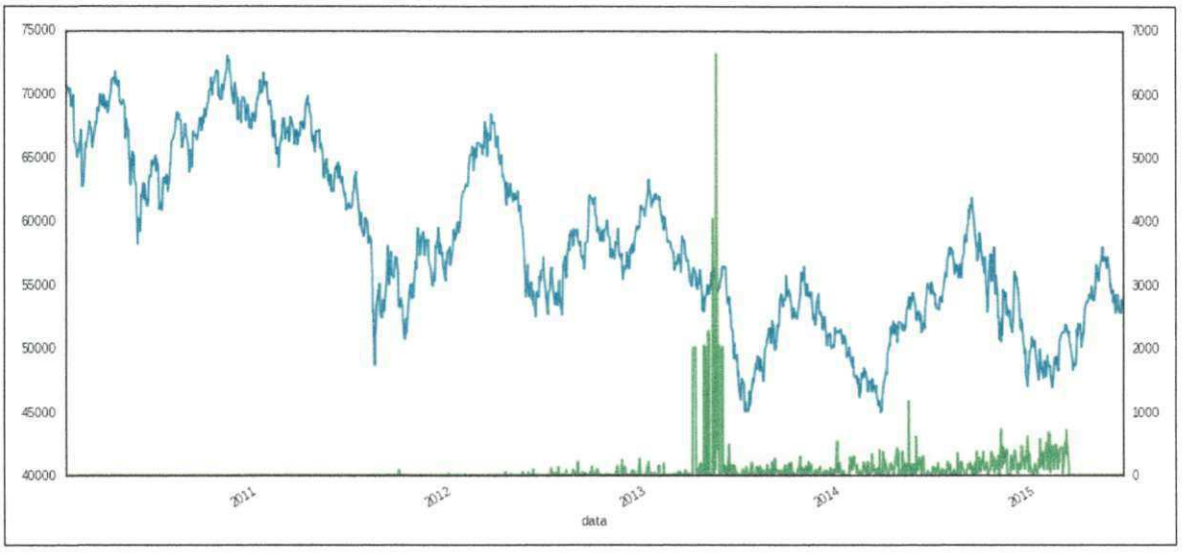

Figura 7.3: Sobreposição das séries temporais do índice Bovespa (em azul) e da quantidade de publicações do jornal Estadão compartilhadas via Google Plus (em verde).

referente a diferença entre positivas e negativas, que foram publicadas diariamente ao longo dos anos. Percebe-se que novamente as maiores correlações obtidas devem-se ao jornal Estadão com 38% e as menores do jornal G1 tendo uma correlação de praticamente zero em todas as análises.

7.3 Modelos de Previsão do Índice Bovespa

Esta seção mediu a capacidade preditiva do índice Bovespa por meio de algoritmos de aprendizagem de máquina que usam atributos extraídos de notícias.

A seguir, são detalhados a formalização do problema, o *design* dos experimentos e os modelos construídos para prever o índice Bovespa.

7.4 Preparação dos Dados

O capítulo 4 apresentou em detalhes todo o processo de coleta e armazenamento de dados para realização desta investigação.

A seguir serão detalhados os processos que envolveram a preparação de dados antes de serem utilizados na construção dos modelos preditivos.

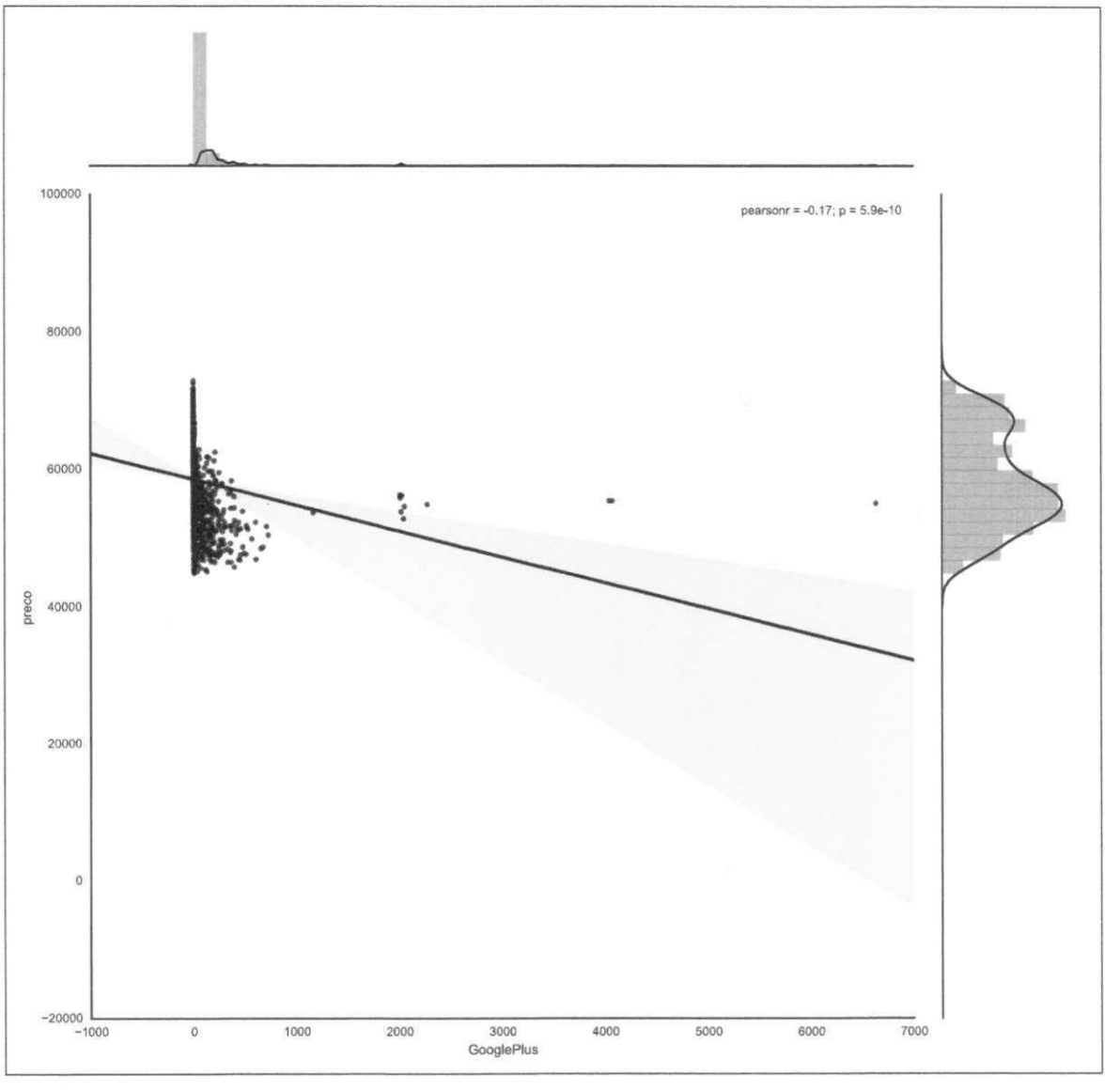

Figura 7.4: Gráfico de dispersão entre IBOVE e quantidade de publicações do jornal Estadão compartilhada via Google Plus.

Atributos	Análise	Normalidade	Correlação	p-value
Quantidade de Notícias Positivas	**Jornais**	*não*	0.11	1.8e-10
	G1	*não*	-0.001	0.91
	Folha	*não*	0.15	7.1e-17
	Estadão	***sim***	**0.38**	**2.7e-104**
Quantidade de Notícias Negativas	**Jornais**	*não*	0.021	0.23
	G1	*não*	-0.08	8.0e-06
	Folha	*não*	0.013	0.46
	Estadão	***não***	**0.24**	**2.8e-39**
Saldo de Notícias do Dia	**Jornais**	*não*	0.13	6.8e-14
	G1	*não*	0.01	0.45
	Folha	*não*	0.16	6.3e-19
	Estadão	***não***	**0.34**	**2.7e-82**

Tabela 7.3: Análise de correlação entre as polaridades das notícias econômicas publicadas e o índice Bovespa

1. Cada linha do arquivo de treino contém informações sobre uma notícia publicada entre 16/06/2005 e 02/03/2015. Os atributos para cada notícia são:

   - **Data e Hora:** Este atributo contém o ano, o mês, o dia e a hora em que a notícia foi publicada no formato *YYYYMMDDHHHH*. Este atributo é do tipo *numérico* com escala *intervalar*. Seu domínio compreende os números naturais entre 200516060000 à 201502032359.
   - **Quantidade de notícias:** A quantidade de notícias que foram publicadas em uma determinada janela de tempo. Este atributo é do tipo *numérico* com escala de *razão*. Seu domínio compreende os números naturais.
   - **Média da Quantidade de Comentários:** A média da quantidade de comentários publicados em uma determinada janela de tempo. Este atributo é do tipo *numérico* com escala de *razão*. Seu domínio compreende os números reais.
   - **Mediana da Quantidade de Comentários:** A mediana da quantidade de comentários publicados em uma determinada janela de tempo. Este atributo é do tipo *numérico* com escala de *razão*. Seu domínio compreende os números naturais.
   - **Desvio Padrão da Quantidade de Comentários:** O desvio padrão da quan-

tidade de comentários publicados em uma determinada janela de tempo. Este atributo é do tipo *numérico* com escala de *razão*. Seu domínio compreende os números reais.

- **Variância da Quantidade de Comentários:** A variância da quantidade de comentários publicados em uma determinada janela de tempo. Este atributo é do tipo *numérico* com escala de *razão*. Seu domínio compreende os números reais.
- **Somatório da Quantidade de Comentários:** A média da quantidade de comentários publicados em uma determinada janela de tempo. Este atributo é do tipo *numérico* com escala de *razão*. Seu domínio compreende os números naturais.
- **Média de Compartilhamentos via Twitter:** A média da quantidade de compartilhamentos que uma notícia recebeu via Twitter após sua publicação. Este atributo é do tipo *numérico* com escala de *razão*. Seu domínio compreende os números reais.
- **Mediana da Quantidade de Twitter:** A mediana da quantidade de compartilhamentos que uma notícia recebeu via Twitter após sua publicação. Este atributo é do tipo *numérico* com escala de *razão*. Seu domínio compreende os números naturais.
- **Desvio Padrão da Quantidade de Twitter:** O desvio padrão da quantidade de compartilhamentos que uma notícia recebeu via Twitter após sua publicação. Este atributo é do tipo *numérico* com escala de *razão*. Seu domínio compreende os números reais.
- **Variância da Quantidade de Twitter:** A variância da quantidade de compartilhamentos que uma notícia recebeu via Twitter após sua publicação. Este atributo é do tipo *numérico* com escala de *razão*. Seu domínio compreende os números reais.
- **Somatório da Quantidade de Twitter:** A soma total da quantidade de compartilhamentos que uma notícia recebeu via Twitter após sua publicação. Este atributo é do tipo *numérico* com escala de *razão*. Seu domínio compreende os números naturais.

- **Média da Quantidade de Facebook:** A média da quantidade de compartilhamentos que uma notícia recebeu via Facebook após sua publicação. Este atributo é do tipo *numérico* com escala de *razão*. Seu domínio compreende os números reais.
- **Mediana da Quantidade de Facebook:** A mediana da quantidade de compartilhamentos que uma notícia recebeu via Facebook após sua publicação. Este atributo é do tipo *numérico* com escala de *razão*. Seu domínio compreende os números naturais.
- **Desvio Padrão da Quantidade de Facebook:** O desvio padrão da quantidade de compartilhamentos que uma notícia recebeu via Facebook após sua publicação. Este atributo é do tipo *numérico* com escala de *razão*. Seu domínio compreende os números reais.
- **Variância da Quantidade de Facebook:** A variância da quantidade de compartilhamentos que uma notícia recebeu via Facebook após sua publicação. Este atributo é do tipo *numérico* com escala de *razão*. Seu domínio compreende os números reais.
- **Somatório da Quantidade de Facebook:** A soma total da quantidade de compartilhamentos que uma notícia recebeu via Facebook após sua publicação. Este atributo é do tipo *numérico* com escala de *razão*. Seu domínio compreende os números naturais.
- **Média da Quantidade de LinkedIn:** A média da quantidade de compartilhamentos que uma notícia recebeu via LinkedIn após sua publicação. Este atributo é do tipo *numérico* com escala de *razão*. Seu domínio compreende os números reais.
- **Mediana da Quantidade de LinkedIn:** A mediana da quantidade de compartilhamentos que uma notícia recebeu via LinkedIn após sua publicação. Este atributo é do tipo *numérico* com escala de *razão*. Seu domínio compreende os números naturais.
- **Desvio Padrão da Quantidade de LinkedIn:** O desvio padrão da quantidade de compartilhamentos que uma notícia recebeu via LinkedIn após sua publicação.

Este atributo é do tipo *numérico* com escala de *razão*. Seu domínio compreende os números reais.

- **Variância da Quantidade de LinkedIn:** A variância da quantidade de compartilhamentos que uma notícia recebeu via LinkedIn após sua publicação. Este atributo é do tipo *numérico* com escala de *razão*. Seu domínio compreende os números reais.
- **Somatório da Quantidade de LinkedIn:**A soma total da quantidade de compartilhamentos que uma notícia recebeu via LinkedIn após sua publicação. Este atributo é do tipo *numérico* com escala de *razão*. Seu domínio compreende os números naturais.
- **Média da Quantidade de Google Plus:** A média da quantidade de compartilhamentos que uma notícia recebeu via Google Plus após sua publicação. Este atributo é do tipo *numérico* com escala de *razão*. Seu domínio compreende os números reais.
- **Mediana da Quantidade de GooglePlus:** A mediana da quantidade de compartilhamentos que uma notícia recebeu via Google Plus após sua publicação. Este atributo é do tipo *numérico* com escala de *razão*. Seu domínio compreende os números naturais.
- **Desvio Padrão da Quantidade de GooglePlus:** O desvio padrão da quantidade de compartilhamentos que uma notícia recebeu via Google Plus após sua publicação. Este atributo é do tipo *numérico* com escala de *razão*. Seu domínio compreende os números reais.
- **Variância da Quantidade de GooglePlus:** A variância da quantidade de compartilhamentos que uma notícia recebeu via Google Plus após sua publicação. Este atributo é do tipo *numérico* com escala de *razão*. Seu domínio compreende os números reais.
- **Somatório da Quantidade de GooglePlus:** A soma total da quantidade de compartilhamentos que uma notícia recebeu via Google Plus após sua publicação. Este atributo é do tipo *numérico* com escala de *razão*. Seu domínio compreende os números naturais.

- **Média da Polaridade:** O somatório das polaridades de todas as notícias publicadas em um determinado intervalo de tempo dividido pela quantidade de notícias publicadas. Este atributo é do tipo *numérico* com escala de *razão*. Seu domínio compreende os números reais.
- **Mediana da Polaridade:** A mediana das polaridades das notícias publicadas em um determinado intervalo de tempo. Este atributo é do tipo *numérico* com escala de *razão*. Seu domínio compreende os números naturais.
- **Desvio Padrão da Polaridade:** O desvio padrão da polaridade das notícias publicadas. Este atributo é do tipo *numérico* com escala de *razão*. Seu domínio compreende os números reais.
- **Variância da Polaridade:** A variância da polaridade das notícias publicadas em um determinado intervalo de tempo. Este atributo é do tipo *numérico* com escala de *razão*. Seu domínio compreende os números reais.
- **Somatório da Polaridade:** O somatório das polaridades de todas as notícia publicadas em um determinado intervalo de tempo. Este atributo é do tipo *numérico* com escala de *razão*. Seu domínio compreende os números naturais.
- Há quatro variáveis alvo que foram utilizadas conforme descrito na formalização presente na seção 2.3.

2. O termo **janelamento** que será descrito a seguir refere-se ao treinamento realizado com uma porção de dados limitada (janela) e ordenada no tempo. A hipótese neste tipo de treinamento é de que toda a informação necessária para compreender o cenário que pretende-se prever é concebida por eventos recentes, contemplados pelo tamanho da janela (15 minutos, 7 dias, 15 dias, 12 horas, e assim por diante) e que eventos anteriores ao tamanho da janela possuem pouca, ou nenhuma influência sobre o cenário de interesse. O termo **convencional** é utilizado aqui para descrever o procedimento comum de construção de modelos baseado em aprendizagem de máquina, onde, aleatoriamente são selecionadas porções dos dados para comporem as etapas de treino e teste. A hipótese dessa abordagem é de que algumas características são sempre preservadas independente da recência da informação sendo avaliada. Na tentativa de compreender o mercado acionário, ambas as interpretações são válidas. Se por um lado

é possível crer que notícias antigas não impactam mais no mercado acionário porque as suas informações já foram incorporadas aos preços das ações (encorajado, assim, o treinamento janelado), também é possível crer que os investidores do mercado acionário reagem sempre de forma similar as mesmas informações, encorajando, assim, o treinamento convencional. Em certa medida, ambas as análises foram verificadas.

3. O termo **acurácia** e **melhor acurácia** são definidos a seguir:

   Sejam:

   c : O número de classificações corretas que um modelo obteve ao realizar previsões.

   t : O total de classificações realizadas.

   z : O maior tamanho de janela possível no experimento.

   A **acurácia** que será discutida a seguir é definida na equação 7.1 a qual apresenta a função a que calcula a acurácia de um determinado modelo com tamanho de janela p.

$$a(p) = \frac{|\ c\ |}{|\ \frac{t}{p}\ |} \tag{7.1}$$

   A **melhor acurácia** é definida como o maior valor de acurácia obtido por um modelo ao variar todos os tamanhos de janelas possíveis, como apresentado pela equação 7.2.

$$max(a(1), \ldots, a(z)) \tag{7.2}$$

7.5 Planejamento dos Experimentos

Neste trabalho deseja-se investigar o impacto de notícias econômicas sendo publicadas na mídia nacional sobre a performance da BM&FBOVESPA. Como comentado na seção 2.1.3 o indicador IBOVE mede a performance média das empresas mais importantes e mais negociadas da BM&FBOVESPA e por isso é amplamente utilizado como forma de medir o desempenho do mercado de ações nacional. É possível analisar o IBOVE com base em várias outras características, dentre elas a *quantidade de negociações*, *quantidade de contratos negociados* e o *volume financeiro*. Ampliar o conjunto de alvos remete a outra perspectiva do problema onde, sendo, as variáveis preditivas insuficientes para preverem com boa acurácia a classificação futura dos valores do IBOVE, ainda assim, podem ser bastante precisas

em prever algumas das outras variáveis alvo e, por transitividade, acaba-se por compreender também como será o comportamento das outras variáveis alvo.

A definição e a formalização do problema multi-classe que envolve a previsão de variáveis relacionadas ao IBOVE encontra-se presente na seção 2.3 do capítulo 2.

Nesta seção são apresentados os experimentos que buscaram esclarecer hipóteses referentes a previsibilidade do índice Bovespa por meio de informações coletadas das notícias econômicas publicadas pelos jornais G1, Folha de São Paulo e Estadão.

Os modelos produzidos foram comparados com dois *baselines*. O primeiro deles é o modelo randômico para o qual todas as classes tem iguais chances de serem preditas. E, o modelo *Keep Trend* que prediz para o momento $t+1$ o que ocorreu no momento t. Em outras palavras, este modelo prevê que as tendências serão mantidas sejam elas de crescimento, queda ou consolidação.

Os dados coletados foram submetido a cinco classificadores estado-da-arte: Gaussian [Chan et al., 1982], Naive Bayes [Zhang, 2004], Decision Tree [Dumont et al., 2009], Random Forest [Breiman, 2001], Extra Trees [Geurts et al., 2006], Adaptive Boosting [Zhu et al., 2009] e Gradient Boosting [Friedman, 2001] em uma vasta diversidade de configurações. Para todos os classificadores considerados foram verificadas diferentes proporções de treino e teste: $60/40$, $70/30$ e $75/25$. A proporção que apresentou os melhores para todos os experimentos foi o de $70/30$. Até esse ponto procedeu-se de forma semelhante ao que foi executado para a análise setorial já explicada no capítulo 8

Um dos problemas desafiadores no desenvolvimento do protocolo experimental para a predição do IBOVE foi a definição da janela de tempo ideal para o funcionamento dos modelos de predição. Em um primeiro momento, todos os atributos envolvidos no experimento foram agregados por dia. Isto é, onde cada linha do arquivo de treino e teste representa a sumarização de todas as variáveis por dia. Em outras palavras, a quantidade de repercussões via Facebook, LinkedIn, a quantidade de notícias publicadas e os próprios valor do IBOVE foram todos agrupadas por dia. Seguindo essa configuração inicial, o melhor resultado de acurácia obtido foi de 44%, superando estatisticamente apenas o modelo randômico. Do ponto de vista prático, isso significa que mesmo após realizar inúmeras considerações sobre notícias e realizar inúmeros processamentos, a previsão do IBOVE diária via modelo de notícias não supera significativamente o *Keep Trend*. Porém, esse experimento apresenta

evidências da utilidade das notícias em relação ao modelo randômico e como elas podem ser utilizadas de modo a alavancar a qualidade das predições. Em um segundo momento, todos os atributos foram agrupados a cada 15 minutos e todo o processo foi novamente executado levando em consideração as hipóteses de treinamento por hora, dia e convencional explicadas no capítulo 8. Isso verificou se a mesma dinâmica de funcionamento em um nível setorial também poderia aplicar-se a um agrupamento mais genérico como é o caso dos índices de modo a melhorar a qualidade da predição. Neste momento, o arquivo de treino e teste ampliaram significativamente para 423.000 instâncias. De modo a calcular a variabilidade, todas as instâncias de notícias também foram divididas em 10 grupos cada grupo com 42.300 instâncias e submetidas simultaneamente a todos os modelos com as mesmas partições de treino e teste. A Figura 7.5 apresenta um exemplo de um grupo (uma porção) de instâncias do arquivo de treino submetida aos modelos que serão descritos a seguir. Cada um dos 10 grupos foi submetido para cada um dos classificadores.

Finalmente, os resultados obtidos com cada classificador foram verificados. Aplicou-se o teste-T de modo a avaliar as diferenças entre cada par de médias em todas as configurações. Todos os resultados foram estatisticamente significantes para $\alpha = 0.95$. A Tabela 7.4 sumariza a performance entre os *baselines* os métodos de melhor desempenho sendo comparados. Por sua vez, a Figura 7.7 apresenta em detalhes o desempenho de todos os métodos utilizados no experimento. A Figura 7.6 apresenta a matriz confusão do desempenho médio da previsão dos valores alvo para o Índice Bovespa. Percebe-se que de forma geral, todos os métodos, apesar de superarem o *baseline* são ainda ineficientes para determinarem janelas de consolidação. Por fim, nota-se que o enfoque proposto aqui superou significativamente tanto o modelo randômico quanto o *Keep Trend* para todos os alvos do IBOVE demonstrando assim o potencial da utilização de notícias na predição de eventos relacionados também ao IBOVE.

7.6 Discussão

Inicialmente é válido enfatizar ainda que em ambientes amplamente competitivos como os mercados financeiros, pequenas diferenças na capacidade preditiva entre modelos (aparentemente insignificantes em outros cenários) podem tornar-se amplamente lucrativos na prá-

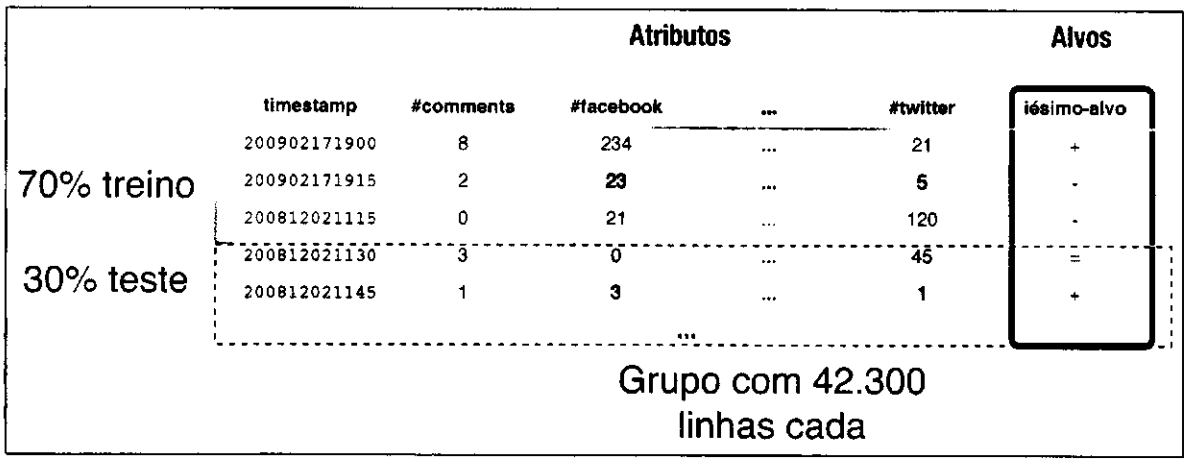

Figura 7.5: Exemplo de grupo de dados que foi dividido em treino e teste e submetido aos método de classificação.

tica. Também é válido mencionar que para o IBOVE os experimentos realizados apresentam evidências de sua eficiência em prever os próximos 15 minutos assim como os resultados encontrados para os setores. Esses modelos apresentam vantagem significativa em relação aos que o tempo de previsão supera os 15 minutos. O ponto negativo do modelo apresentado é sua necessidade por notícias econômicas. Isto é, sem notícias econômicas para serem analisadas, não há parâmetros de entrada para que a função possa gerar resultados. Isso não é frequente dado que em média cada um dos jornais analisados publicam $\sim$ 60 notícias econômicas diárias e grande parte disso em períodos de operação da bolsa.

7.7 Considerações Finais

A seguir são apresentadas algumas conclusões para este capítulo:

1. O Estadão é o jornal que apresenta o maior índice de correlação com o índice Bovespa enquanto o jornal G1 as menores correlações, apresentando resultados nulos ou próximos de nulidade.

2. A correlação medida entre o índice Bovespa e o compartilhamento de notícias nas redes sociais foi inversamente proporcional em todas as análises individuais de jornais. Ou seja, a medida que o valor do índice desce, o número de compartilhamentos de notícias econômicas aumenta (e vice-versa).

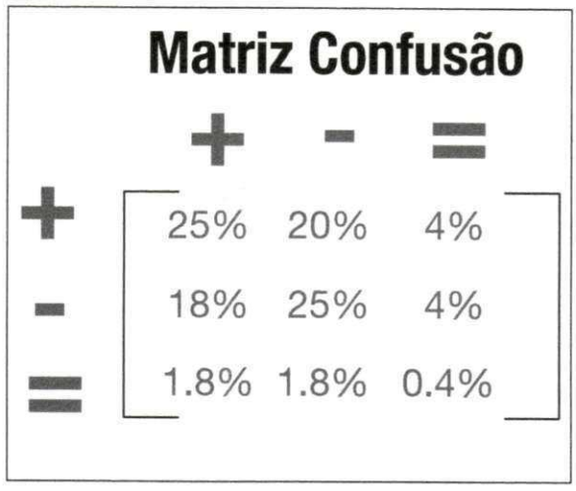

Figura 7.6: Matriz confusão dos desempenhos médios dos alvos relacionados ao IBOVE

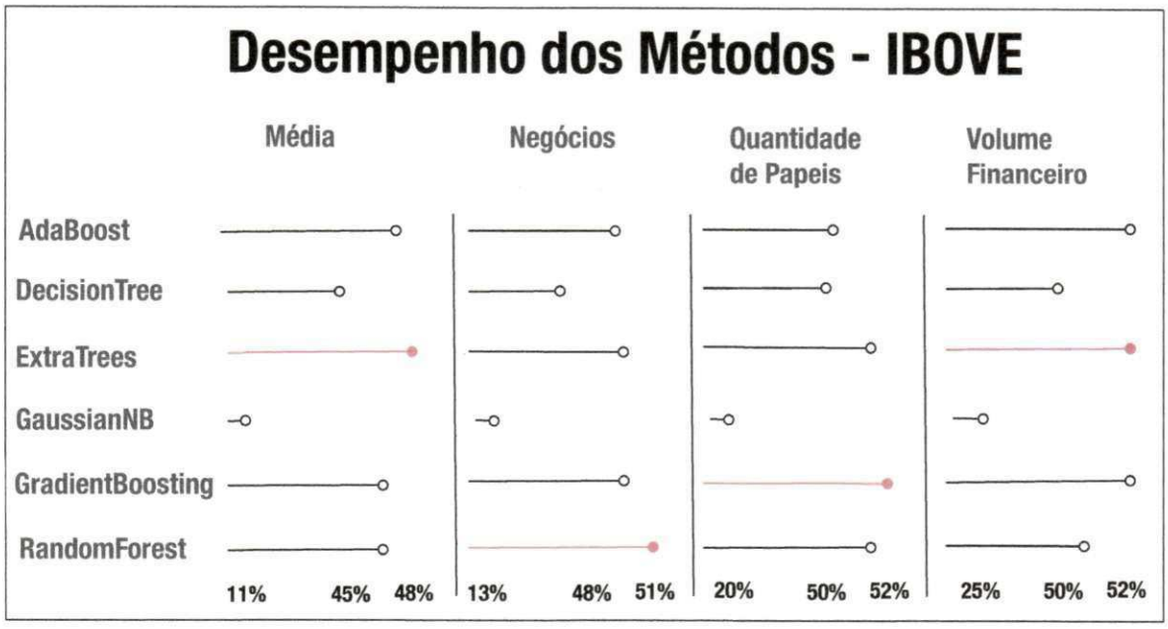

Figura 7.7: Desempenho detalhado dos métodos na predição das variáveis relacionadas ao IBOVE

Capítulo 8

Analise Setorial

Enquanto no capítulo anterior foi investigada a previsibilidade da BM&FBOVESPA em um grau mais amplo de granularidade por meio do índice Bovespa, aqui investigou-se o que acontece quando esta granularidade diminui, passando a considerar os diversos setores que compõem o índice.

De forma geral, a BM&FBOVESPA possui um mercado com centenas de ações sendo negociadas constantemente, para as quais, os preços são diretamente influenciados pela dinâmica de fatores econômicos como por exemplo, eventos domésticos e internacionais. Dessa forma, identificar boas ações e fazer bons negócios de forma sistemática é reconhecidamente uma tarefa complexa.

Uma maneira menos complexa de analisar o mercado de ações é desmembrá-lo em setores. Os setores apresentam em essência um grupo de ações que desempenham um conjunto de atividades semelhantes e atuam na mesma esfera ou ramo econômico. O desmembramento do mercado acionário em setores ajuda a posicionar melhor as companhias em torno do cenário econômico analisado, permitindo dessa forma, compreender melhor sobre riscos e benefícios das respectivas ações que estão associadas a eles. Apesar de não sabermos exatamente como cada uma das ações de um determinado setor irá se comportar no futuro, a análise setorial nos permite delimitar um panorama de possibilidades e ter bons direcionamentos sobre ele. Por exemplo, mesmo sem saber sobre a performance futura da empresa farmacêutica *Medley*, ou da empresa de alimentos *JBS*, é sabido que setores de alimentos e serviços médicos são menos sensíveis as condições econômicas que outros setores como Petróleo e Gás, uma vez que são fundamentais e de necessidade constante por parte da po-

pulação. Obviamente, essas afirmações sofrem ajustes mercado a mercado, porém, no geral, o funcionamento setorial acontece em grande medida em todos eles. Nessa perspectiva, entender a sensibilidade de diferentes setores do mercado financeiro em relação as notícias econômicas sendo publicadas é potencialmente útil para adquirir uma compreensão mais acurada do cenário econômico onde estes setores estão inseridos.

Neste capítulo serão descritas uma série de análises em nível setorial. Muitas delas já foram realizadas anteriormente para o índice BOVESPA, e aqui foram repetidas com dados em uma granularidade mais específica. Objetivamente busca-se responder ao longo desse capítulo a seguinte pergunta: *De que forma o desempenho dos setores da BM&FBOVESPA é afetado pelas notícias econômicas publicadas nos jornais de alta circulação no Brasil ?* A hipótese planteada é de que quanto maior for a variação de desempenho em um determinado setor face a eventos econômicos publicados via notícias econômicas maior será a influencia de notícia em impactá-lo e consequentemente sua sensibilidade em relação a elas. Desse modo, os resultados obtidos por meio dos experimentos realizados abrem novas perspectivas sobre a utilização de notícias em modelos de previsão a serem desenvolvidos, complementando técnicas já existentes ou seguindo novos enfoques cada vez mais fundamentalistas.

A definição e a formalização do problema multi-classe que envolve a previsão de variáveis relacionadas ao setores encontra-se presente na seção 2.3 do capítulo 2.

8.1 Preparação e Análise

Nesta seção serão descritos todas as etapas de preparação de dados e os protocolos experimentais que foram utilizados.

8.1.1 Setores

Todas as ações negociadas na BM&FBOVESPA pertencem a algum dos seguintes setores: Construção e Transporte, Consumo Cíclico, Consumo não Cíclico, Utilidade Pública, Bens Industriais, Financeiro e Outros, Materiais Básicos, Telecomunicações, Petróleo Gás e Biocombustíveis, Tecnologia da Informação e Não Classificados. Entretanto, nenhuma ação dos setores de Tecnologia da Informação e Não Classificados compõe o índice Bovespa, que sumariza as empresas mais negociadas no mercado. Na prática isso significa que a quanti-

dade de negociações praticada para essas empresas é bastante tímido ou com pouquíssima relevância. Dessa forma, esses setores não foram incluídos nas análise subsequentes.

Os dados de todas as empresas e suas variações a cada 15 minutos foram generosamente concedidos sem quaisquer custos pela equipe que compõe Observatório da Web na Universidade Federal de Minas Gerais [43].

8.1.2 Análise de Autocorrelação

A autocorrelação consiste na análise de correlação de uma série temporal com ela mesma em diferentes pontos no tempo. Essa técnica apresenta-se como uma ferramenta matemática hábil em detectar padrões de repetição apesar de, nem sempre, ser suficiente para explicá-los.

Vale a pena enfatizar que há inúmeros tamanhos de janelas para os quais os dados podem ser agrupados para a realização de análises sobre eles (de segundos a anos). Entretanto, com recursos e tempo limitados para explorar todas essas possibilidades a análise de autocorrelação foi utilizada para lançar alguma luz sobre os tamanhos de janelas de tempo potencialmente úteis, para as quais, cada um dos alvos apresentava uma repetição sistemática.

A análise de autocorrelação foi construída para todas as series temporais de todas as variáveis alvo para todos os setores considerados. Para isso, fixou-se a serie temporal original e calculou-se a correlação entre a série original com ela mesma para um movimento de janela em múltiplos de 15 minutos, 30 minutos, 45 minutos, e assim por diante. A análise de autocorrelação foi bastante útil para fundamentar a escolha da granularidade de tempo com maior potencial de predição que no caso foi a janela de 15 minutos.

A Figura 8.1 apresenta o gráfico de autocorrelação para todas variáveis alvo em todos os setores. O eixo horizontal apresenta os tamanhos de janela múltiplos de 15 minutos. Isto é, o valor 1, um tamanho de janela de 15 minutos, o valor 2, 30 minutos, e assim sucessivamente. Já o eixo vertical apresenta os resultados da autocorrelação entre a série temporal sem deslocamentos no tempo e a série com deslocamentos no tempo proporcionais ao tamanho de janela. Ou seja, o par ordenado (12, 0.8), por exemplo, indica que ao deslocar a série temporal 12 janelas de 15 minutos, isto é, 3 horas, a correlação entre a série sem alteração e a série com deslocamento são 80% similares. É possível verificar que o melhor

[43] http://observatorio.inweb.org.br/

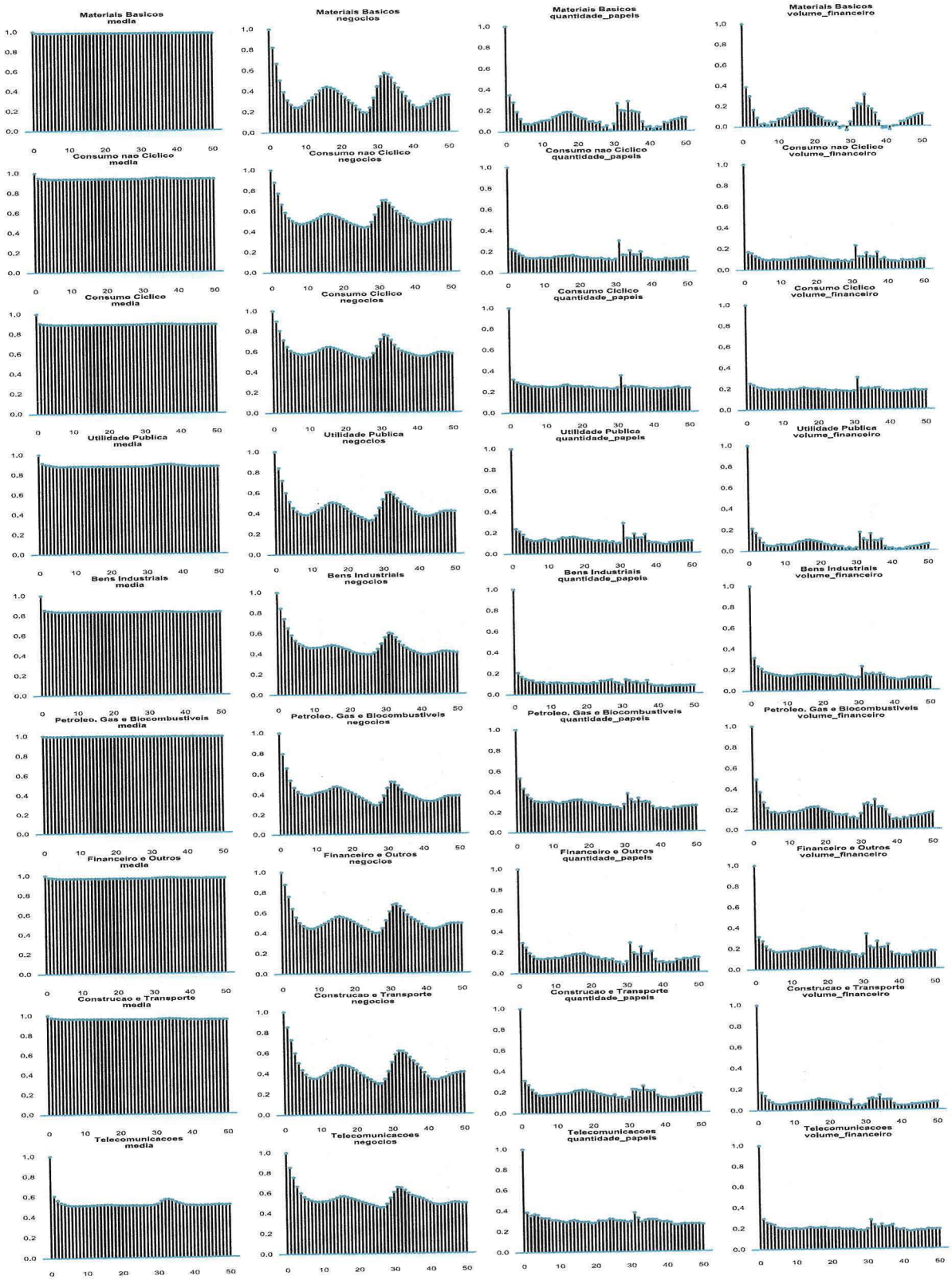

Figura 8.1: Resultado da análise de correlação entre todos os alvos e setores.

valor de correlação em todos os cenários ocorre quando a janela de tempo é ajustada para os próximos 15 minutos.

Após a análise de autocorrelação o próximo passo foi agrupar as características extraídas das notícias em 15 minutos e verificar se elas são capazes de explicar alguma parcela dos padrões temporais que foram encontrados.

8.1.3 Análise de Sensibilidade

De modo a investigar a sensibilidade de cada setor da BM&FBOVESPA às notícias econômicas mediu-se a correlação entre cada atributo relacionado às notícias e cada variável alvo para cada um dos setores considerados em três janela de tempo (15 minutos, 1 hora e 1 dia).

Apesar de a análise de autocorrelação indicar o período de 15 minutos como provável janela ideal para as demais análises, o fato é que na prática alguns *day-traders* e *swing-traders* de varejo para apoiarem suas decisões em análise fundamentalista de notícias investem um tempo superior a 15 minutos para se informarem sobre fatos econômicos. De modo a tentar englobar parte desses perfis, além da janela de tempo de 15 minutos, também foram incorporadas as janelas de tempo de 1 hora e 1 dia.

A seguir introduzimos formalmente a métrica de sensibilidade usada nesse trabalho.

Seja:

M : o conjunto de preditores extraídos das notícias econômicas.

N : o conjunto de variáveis alvo.

T : o conjunto de diferentes janelas de tempo para as quais o teste foi realizado.

S : o conjunto de setores da BM&FBOVESPA.

$C = N \times M = \{(n, m) | n \in N \wedge m \in M\}$: o conjunto de todas as combinações par-a-par entre variáveis preditoras e variáveis alvo para uma dada janela de tempo específica.

A sensibilidade de um setor $s \in S$ em uma janela de tempo $t \in T$ é definida como:

$$sens(s)_t = \frac{\sum_{c \in C} (\tau_{c,s})_t^2}{|C|} \tag{8.1}$$

A equação 8.1 define a medida de sensibilidade como sendo a soma de todas as correlações de Kendall de todas as combinações entre as variáveis alvo e as variáveis preditoras para os três tamanhos de janelas: 15 minutos, 1 hora e 1 dia. Como já foi explicado an-

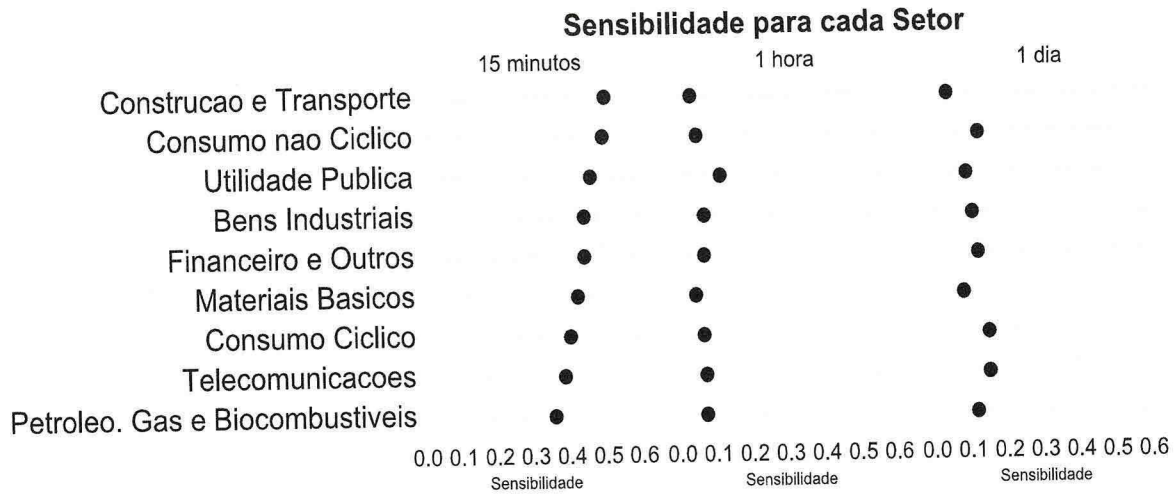

Figura 8.2: Resultado da sensibilidade medida para todos os setores em 15 minutos 1 hora e 1 dia.

teriormente, a correlação de Kendall foi utilizada porque ao contrário das correlações de Pearson e Spearman ela não é afetada pela magnitude dos valores ou pela organização do ranqueamento.

Agora, cada observação passou a conter todas as variáveis preditoras, extraídas das notícias, agregadas segundo o período determinado pela janela de tempo selecionada. Por exemplo, a variável *quantidade de repercussão no Facebook* para a janela de tempo de 15 minutos passou a conter a soma de todos os compartilhamentos de todas as notícias publicadas para cada 15 minutos de janela, e assim sucessivamente. As Figuras 8.3, 8.4, 8.5 apresentam os resultados dos cálculos da correlação entre variáveis relacionadas das notícias e variáveis alvos para todos os setores para as janelas de tempo de 15 minutos, 1 hora e 1 dia respectivamente. Em todas as figuras, as variáveis do eixo vertical representam as variáveis extraídas das notícias, enquanto que as do eixo horizontal representam as variáveis alvos, relacionadas ao setores. Os prefixos *dp,m,me,s* e *v* representam respectivamente o desvio padrão, média, mediana e os valores da soma absoluta dos valores sendo pedidos. Os sufixos *Pol, Com, Fb, Gp, Lk* e *Tw* correspondem respectivamente a polaridade, comentários, Facebook, Google Plus, LinkedIn e Twitter. A abreviação a *qt* refere-se a quantidade. Como exemplo, a variável *dpQtLk* traduz-se como o desvio padrão da quantidade de compartilhamentos via LinkedIn. Em todas as imagens, cada célula contém a cor proporcional ao cáclulo da correlação de Kenddal entre uma variável extraídas das notícias e uma variável alvo, extraída

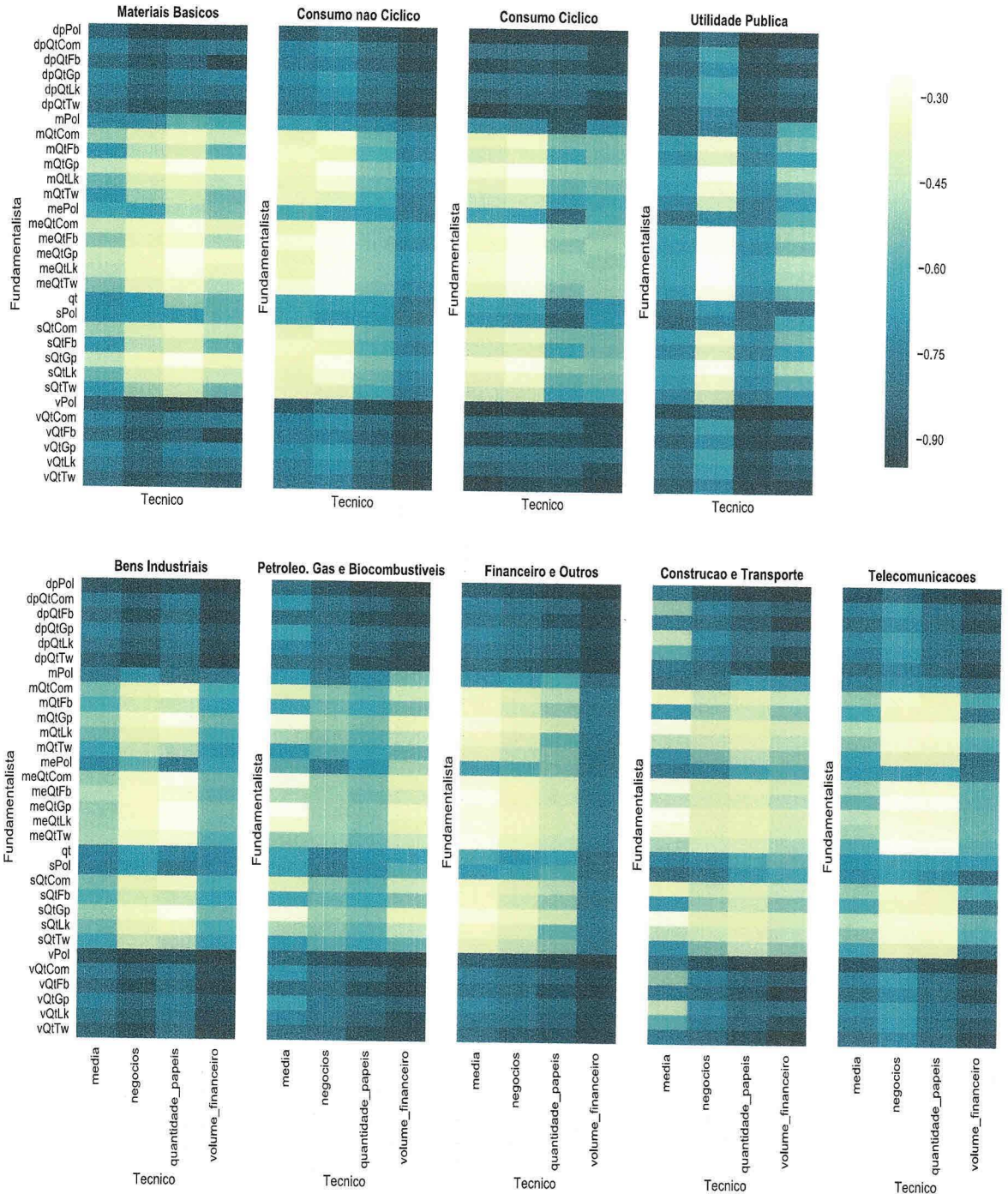

Figura 8.3: Cálculo da correlação de Kendall entre variáveis preditivas extraídas das notícias (eixo vertical) e variáveis alvo (eixo horizontal) para janela temporal de 15 minutos.

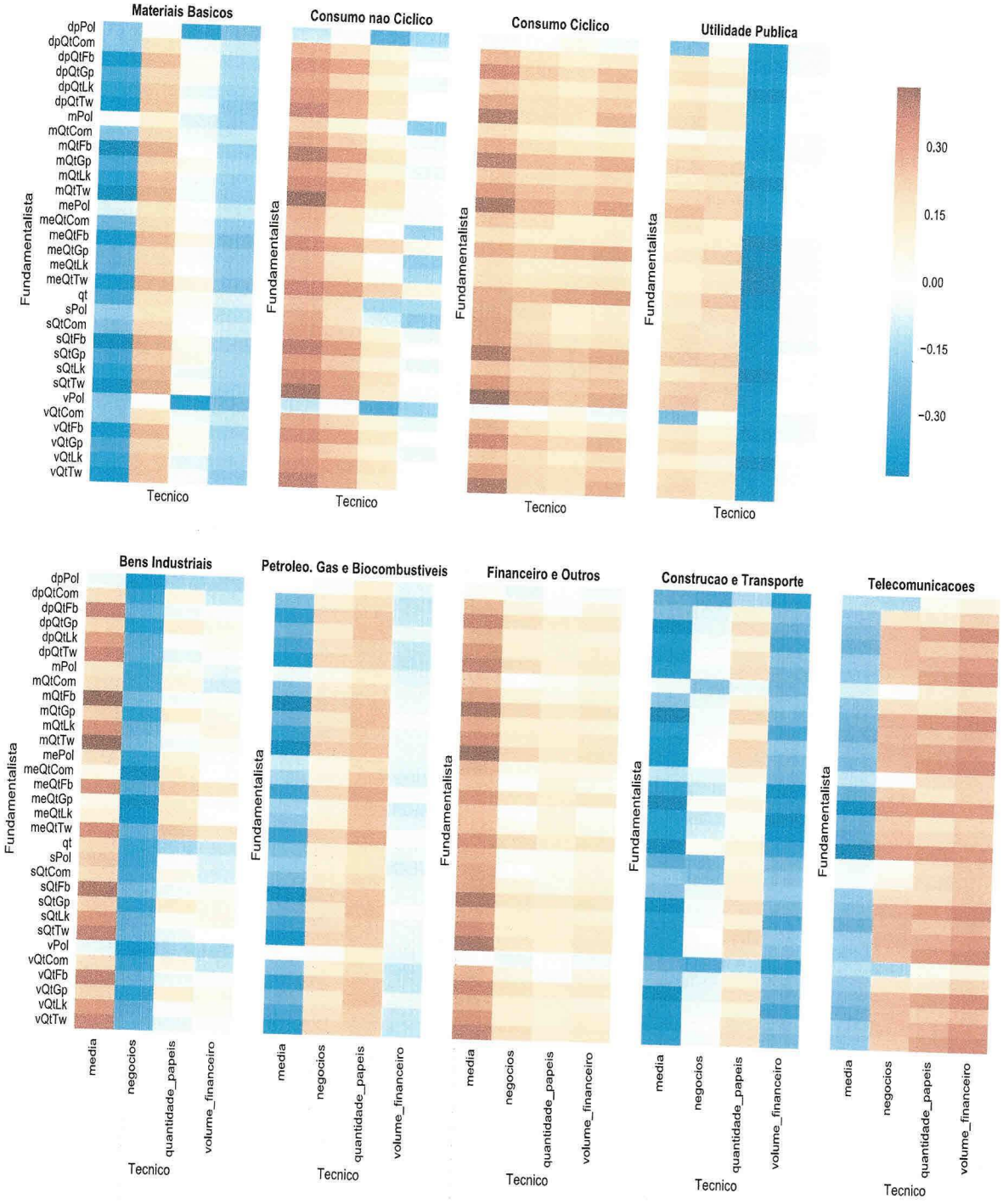

Figura 8.4: Cálculo da correlação de Kendall entre variáveis preditivas extraídas das notícias (eixo vertical) e variáveis alvo (eixo X) para janela temporal de 1 hora.

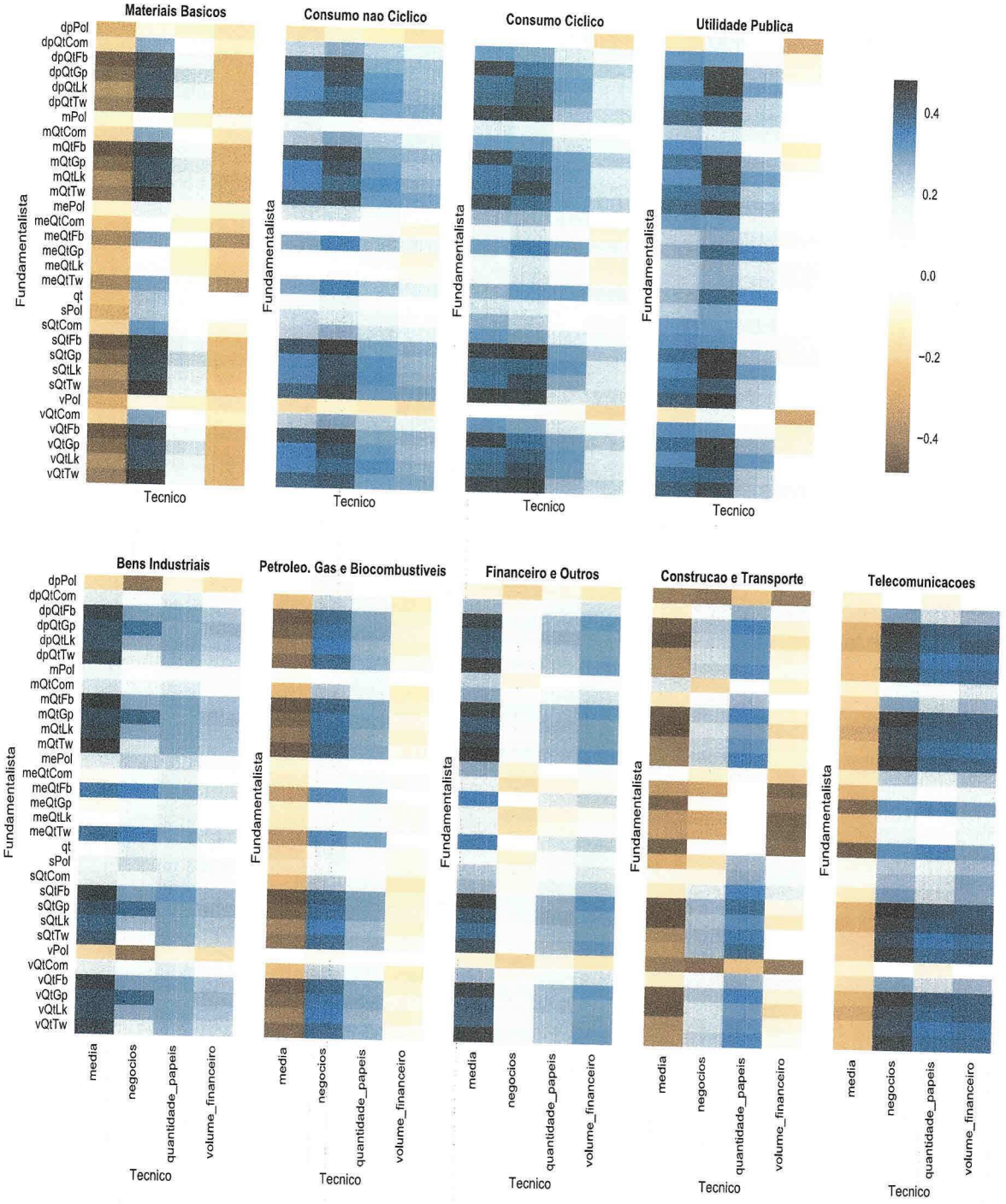

Figura 8.5: Cálculo da correlação de Kendall entre variáveis preditivas extraídas das notícias (eixo Y) e variáveis alvo (eixo horizontal) para janela temporal de 1 dia.

tecnicamente de cada setor sendo analisado.

A Figura 8.2 apresenta o resultado do cálculo da sensibilidade para todos os setores. Pode-se perceber, inicialmente, que os valores mais altos para a medida de sensibilidade foram encontrados para a janela de tempo de 15 minutos o que o torna bastante consistente com a análise de autocorrelação descrita anteriormente. O setor de *Construção e Transporte* ultrapassou os 50% de sensibilidade e foi o setor com maior sensibilidade

De maneira geral, a análise de sensibilidade tanto para 1 hora quanto 1 dia não apresentou resultados promissores. Em ambos os casos nenhum setor ultrapassou os 15% de sensibilidade.

A análise de correlação em conjunto com a análise de sensibilidade apresentaram evidências para decidir sobre a janela de tempo ideal para considerar na predição da performance. É digno de ênfase que no mundo real, eficiência e eficácia são vetores paralelos e que quanto mais rápida ocorra a predição maiores são as vantagens competitivas que apoiam a decisão do investidor. Desse modo, modelos que predição que executam a cada 15 minutos proveem 45 minutos de vantagem competitiva se comparados com modelos que executam a cada hora, por exemplo.

8.2 Experimentos e Resultados

Nesta seção são apresentados em detalhes os experimentos realizados para compor os modelos de predição assim como as configurações ideais que foram estabelecidas.

8.2.1 Modelos de Predição

Cada instância de treino, isto é, cada linha do conjunto de dados corresponde ao agrupamento de informações para cada variável considerada, seja ela preditiva ou variável alvo, que ocorreram em um intervalo de 15 minutos. A lista dessas variáveis podem ser vistas na seção 7.4. Todas as linhas contendo valores faltantes (N/A) em ao menos uma das variáveis consideradas foram removidas. Há dois pontos importantes decorrentes dessas remoções que merecem ser considerados. O primeiro deles é a porcentagem de linhas removidas. Aproximadamente 25% do total dos intervalos gerados com a janela de tempo de 15 minutos ou não possuem notícias ou possuem uma única notícia publicada. Em ambos os casos, todas

as variáveis relacionadas a desvio-padrão e variância ficam impossibilitadas de serem calculadas e são então valoradas com N/A, o que gera sua posterior remoção. O segundo ponto é que seria possível utilizar as janelas de 15 minutos que apresentaram apenas 1 notícia por intervalo. De certa forma, esta abordagem além de diminuir o percentual de descartes, acabaria por aproveitar melhor as instâncias de treino disponíveis passando a ser útil no processo de aprendizagem. Entretanto, a observação do gráfico de correlações presente na Figura 8.3 apresenta valores de correlação bastante altos para as variáveis preditivas relativas tanto ao desvio-padrão quanto variância. Dessa forma, o *trade-off* que envolve essa temática priorizou a qualidade das correlações encontradas por meio do cálculo da variabilidade das variáveis preditivas ao invés de maximizar o aproveitamento da base de dados disponíveis.

Foram comparados três abordagens principais: *O estado-da-arte de classificação*, *Random Model* e o *Keep Trend*. A abordagem *Keep Trend* assume que o mercado manterá a tendência atual para os próximos 15 minutos. Isto é, dada a situação atual de algum alvo para algum setor considerado no estudo, o seu comportamento para os próximos 15 minutos será o mesmo que o seu estado atual. Por sua vez, o Random Model, seleciona aleatoriamente uma das três possibilidades de classificação possíveis e considera o resultado obtido como o valor da previsão para os próximos 15 minutos futuros.

Os modelos de aprendizagem de máquina foram treinados sob três configurações. A primeira delas considerou todo o histórico de dados disponíveis apenas particionando-a em treino e teste em diferentes porções. A segunda estratégia também usou apenas o histórico de dados porém o treino e o teste foram realizados apenas entre horas específicas do dia. Por exemplo, para predizer o que acontece às 10:00h foram utilizadas apenas as informações do histórico de notícias que fora publicadas entre 10:00h e 10:59h. Em essência, essa hipótese pressupõe que cada variável alvo relacionada aos setores da BM&FBOVESPA apresentarão performance similar durante as mesmas horas de operação de mercado. Isto é, enquanto os horários de aberta, em grande medida, possuem dinâmica de funcionamento semelhante entre si, o mesmo acontece com os horários de fechamento. Nessa perspectiva, soa coerente a construção de um modelo que leva em consideração o horário de operação para realizar a previsão. A terceira configuração é análoga a segunda, porém, amplia a perspectiva de hora para uma perspectiva de mês, dado que alguns eventos econômicos importantes estão presentes no calendário econômico anual. Por exemplo, Janeiro é o mês onde grande parte

dos brasileiros acertam suas contas com a Receita Federal e, assim sendo, é esperado que o volume de negociações no mercado acionário possa ser afetado de alguma forma. Além disso, como já mencionado anteriormente, Agosto é o mês de apresentação do novo plano econômico anual para o país e isso impacta de várias formas as decisões dos acionistas, entre outros. Assim sendo, segundo essa perspectiva, para predizer a performance de setores em um determinado mês foram utilizadas informações históricas deste mês específico.

Após agrupar todas as variáveis em períodos de 15 minutos dividiu-se o conjunto de dados (423.000 instâncias) em 10 subgrupos contendo cada um deles $42,300$ todas elas ordenadas no tempo. Isso foi feito com o objetivo de medir a variabilidade de todos resultados para os métodos avaliados. A Figura 7.5, apresentada no capítulo anterior, apresenta um exemplo de organização de treino e teste para um único grupo de instâncias.

Cada subgrupo de dados foi submetido a cinco classificadores estado-da-arte: Gaussian [Chan et al., 1982], Naive Bayes [Zhang, 2004], Decision Tree [Dumont et al., 2009], Random Forest [Breiman, 2001], Extra Trees [Geurts et al., 2006], Adaptive Boosting [Zhu et al., 2009] e Gradient Boosting [Friedman, 2001] em uma vasta diversidade de configurações.

Para todos os classificadores considerados foram verificadas diferentes proporções de treino e teste: $60/40$, $70/30$ e $75/25$. A proporção que apresentou os melhores para todos os experimentos foi o de $70/30$.

Para alguns dos métodos utilizados no experimento como: Random Forest, Extra Trees, Adaptive Boosting and Gradient Boosting o número de árvores configurados inicialmente está diretamente relacionado com a performance obtida. Sendo assim, foi executado um *grid search* com validação cruzada de 5 *folds* para as seguintes quantidades de árvores: $100, 150, 500$ e $1,000$. Em seguida, o resultado foi comparado com todos os modelos considerados e os melhores resultados obtidos ocorreram com 150 árvores de inicio.

A Figura 8.6 apresenta o fluxo de todo o processo que foi desenvolvido desde a geração do arquivo contendo todas as variáveis coletadas a cada 15 minutos até os resultados finais comparando os métodos utilizados. A Figura 8.7 detalha de forma geral os resultados obtidos para todos os métodos sendo comparados para a abordagem baseada em hora de operação da BM&FBOVESPA, o qual apresentou os melhores resultados quando comparados com as demais abordagens. É importante enfatizar que as classes estavam relativamente bem balanceadas ao longo de experimento de modo a evitar o viés de classificação em direção a

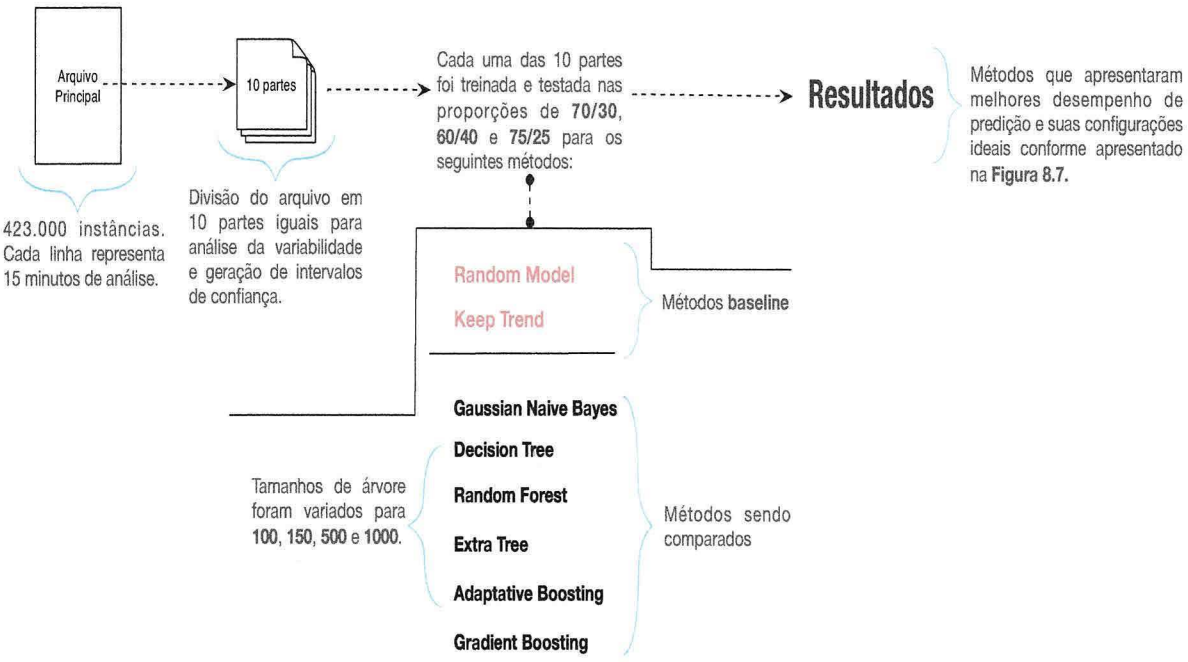

Figura 8.6: Fluxo do processo desenvolvolvido entre a geração do arquivo contendo as informações das variáveis até o resultado final de comparação entre os métodos.

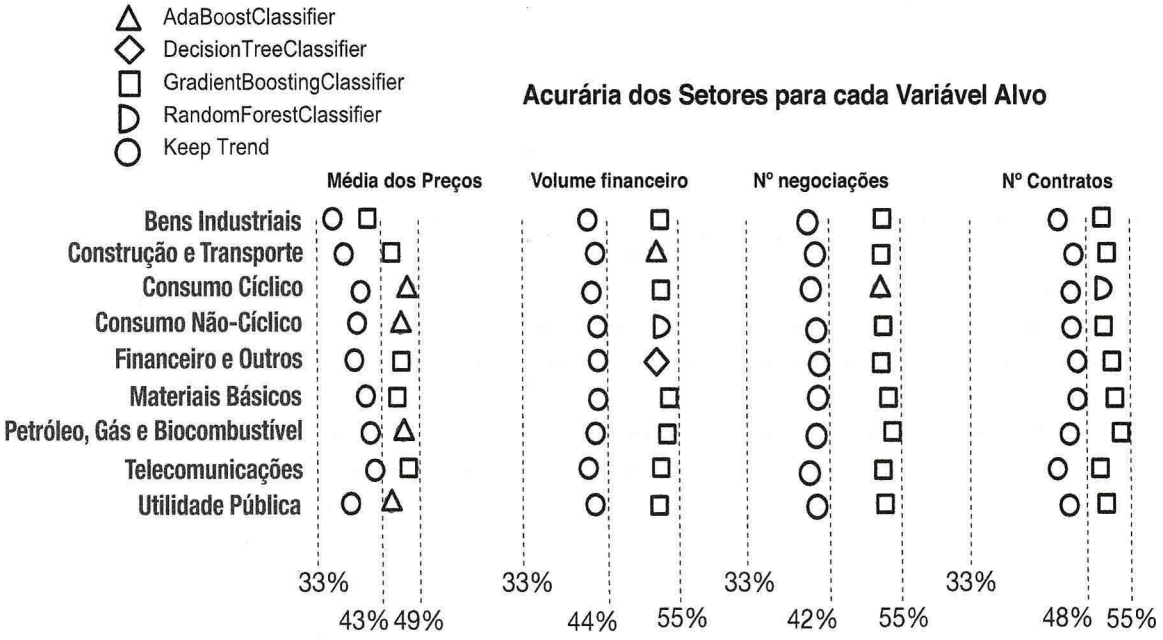

Figura 8.7: Comparação entre a acurácia obtida pelo modelo desenvolvido e demais modelos sendo comparados.

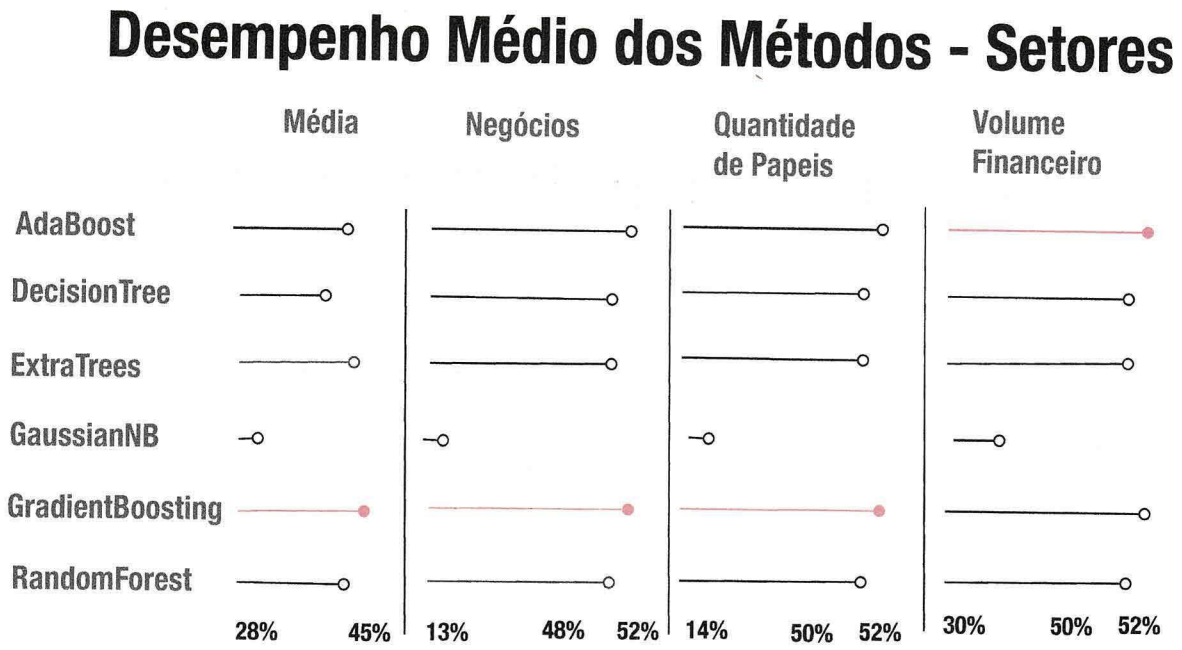

Figura 8.8: Detalhamento do desempenho médio de todos os métodos utilizados por alvo entre todos os setores.

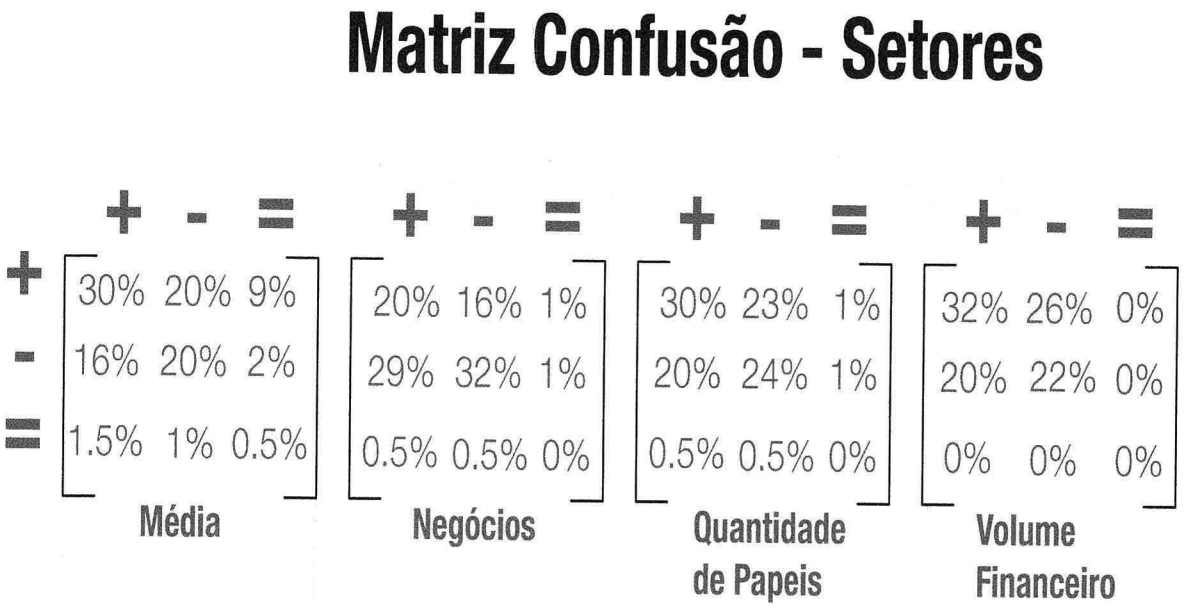

Figura 8.9: Matriz Confusão das médias de proporções entre todos os setores.

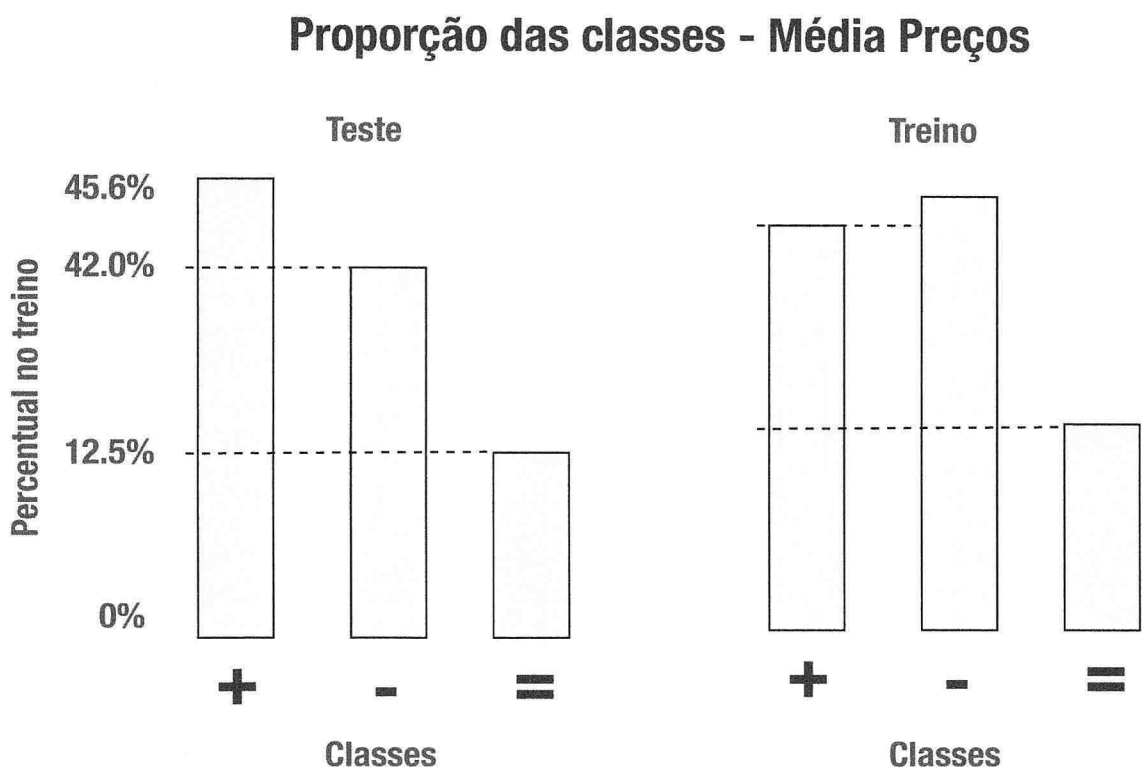

Figura 8.10: Proporção entre treino e teste para o alvo de preço médio entre os setores.

classe majoritária conforme o exemplo apresentado pela Figura 8.10. Aplicou-se o teste-T de modo a avaliar as significâncias das diferenças encontradas entre cada um dos enfoques propostos. Todos os resultados foram estatisticamente significativos para $\alpha = 0.95$. A Figura 8.8 apresenta detalhes sobre o desempenho médio de todos os classificadores utilizados na predição de cada um dos alvos de interesse complementando a Figura 8.7 na qual apenas os métodos de maior desempenho não confrontados contra o *baseline*. Por fim, a Figura 8.9 apresenta a matriz confusão média dos resultados obtidos para cada alvo.

Ao observar os resultados é possível notar que o setor de *Petróleo, Gás e Biocombustível* obteve os melhores resultados de predição. Uma possível explicação para o fato é de que este setor contempla a Petrobrás, uma empresa estatal amplamente impactada tanto pelo cenário econômico quanto pelo cenário político nacional e constantemente presente na mídia sendo acompanhada de perto não apenas por investidores como por todos os brasileiros. Nesta perspectiva, qualquer informação seja ela positiva ou negativa acaba por impactar não apenas o setor mas a economia nacional como um todo. Além disso, de forma efetiva, foi observado empiricamente que o termo Petrobrás está presente em uma fração bastante expressiva de

todas as notícias coletadas. Por outro lado, o setor de *Bens Industriais* apresentou os menores valores de previsibilidade. É provável que na prática os jornais analisados contenham poucas informações relacionadas à esses setores. Uma outra possível explicação é notícias relacionadas ao setor de *Bens Industriais* são produzidas em decorrência de acontecimentos no mercado acionário, sendo assim, há uma grande correlação, como verificado na análise de sensibilidade, porém, com baixíssimo poder de predição. E exatamente o contrário acontece com as notícias econômicas referentes ao setor de *Petróleo Gás e Biocombustível*, onde as notícias tendem a comentar sobre cenários futuros ao invés de meramente registrar acontecimentos passados, e dessa forma, todas as variáveis preditoras passam a ser úteis na atividade de prever os próximos movimentos desse setor.

8.3 Deterioração da Predição ao longo do Tempo

Apesar dos resultados encontrados demonstrarem superioridade entre a abordagem sendo proposta e os demais abordagens com os quais foram comparadas houve a necessidade de saber até que ponto a predição se mantém estável sem o recebimento de novas informações. Dito de outra maneira: *Em caso de a coleta de novas notícias ser de alguma forma impossibilitada, durante quanto tempo é possível realizar predições de modo a manter os resultados alcançados até então ?*

Para responder a esse questionamento, foram conduzidos experimentos que visaram medir a deterioração da predição no tempo. Inicialmente, selecionou-se o primeiro ano de dados (2008) para ser divido em duas partes sendo 70% para treino e 30% para teste, de forma similar ao que foi feito na seção anterior. A Figura 8.11 ilustra o processo que foi realizado para realizar o cálculo da deterioração da predição ao longo do tempo. A primeira instância de teste está posicionada 15 minutos após a última instância de treino, a segunda instância de teste está 30 minutos distante, e assim por diante. Este processo foi repetido várias vezes deslizando treino e teste em 30% das instâncias até cobrir toda a base de dados. Isso foi necessário para evitar que o treinamento fosse realizado sempre com os mesmos meses. Ao treinar com os primeiros 70% do ano e testar com os 30% finais o treino sempre é realizado de Janeiro à Setembro e o teste de Outubro à Dezembro. Para evitar esse problema, tanto o treino quanto o teste foram deslocados 30% à frente na linha do tempo, tanto para oportuni-

zar que o processo de treino e teste passasse por todos os meses ao longo dos anos, quanto para evitar que o teste fosse realizado com parcelas já testadas anteriormente.

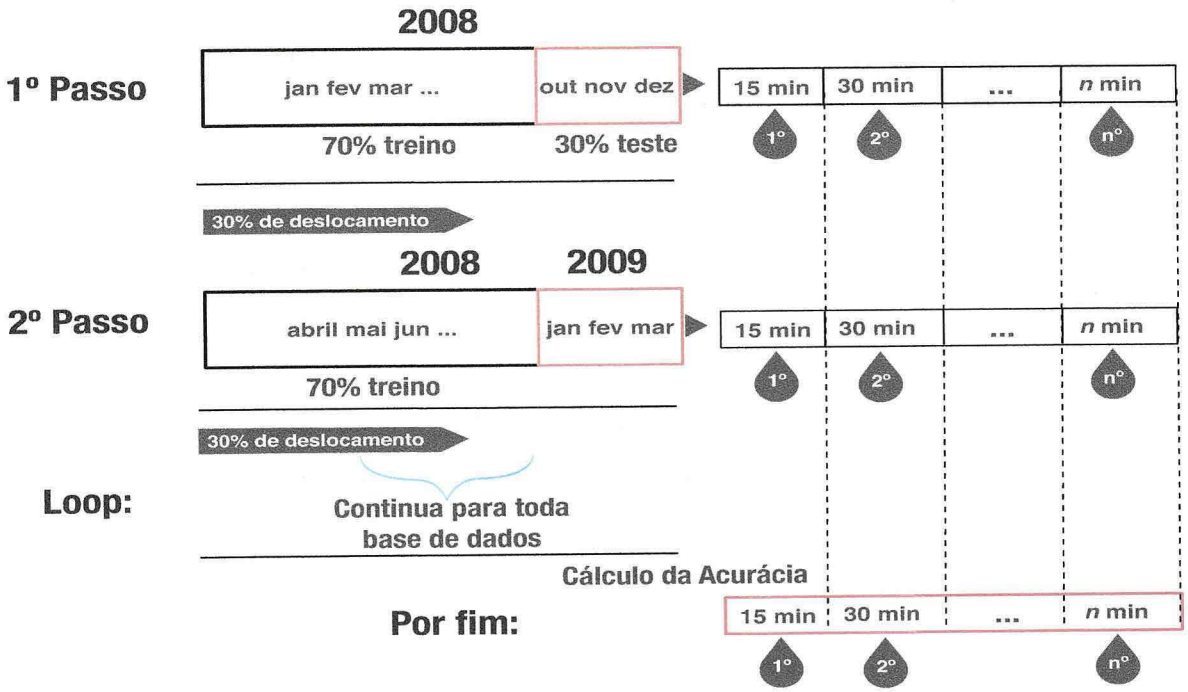

Figura 8.11: Processo de cálculo da deterioração da predição ao longo do tempo.

O modelo de treinamento utilizado foi o Classificador *Gradient Boosting* com 150 árvores dado que foi o modelo que apresentou os melhores resultados de predição para a maioria dos alvos em todos os setores analisados. A Figura 8.12 apresenta os resultados deste experimento. Pode-se observar que para todos os setores, em todas as variáveis alvos consideradas, a acurácia não ficou abaixo dos 40% durante 24 dias seguidos. Após isso, os resultados oscilaram drasticamente até eventualmente sofrerem uma queda drástica.

8.4 Qualidade da Predição sobre a quantidade de Notícias

Uma última questão que foi levantada sobre a análise de setores foi de que maneira se comporta da qualidade da predição diante a quantidade de notícias que ocorreram em um intervalo de 15 minutos. Obviamente, não é comum que a cada 15 minutos existam inúmeros fatos econômicos que precisem ser publicados, além disso, é notória a necessidade de redação e escrita e correção que normalmente ultrapassam o período de 15 minutos para serem publicados.

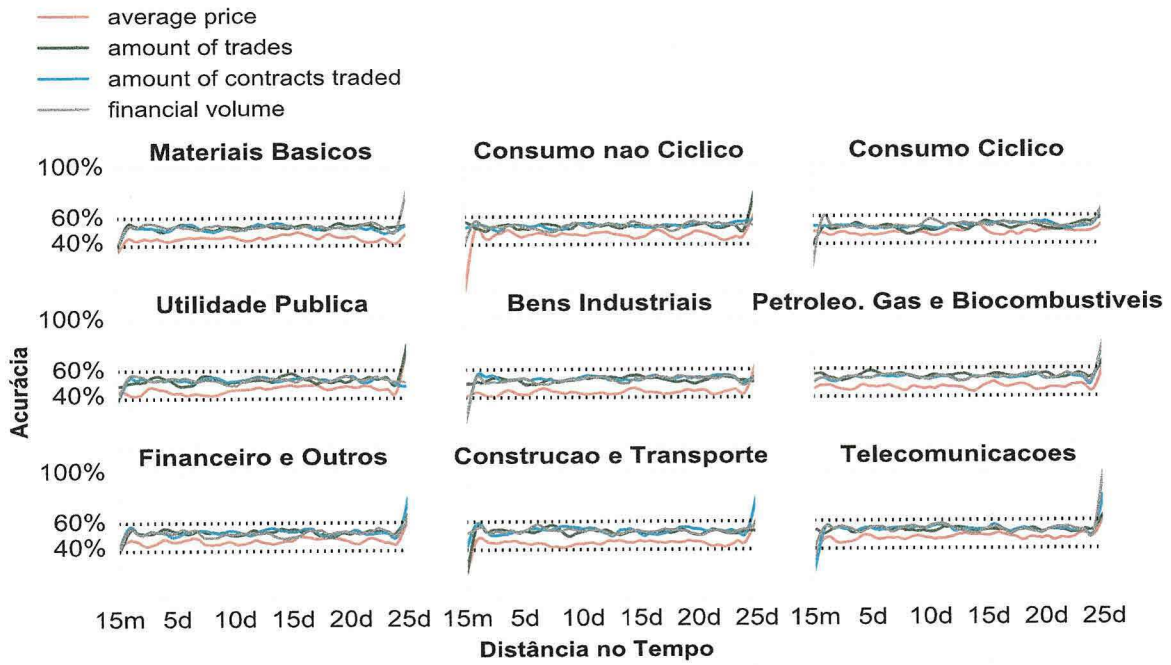

Figura 8.12: Deterioração da predição ao longo do tempo para todos os setores analisados.

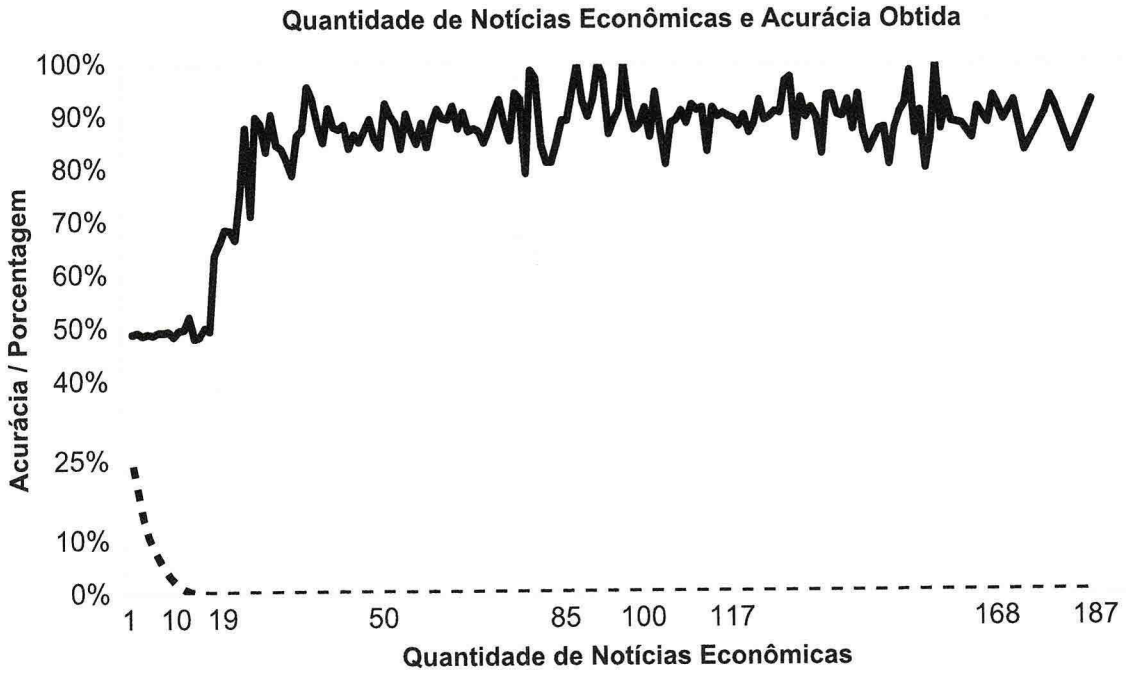

Figura 8.13: Qualidade da predição pela quantidade de notícias publicadas no período de 15 minutos.

Este experimento investigou como as predições são afetadas pela quantidade de notícias em cada instância de 15 minutos. É importante notar que cada instância de 15 minutos contém o agrupamento de todas as notícias publicadas neste intervalo, dessa forma, algumas instâncias podem incluir mais notícias que outras. Inicialmente, todas as instâncias foram agrupadas em relação a quantidade de notícias que foram publicadas naquele intervalo específico. Foram encontrados 171 grupos que variam de instâncias que possuem entre 2 até 187 notícias econômicas. Em seguida, o classificador Gradient Boosting em sua configuração mais exitosa, com a partição de treino e teste de 70/30 foi executado para cada um dos grupos tendo como resultado final a acurácia para cada grupo. A Figura 8.13 apresenta os resultados obtidos entre acurácia e quantidade de notícias. A curva abaixo representa a frequência da quantidade de notícias para janelas de 15 minutos. A curva acima representa o valor de acurácia pelo número de notícias. Percebe-se que quando a quantidade de notícias varia entre 2 até 10 isso é suficiente para que a acurácia alcance $\sim 50\%$ (o que é melhor que o modelo randômico). Entretanto, acima de 19 notícias econômicas há um grande incremento da acurácia superando os 70% em todos os cenários.

8.5 Seleção de Características

Este experimento buscou verificar dentre todo o conjunto de variáveis sendo utilizado quais delas apresentavam maior capacidade discriminativa entre as classes. A Figura 8.14 apresenta a lista de todas as variáveis ordenadas por sua escala de importância. Percebe-se que de forma geral, as variáveis relacionadas com a ferramenta Twitter apesentam, em grande medida, grande capacidade discriminativa em relação à demais variáveis e que ferramentas como quantidade de comentários, LinkedIn e GooglePlus contribui de forma muito tímida para compor o resultado final.

8.6 Considerações Finais

Este capítulo se propôs a explorar a relação entre notícias econômicas e a BM&FBOVESPA em um nível mais fino de granularidade considerando os diversos setores que a compõe. Em grande medida foi possível concluir que as notícias econômicas publicadas em jornais de

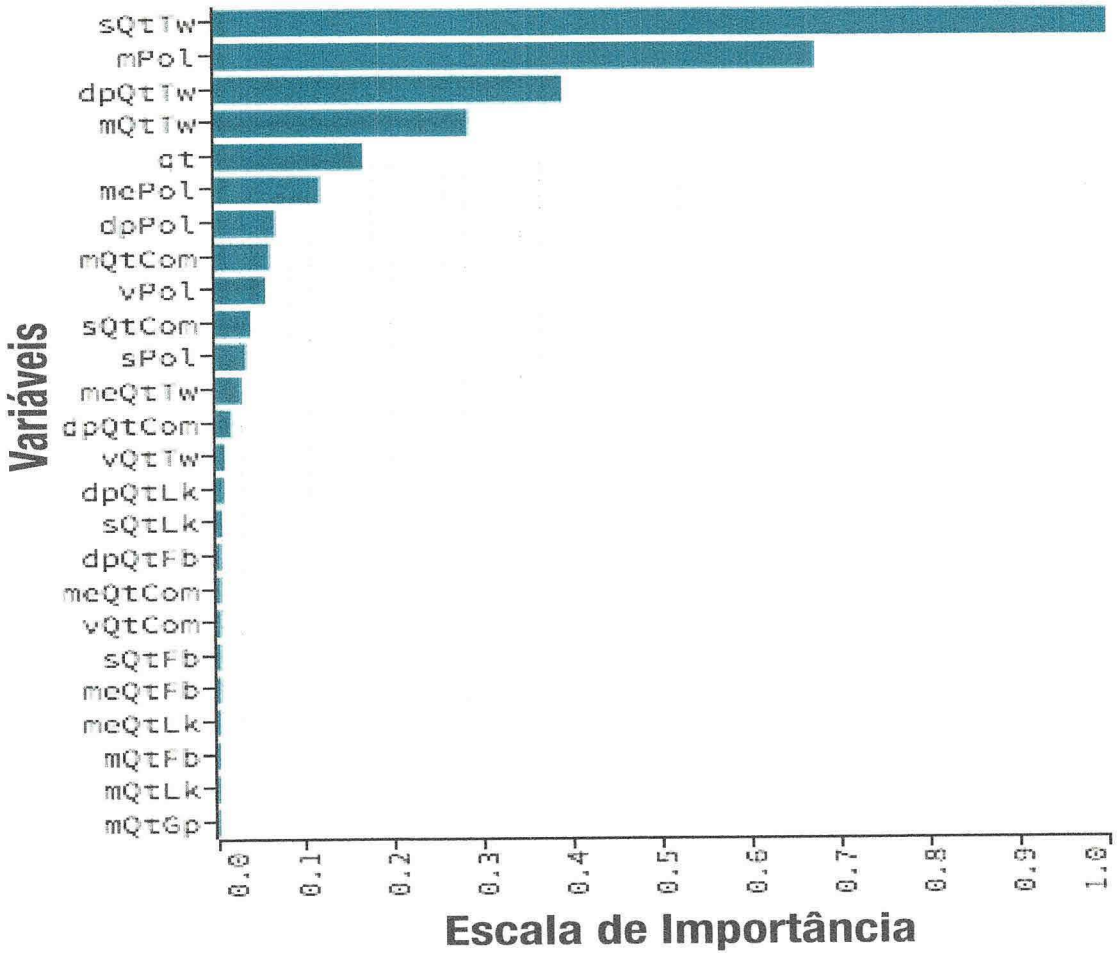

Figura 8.14: Qualidade da predição pela quantidade de notícias publicadas no período de 15 minutos.

domínio público possuem informação útil que podem ser utilizadas em modelos de previsão de modo a ampliar suas potencialidades e melhorar seus resultados. Foram apresentadas inúmeras evidências que apontam rumo a previsibilidade dos setores da BM&FBOVESPA e ortogonais à hipótese da eficiência de mercado para o cenário nacional. Entre elas pode-se enfatizar a superioridade dos modelos de previsão quanto comparados com o modelo randômico. Outras contribuições que esse capítulo apresentou foram:

- Foi verificado que a janela de tempo de 15 minutos é mais sensível às notícias que períodos maiores como 1 hora ou 1 dia.
- Para cada variável alvo em cada setor da BM&FBOVESPA, foi identificado o classificador estado-da-arte em sua configuração ideal que apresentou os melhores resultados de predição. Em todos os casos, tanto o modelo randômico quanto o modelo de manutenção da tendência foram superados significativamente.
- Foi verificado que o método proposto apresenta-se estável durante o prazo máximo de 24 dias dado um treinamento mínimo de 9 meses imediatamente anteriores.
- O método é capaz de gerar predições a partir de uma janela de tempo de 15 minutos onde tenham ocorrido ao menos 2 notícias econômicas. E, além disso, quando a quantidade de notícias em uma janela de 15 minutos é igual ou superior a 19 notícias, os resultados para acurácia tendem a superar os 70%.

Capítulo 9

Conclusões e Trabalhos Futuros

Neste capítulo são apresentados as considerações finais e diretrizes futuras que se pretende explorar em sequência.

9.1 Conclusões

O trabalho revelou evidências de um cenário até então desconhecido sobre as relações existentes entre a BM&FBOVESPA e as notícias econômicas publicadas em jornal de alta circulação no Brasil. Foram avaliados aspectos de quantidades de publicações, comentários, polaridade, engajamento em redes sociais e correlações com mercado acionário nacional. As evidências apresentadas contradizem, em grande medida, a hipótese de eficiência de mercado confirmando para o Brasil o que já foi evidenciado em outros mercados como Estados Unidos, Hong Kong, China, Turquia e Tehan.

Em todos os casos, os modelos de predição criados a partir do processamento de dados de notícias econômicas foram capaz de prover previsões para os próximos 15 minutos significativamente superiores as do *baselines* de comparação tanto para o *índice Bovespa* quanto para cada um dos setores da BM&FBOVESPA, em todas as suas características alvo. Tanto a análise de correlação quanto o desempenho desses modelos em janelas de tempo superiores a 15 minutos foram inferiores demonstrando com isso que uma ferramenta derivada deste trabalho trará mais benefícios para *day-traders* que outros perfis de atuação. Isso pode ser visto como algo bastante positivo, dado que na prática saber antecipadamente sobre algo abre vantagem ampla estratégica. Dito de outra forma, ter consciência da tendência futura a cada

15 minutos supera todos os modelos que necessitam de mais tempo para chegar ao mesmo resultado.

Uma vez que todos os modelos de predição recebem notícias como entrada, a falta de notícias acarreta, por óbvio, a falta de capacidade de utilização dos modelos tal como encontram-se atualmente. Apesar de ser uma fragilidade, esta condição é incomum. Conforme apresentado pela análise descritiva, em média, cada um dos jornais analisados publicam diariamente mais de 50 notícias econômicas e parte disso em horário de negociação.

É importante enfatizar que as variáveis de maior correlação e entropia são variáveis relacionadas a quantidades. Isto é, não há a necessidade de processar texto, executar complexos algoritmos e construir uma infraestrutura ampla de modo a dar suporte aos modelos de previsão construídos. Com exceção da variável relacionada à polaridade, todas as demais medidas são simplesmente baseadas em contagem. Isso torna simples e eficiente a implementação dos modelos gerados em ambientes reais.

Por fim, este trabalho lança luz sobre o impacto que notícias econômicas, mesmo as de ampla circulação nacional, trazem para a BM&FBOVESPA, apresentando evidências contrárias a hipótese de eficiência de mercado. Os resultados encontrados mostram que utilizar notícias econômicas de domínio público pode ampliar em 29% as chances de melhorar os retornos financeiros com investimentos em setores e no próprio índice Bovespa, revelando inclusive a sensibilidade dos setores e do IBOVE às publicações. No futuro pretende-se desenvolver uma ferramenta, baseada nos modelos desenvolvidos com o objetivo de tornar pequenos investidores, que não podem pagar por informações restritas aos grandes veículos de comunicação, mais conscientes quanto aos riscos e oportunidades, dando-lhes alguma vantagem e talvez incentivando maior inserção dos brasileiros em seu próprio mercado de capitais.

9.2 Trabalhos Futuros

No futuro pretende-se explorar os seguintes pontos:

1. Repetir os protocolos experimentais para as granularidades de *subsetor*, *segmento* e ações individuais de modo a verificar se se aplicam as mesmas conclusões;

2. Repetir os protocolos experimentais para outros índices diferentes do índice Bovespa, como o índice *IBr50* e *IBr100*, por exemplo;

3. Utilizar as notícias publicadas pela própria BM&FBOVESPA sobre as empresas participantes do mercado ao invés de apenas notícias de jornais;

4. Repetir os experimentos realizados para notícias de domínio específico, como Reuters e Infomoney;

5. Aplicar outras técnicas de aprendizagem de máquina, tais como redes neurais multicamadas sob os resultados;

6. Repetir os protocolos experimentais sem considerar as notícias do jornal G1 pela baixíssima correlação de seus parâmetros com o índice Bovespa e os setores;

7. Medir o impacto e a capacidade preditiva considerando apenas jornais individualmente e acrescentar para cada um deles uma forma de ponderação às variáveis sendo extraídas. Isto é, a depender do jornal sendo considerado, as variáveis sendo coletadas passariam a receber ajustes mediante um fator de impacto estabelecido para cada fonte.

8. Ajustar o limiar de consolidação para os alvos *quantidade de negócios*, *quantidade de contratos* e *volume financeiro*. Atualmente o limiar de consolidação para todos os alvos é de 0.5%. Apesar de funcionar bem para o alvo relacionado a preço, este liminar não apresenta-se como ideal para os demais alvos e um novo estudo para determinar limiares mais robustos surge como potencialmente útil;

9. Criar um protótipo de uma ferramenta capaz de processar as informações em tempo real e gerar previsões sobre o índice Bovespa e cada um dos setores da BM&FBOVESPA.

10. Combinar a perspectiva fundamentalista discutida neste trabalho com enfoques técnicos de modo a criar modelos híbridos hábeis a avaliar o melhor cenário mediante diferentes perspectivas do mercado.

Referências Bibliográficas

[Al Nasseri et al., 2015] Al Nasseri, A., Tucker, A., and de Cesare, S. (2015). Quantifying stocktwits semantic terms' trading behavior in financial markets: An effective application of decision tree algorithms. *Expert Systems with Applications*, 42(23):9192–9210.

[Araújo et al., 2016] Araújo, M., Reis, J. C., Pereira, A. C., and Benevenuto, F. (2016). An evaluation of machine translation for multilingual sentence-level sentiment analysis. In *Proceedings of the 31th Annual ACM Symposium on Applied Computing*. ACM.

[Asad, 2015] Asad, M. (2015). Optimized stock market prediction using ensemble learning. In *Application of Information and Communication Technologies (AICT), 2015 9th International Conference on*, pages 263–268. IEEE.

[Awasthi and Malafeyev, 2015] Awasthi, A. and Malafeyev, O. (2015). Is the indian stock market efficient-a comprehensive study of bombay stock exchange indices. *arXiv preprint arXiv:1510.03704*.

[Boudoukh et al., 2013] Boudoukh, J., Feldman, R., Kogan, S., and Richardson, M. (2013). Which news moves stock prices? a textual analysis. Technical report, National Bureau of Economic Research.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[Bulkowski, 2011] Bulkowski, T. N. (2011). *Encyclopedia of chart patterns*, volume 225. John Wiley & Sons.

[Buscaldi and Hernandez-Farias, 2015] Buscaldi, D. and Hernandez-Farias, I. (2015). Sentiment analysis on microblogs for natural disasters management: a study on the 2014 genoa floodings. In *Proceedings of the 24th International Conference on World Wide*

Web Companion, pages 1185–1188. International World Wide Web Conferences Steering Committee.

[Cambria et al., 2010] Cambria, E., Speer, R., Havasi, C., and Hussain, A. (2010). Senticnet: A publicly available semantic resource for opinion mining. In *AAAI fall symposium: commonsense knowledge*, volume 10, page 02.

[Chan and Franklin, 2011] Chan, S. W. and Franklin, J. (2011). A text-based decision support system for financial sequence prediction. *Decision Support Systems*, 52(1):189–198.

[Chan et al., 1982] Chan, T. F., Golub, G. H., and LeVeque, R. J. (1982). Updating formulae and a pairwise algorithm for computing sample variances. In *COMPSTAT 1982 5th Symposium held at Toulouse 1982*, pages 30–41. Springer.

[Chan, 2003] Chan, W. S. (2003). Stock price reaction to news and no-news: drift and reversal after headlines. *Journal of Financial Economics*, 70(2):223–260.

[Chmielewski et al., 2015] Chmielewski, L. J., Janowicz, M., Ochnio, L., and Orłowski, A. (2015). Clusterization of indices and assets in the stock market. In *Intelligent Data Engineering and Automated Learning–IDEAL 2015*, pages 541–550. Springer.

[Chok, 2010] Chok, N. S. (2010). *Pearson's versus Spearman's and Kendall's correlation coefficients for continuous data.* PhD thesis, University of Pittsburgh.

[Da Fonseca and Wang, 2015] Da Fonseca, J. and Wang, P. (2015). A joint analysis of market indexes in credit default swap, volatility and stock markets. *Applied Economics*, pages 1–18.

[de Mattos Neto et al., 2013] de Mattos Neto, P. S., Cavalcanti, G. D., Madeiro, F., and Ferreira, T. A. (2013). An ideal gas approach to classify countries using financial indices. *Physica A: Statistical Mechanics and its Applications*, 392(1):177–183.

[De Smedt and Daelemans, 2012] De Smedt, T. and Daelemans, W. (2012). Pattern for python. *The Journal of Machine Learning Research*, 13(1):2063–2067.

[Ding et al., 2008] Ding, X., Liu, B., and Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 231–240. ACM.

[Dodds and Danforth, 2010] Dodds, P. S. and Danforth, C. M. (2010). Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies*, 11(4):441–456.

[Dumont et al., 2009] Dumont, M., Marée, R., Wehenkel, L., and Geurts, P. (2009). Fast multi-class image annotation with random subwindows and multiple output randomized trees. In *Proc. International Conference on Computer Vision Theory and Applications (VISAPP)*, volume 2, pages 196–203.

[Esuli and Sebastiani, 2006] Esuli, A. and Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422. Citeseer.

[Fama, 1970] Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work*. *The journal of Finance*, 25(2):383–417.

[Feldman, 2013] Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89.

[Feldman et al., 2011] Feldman, R., Rosenfeld, B., Bar-Haim, R., and Fresko, M. (2011). The stock sonar—sentiment analysis of stocks based on a hybrid approach. In *Twenty-Third IAAI Conference*.

[Friedman, 2001] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

[Gallegos and Hau, 2015] Gallegos, D. and Hau, A. (2015). Predicting stock prices through textual analysis of web news. Technical report, Technical Report, Stanford.

[Geurts et al., 2006] Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63(1):3–42.

[Go et al., 2009] Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1:12.

[Gonçalves et al., 2013a] Gonçalves, P., Araújo, M., Benevenuto, F., and Cha, M. (2013a). Comparing and combining sentiment analysis methods. In *Proceedings of the first ACM conference on Online social networks*, pages 27–38. ACM.

[Gonçalves et al., 2013b] Gonçalves, P., Benevenuto, F., and Cha, M. (2013b). Panas-t: A pychometric scale for measuring sentiments on twitter. *arXiv preprint arXiv:1308.1857*.

[Gong and Sriboonchitta, 2016] Gong, X. and Sriboonchitta, S. (2016). The causal relationship between government opinions and chinese stock market in social media era. In *Causal Inference in Econometrics*, pages 481–493. Springer.

[González, 2016] González, M. (2016). Asymmetric causality in-mean and in-variance among equity markets indexes. *The North American Journal of Economics and Finance*, 36:49–68.

[Gu et al., 2015] Gu, Y., Storey, V. C., and Woo, C. C. (2015). Conceptual modeling for financial investment with text mining. In *Conceptual Modeling*, pages 528–535. Springer.

[Hagströmer and Norden, 2013] Hagströmer, B. and Norden, L. (2013). The diversity of high-frequency traders. *Journal of Financial Markets*, 16(4):741–770.

[Hutto and Gilbert, 2014] Hutto, C. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International AAAI Conference on Weblogs and Social Media*.

[Jayawardena et al., 2015] Jayawardena, N. I., Todorova, N., Li, B., and Su, J.-J. (2015). Forecasting the volatility of the japanese stock market using after-hour information in other markets. Technical report, Griffith University, Department of Accounting, Finance and Economics.

[Kadambari et al., 2015] Kadambari, S., Jaswal, K., Kumar, P., and Rawat, S. (2015). Using twitter for tapping public minds, predict trends and generate value. In *Advanced Computing & Communication Technologies (ACCT), 2015 Fifth International Conference on*, pages 586–589. IEEE.

[Kruse et al., 2013] Kruse, R., Borgelt, C., Klawonn, F., Moewes, C., Steinbrecher, M., and Held, P. (2013). *Computational intelligence: a methodological introduction*. Springer Science & Business Media.

[Levallois, 2013] Levallois, C. (2013). Umigon: sentiment analysis for tweets based on lexicons and heuristics. In *Proceedings of the International Workshop on Semantic Evaluation, SemEval*, volume 13.

[Li, 2015] Li, K. L. (2015). Automatic method for stock trading strategy. Technical report.

[Li et al., 2014] Li, X., Xie, H., Wang, R., Cai, Y., Cao, J., Wang, F., Min, H., and Deng, X. (2014). Empirical analysis: stock market prediction via extreme learning machine. *Neural Computing and Applications*, pages 1–12.

[Linstone et al., 1975] Linstone, H. A., Turoff, M., et al. (1975). *The Delphi method: Techniques and applications*, volume 29. Addison-Wesley Reading, MA.

[Ma and Liang, 2015] Ma, C. and Liang, X. (2015). Online mining in unstructured financial information: An empirical study in bulletin news. In *Service Systems and Service Management (ICSSSM), 2015 12th International Conference on*, pages 1–6. IEEE.

[Machado et al., 2015] Machado, E. J., Pereira, A., Castilho, D., Silva, E., and Brandão, H. (2015). Proposal and implementation of new trading strategies for stock markets using web data. In *Proceedings of the 21st Brazilian Symposium on Multimedia and the Web*, pages 113–120. ACM.

[Minev, 2013] Minev, M. (2013). Quantification of financial news for economic surveys. In *Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on*, pages 1141–1144. IEEE.

[Minev et al., 2012] Minev, M., Schommer, C., and Grammatikos, T. (2012). News and stock markets: A survey on abnormal returns and prediction models. Technical report, Technical Report, UL.

[Mohammad and Turney, 2013a] Mohammad, S. M. and Turney, P. D. (2013a). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

[Mohammad and Turney, 2013b] Mohammad, S. M. and Turney, P. D. (2013b). Nrc emotion lexicon. Technical report, NRC Technical Report.

[Nguyen et al., 2015] Nguyen, T. H., Shirai, K., and Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, 42(24):9603–9611.

[Nielsen, 2011] Nielsen, F. Å. (2011). A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

[Nunes Ribeiro et al., 2015] Nunes Ribeiro, F., Araújo, M., Gonçalves, P., Benevenuto, F., and André Gonçalves, M. (2015). A benchmark comparison of state-of-the-practice sentiment analysis methods. *arXiv preprint arXiv:1512.01818*.

[Oliveira et al., 2013] Oliveira, N., Cortez, P., and Areal, N. (2013). On the predictability of stock market behavior using stocktwits sentiment and posting volume. In *Progress in Artificial Intelligence*, pages 355–365. Springer.

[Park et al., 2013] Park, J., Barash, V., Fink, C., and Cha, M. (2013). Emoticon style: Interpreting differences in emoticons across cultures. In *ICWSM*.

[Piñeiro-Chousa et al., 2015] Piñeiro-Chousa, J. R., López-Cabarcos, M. Á., and Pérez-Pico, A. M. (2015). Examining the influence of stock market variables on microblogging sentiment. *Journal of Business Research*.

[Rao et al., 2015] Rao, A., Hule, S., Shaikh, H., Nirwan, E., and Daflapurkar, P. (2015). Survey: Stock market prediction using statistical computation and artificial neural networks. *International Research Journal of Engineering and Technology*.

[Schumaker and Chen, 2010] Schumaker, R. P. and Chen, H. (2010). A discrete stock price prediction engine based on financial news. *Computer*, 43(1):51–56.

[Seker et al., 2014] Seker, S. E., Mert, C., Al-Naami, K., Ozalp, N., and Ayan, U. (2014). Time series analysis on stock market for text mining correlation of economy news. *arXiv preprint arXiv:1403.2002*.

[Shi et al., 2016] Shi, W., Shang, P., Xia, J., and Yeh, C.-H. (2016). The coupling analysis between stock market indices based on permutation measures. *Physica A: Statistical Mechanics and its Applications*, 447:222–231.

[Skuza and Romanowski, 2015] Skuza, M. and Romanowski, A. (2015). Sentiment analysis of twitter data within big data distributed environment for stock prediction. In *Computer Science and Information Systems (FedCSIS), 2015 Federated Conference on*, pages 1349–1354. IEEE.

[Slivka and Biryol, 2015] Slivka, R. T. and Biryol, C. (2015). Stock index arbitrage in the turkish market. *Indian Journal of Finance*, 9(11):7–18.

[Sukprasert et al., 2015] Sukprasert, A., Kanchymalay, K., Salim, N., and Khan, A. (2015). Social network news sentiments and stock price movement: A correlation analysis. *Jurnal Teknologi*, 77(20).

[Taboada et al., 2008] Taboada, M., Anthony, C., Brooke, J., Grieve, J., and Voll, K. (2008). So-cal: Semantic orientation calculator. *Simon Fraser University, Vancouver*.

[Talarposhti et al., 2016] Talarposhti, F. M., Sadaei, H. J., Enayatifar, R., Guimarães, F. G., Mahmud, M., and Eslami, T. (2016). Stock market forecasting by using a hybrid model of exponential fuzzy time series. *International Journal of Approximate Reasoning*, 70:79–98.

[Tetlock et al., 2008] Tetlock, P. C., SAAR-TSECHANSKY, M., and Macskassy, S. (2008). More than words: Quantifying language to measure firms fundamentals. *The Journal of Finance*, 63(3):1437–1467.

[Thelwall, 2013] Thelwall, M. (2013). Heart and soul: Sentiment strength detection in the social web with sentistrength. *Proceedings of the CyberEmotions*, pages 1–14.

[Vega, 2010] Vega, S. (2010). High holy days provide advantages for the stock market. *Advances in Business Research*, 1(1):45–52.

[Voll and Taboada, 2007] Voll, K. and Taboada, M. (2007). Not all words are created equal: Extracting semantic orientation as a function of adjective relevance. In *AI 2007: Advances in Artificial Intelligence*, pages 337–346. Springer.

[Wang et al., 2012] Wang, H., Can, D., Kazemzadeh, A., Bar, F., and Narayanan, S. (2012). A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, pages 115–120. Association for Computational Linguistics.

[Wilson et al., 2005] Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., and Patwardhan, S. (2005). Opinionfinder: A system for subjectivity analysis. In *Proceedings of hlt/emnlp on interactive demonstrations*, pages 34–35. Association for Computational Linguistics.

[Wu and Olson, 2015] Wu, D. D. and Olson, D. L. (2015). Online stock forum sentiment analysis. In *Enterprise Risk Management in Finance*, pages 49–56. Springer.

[Wu et al., 2014] Wu, D. D., Zheng, L., and Olson, D. L. (2014). A decision support approach for online stock forum sentiment analysis. *Systems, Man, and Cybernetics: Systems, IEEE Transactions on*, 44(8):1077–1087.

[Yin et al., 2015] Yin, S., Wu, F., Luo, H., and Gao, H. (2015). Support vector regression based approach for key index forecasting with applications. In *Industrial Informatics (INDIN), 2015 IEEE 13th International Conference on*, pages 591–596. IEEE.

[Zhang, 2004] Zhang, H. (2004). The optimality of naive bayes. *AA*, 1(2):3.

[Zhao and Wang, 2015] Zhao, L. and Wang, L. (2015). Price trend prediction of stock market using outlier data mining algorithm. In *Big Data and Cloud Computing (BDCloud), 2015 IEEE Fifth International Conference on*, pages 93–98. IEEE.

[Zhu et al., 2009] Zhu, J., Zou, H., Rosset, S., and Hastie, T. (2009). Multi-class adaboost. *Statistics and its Interface*, 2(3):349–360.

Apêndice A

Sistema Financeiro Brasileiro

O Sistema Financeiro Brasileiro (SFN) é o sistema composto por cinco órgãos normativos: o Conselho Monetário Nacional (CMN), Banco central do Brasil (BACEN), a Comissão de Valores Mobiliários (CVM), o Conselho Nacional de Seguros Privados (CNSP) e o Conselho Nacional de Previdência Complementar (CNPC) que tem como responsabilidade a implementação de regras que visem a capitação de recursos financeiros, sua distribuição, circulação e regulação de processos que promovam o desenvolvimento financeiro nacional de forma equilibrada. Tal sistema detém controle sobre todas as instituições que são ligadas às atividades econômicas dentro no Brasil. De forma sintética, o SFN define o que são recursos financeiros e como manipulá-los legalmente no Brasil.

O artigo 192 da Constituição Federal de 1988 diz:

> *O Sistema Financeiro Nacional, estruturado de forma a promover o desenvolvimento equilibrado do País e a servir aos interesses da coletividade, em todas as partes que o compõem, abrangendo as cooperativas de crédito, será regulado por leis complementares que disporão, inclusive, sobre a participação do capital estrangeiro nas instituições que o integram. (Redação dada pela Emenda Constitucional nº 40, de 2003) (Vide Lei nº 8.392, de 1991)*

A Figura A.1 apresenta a composição sintética do SFN e a distinção entre os órgãos normativos e agências supervisoras.

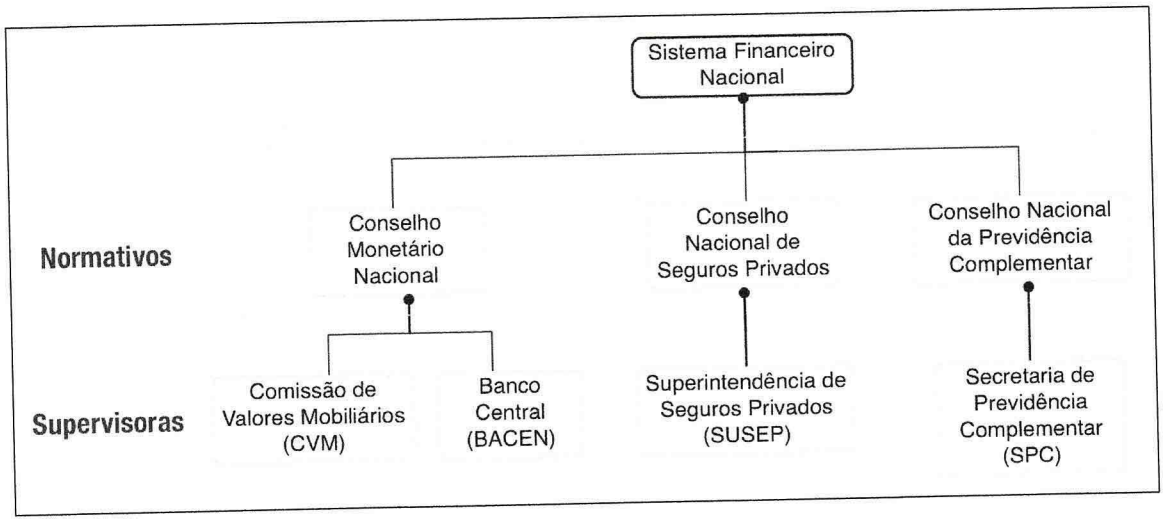

Figura A.1: Composição do Sistema Financeiro Nacional.

A.1 Componentes do Sistema Brasileiro

A seguir serão apresentados em detalhes os órgãos que compõem o SFN.

A.1.1 Conselho Monetário Nacional (CMN)

O Conselho Monetário Nacional (CMN) é o órgão superior do Sistema Financeiro Nacional e tem a responsabilidade de formular a política monetária e crediticia visando a estabilidade da moeda e o desenvolvimento econômico e social do Brasil. É composto por:

- Ministro da Fazenda (Presidente do Conselho);
- Ministro do Planejamento, Orçamento e Gestão;
- Presidente do Banco Central do Brasil.

Os seus membros reúnem-se uma vez por mês para deliberarem sobre assuntos relacionados com as competências do CMN. É órgão que normatiza as regras e leis que em seguida serão executadas pelo Banco Central e pela Comissão de Valores Mobiliários.

A.1.2 Banco Central do Brasil (BACEN)

O Banco Central do Brasil (BACEN[44]) é o orgão responsável por organizar e assessorar as sessões deliberativas pelo CMN (preparar, assessorar e dar suporte durante as reuniões, elaborar as atas e manter seu arquivo histórico). De forma objetiva, o Banco Central do Brasil opera como sendo a Secretaria-Executiva do CMN. Realiza a supervisão das normas do SFN para o sistema bancário e não-bancário (cooperativas, consórcios, etc.).

O presidente do BACEN é indicado pelo Presidente da República e aprovado pelo Senado Federal por meio de votação secreta.

A.1.3 Comissão de Valores Mobiliários (CVM)

A Comissão de Valores Mobiliários (CVM[45] é uma entidade vinculada ao Ministério da Fazenda com personalidade jurídica e patrimônio próprios. Possui autoridade administrativa independente, ausência de subordinação hierárquica, mandato fixo e estabilidade de seus dirigentes, além de autonomia financeira e orçamentária.

A CVM tem como meta prezar pela integridade e desenvolvimento sustentável do mercado de capitais buscando o equilíbrio entre a iniciativa de agentes e a proteção dos investidores. Supervisiona as bolsas de valores e corretoras.

A.1.4 Conselho Nacional de Seguros Privados (CNSP)

O Conselho Nacional de Seguros Privados (CNSP[46]) é o órgão normativo das atividades securitícias do país. Tem como principais atribuições o estabelecimento de diretrizes e normas para a política de seguros privados, fiscalização e penalização dos que exercem atividades subordinadas a CNSP.

A.1.5 Conselho Nacional de Previdência Complementar (CNPC)

O Conselho Nacional de Previdência Complementar (CNPC[47]) é o organismo que regula o regime de previdência complementar, também conhecida como previdência privada, normal-

[44] www.bcb.gov.br/?CMNENTENDA

[45] www.cvm.gov.br

[46] www.fazenda.gov.br/institucional/conselho-nacional-de-seguros-privados

[47] www.previdencia.gov.br/a-previdencia/orgaos-colegiados

mente operado por entidades fechadas.

A.1.6 Mercado de Capitais

O mercado de capitais surge quando a quantidade de dinheiro que se deseja tomar emprestado é superior ao que um banco pode suportar ou assumir como risco. É um sistema de distribuição de valores mobiliários que visa a liquidez de títulos emitidos por empresas. Constitui-se pelas bolsas de valores, corretoras e outras instituições financeiras autorizadas.

No mercado de capitais, os principais títulos negociados são as ações (títulos do capital da empresa) e debêntures (empréstimos tomados por empresas por meio do mercado acionário). Tais títulos permitem a circulação de capital para custear o desenvolvimento econômico (financiamento de novos projetos) ou até mesmo a transformação de patrimônio em dinheiro.

Em geral as empresas abrem capital por:

1. **Acesso a Capital:** Para financiar projetos de investimento[48].

2. **Liquidez Patrimonial:** Como possibilidade de sócio de transformar porcentagem das ações da empresa que possui em dinheiro. Trata-se de transformar patrimônio em dinheiro, vendendo sua parte a vários investidores para evitar monopólio.

3. **Projeção Institucional:** Por ser bastante rigoroso e abranger inúmeras restrições de transparência, qualidade, dentre outros aspectos, muitas empresas acreditam que ao cumprir todo procedimento para abertura de capital, lançando-se ao mercado, conseguem com isso, maior projeção e notoriedade de sua imagem constitucional.

4. **Reestruturação de Passivos:** Para ser forçada a recolocar-se na rota de crescimento.

O mercado de capitais pode ainda ser subdividido em dois mercados:

- **Mercado Primário:** Empresas e/ou Governo emitem títulos para captação de recursos de investidores.

- **Mercado Secundário:** Não envolve emissores nem contempla entrada de novos títulos apenas a transferência de titularidade. Ou seja, resume-se a compra e venda de títulos.

[48]Forma que o empresário Eik Baptista utilizou para financiar o projeto OGX.

A.1.7 Acesso ao Mercado

Para ter acesso ao mercado de capitais faz-se necessário abertura de conta junto a uma corretora, o processo é similar a abrir uma conta corrente. Após a abertura de conta na corretora e transferência de recursos (CDBs, títulos, dinheiro, etc.) é possível ter acesso ao mercado. A corretora, em cumprimento a legislação, é responsável por avaliar a coerência das operações do investidor com sua capacidade de investimento junto as instituições supervisoras (CVM, BACEN), fornecer a seus investidores acesso ao pregão eletrônico via *home broker* – Sistema de software vinculado a BM&FBOVESPA que permite aos investidores acesso ao livro de ofertas e realização de suas operações de compra e venda de qualquer computador.

A corretora lucra com as operações realizadas por seus clientes no mercado de ações (taxa de corretagem) e com a venda de inúmeros serviços de acessoria.

A.1.8 Síntese da Dinâmica do Mercado de Ações

1. Empresas abrem capital via IPO (Initial Public Offering);
2. Após seu lançamento, as ações da empresa passa a ser listada no livro de ofertas junto as ações das demais empresas;
3. Vendedores e compradores fazem negócios ao longo de um dia de mercado e todos os negócios são registrados (quantidade, preço, horário, comprador e vendedor).
4. Ações trocam de mãos no mercado secundário e inicia-se a dinâmica de mudanças de preço baseado no princípio de oferta e demanda baseado na expectativa de futuro dessa empresa.

Apêndice B

Evidências

A seguir serão apresentadas evidências da falta de notícias econômicas fornecidas entre os anos de 2001 e 2006 pelo jornal Folha de São paulo.

B.1 Evidências para o Jornal Folha de São Paulo

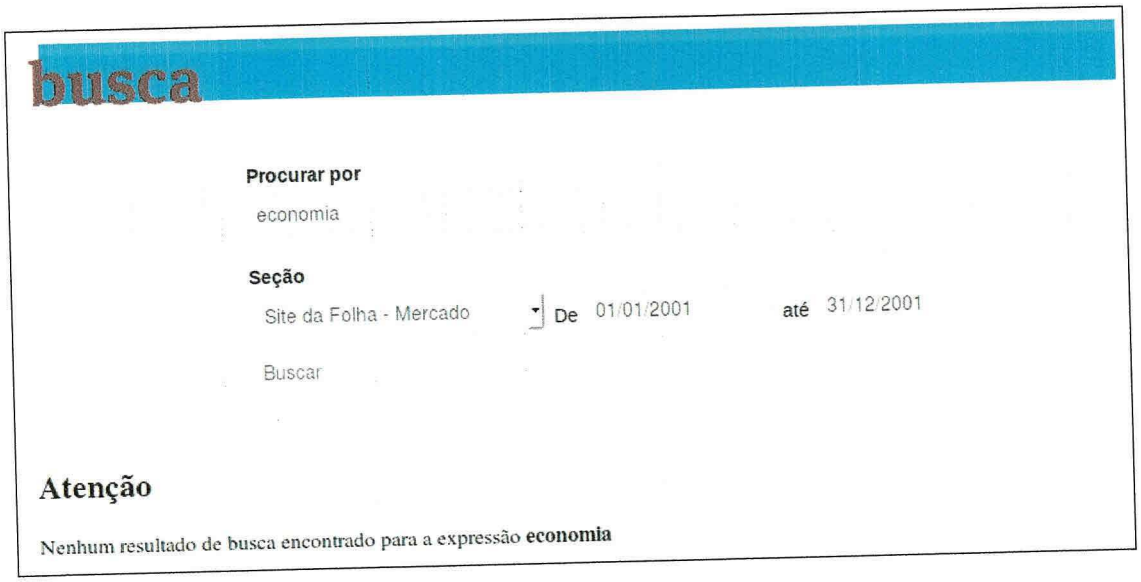

Figura B.1: Evidência da falta de notícias para o ano de 2001

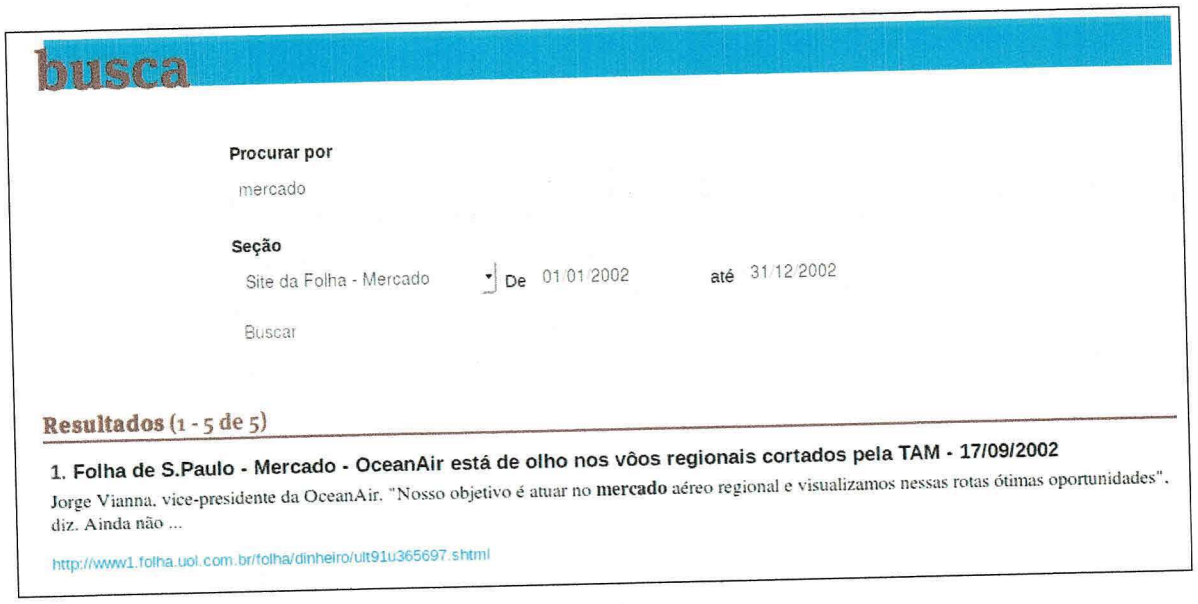

Figura B.2: Evidência da falta de notícias para o ano de 2002

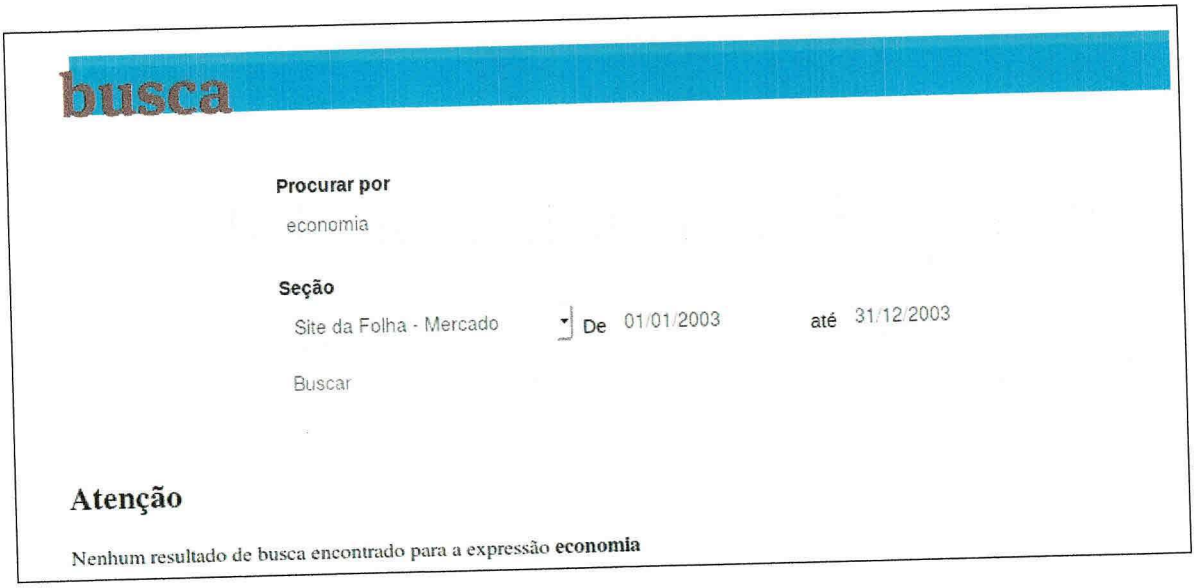

Figura B.3: Evidência da falta de notícias para o ano de 2003

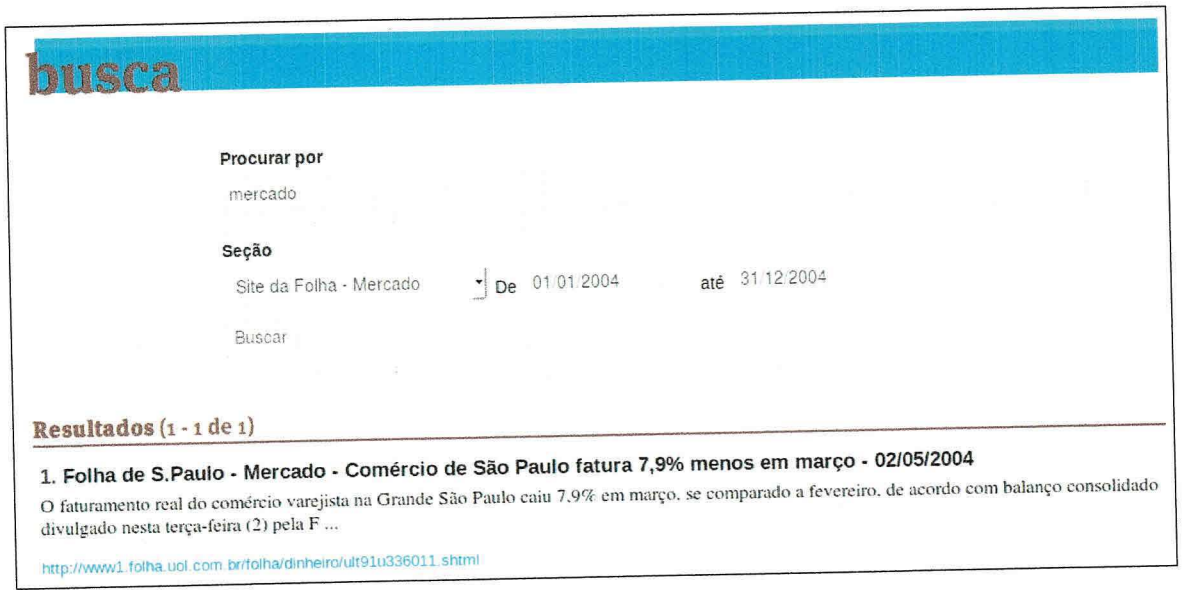

Figura B.4: Evidência da falta de notícias para o ano de 2004

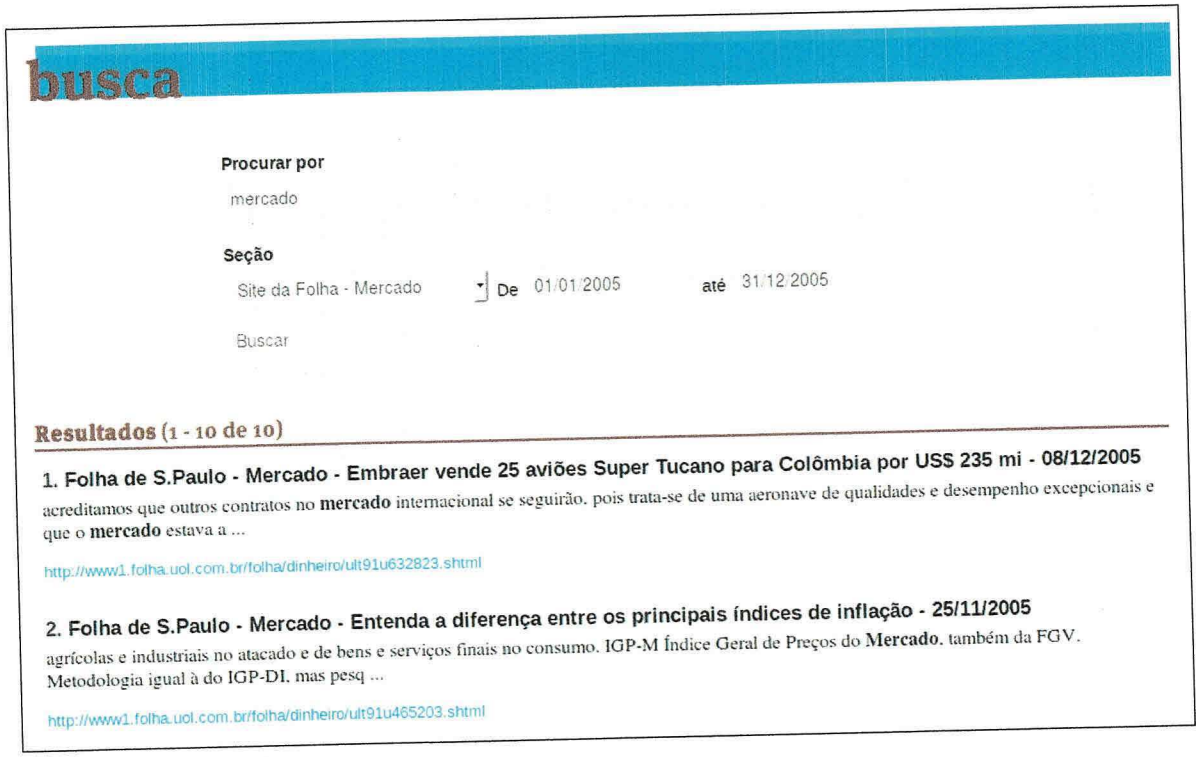

Figura B.5: Evidência da falta de notícias para o ano de 2005

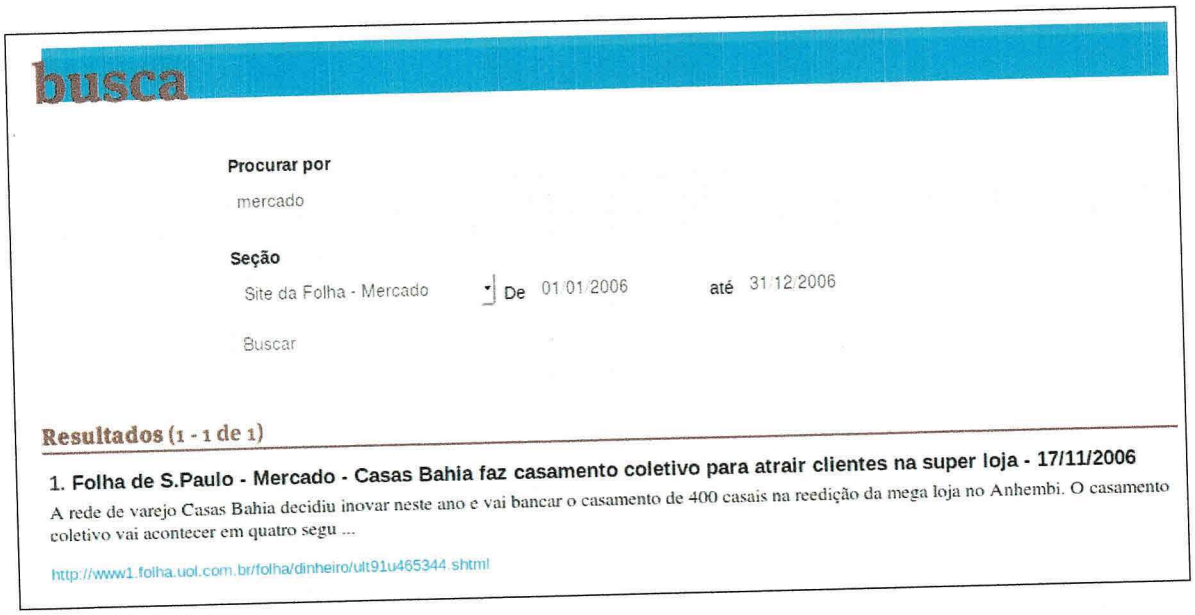

Figura B.6: Evidência da falta de notícias para o ano de 2006

Apêndice C

Análise Descritiva por Jornal

Nesta seção serão apresentadas as características particulares dos jornais G1, Folha de São Paulo e Estadão para cada uma das etapas da análise descritiva proposta no capítulo 5.

C.1 Mês

A seguir são apresentados detalhes da variabilidade para cada um dos jornais analisados por mês separadamente:

Folha

A Figura C.1 apresenta a variabilidade do número de publicações ao longo do meses do ano para o jornal Folha de São Paulo. Não só o número de publicações mas também o índice de tendência central para o jornal Folha de São Paulo foram maiores para o mês de agosto.

Estadão

A Figura C.2 apresenta a variabilidade do número de publicações ao longo do meses do ano para o jornal Estadão. Com exceção dos meses de julho e dezembro, há bastante regularidade quanto ao índice de tendência central de aproximadamente 45 notícias ao dia, o que indica, em certa medida, uma padronização ou regularidade da quantidade de publicações.

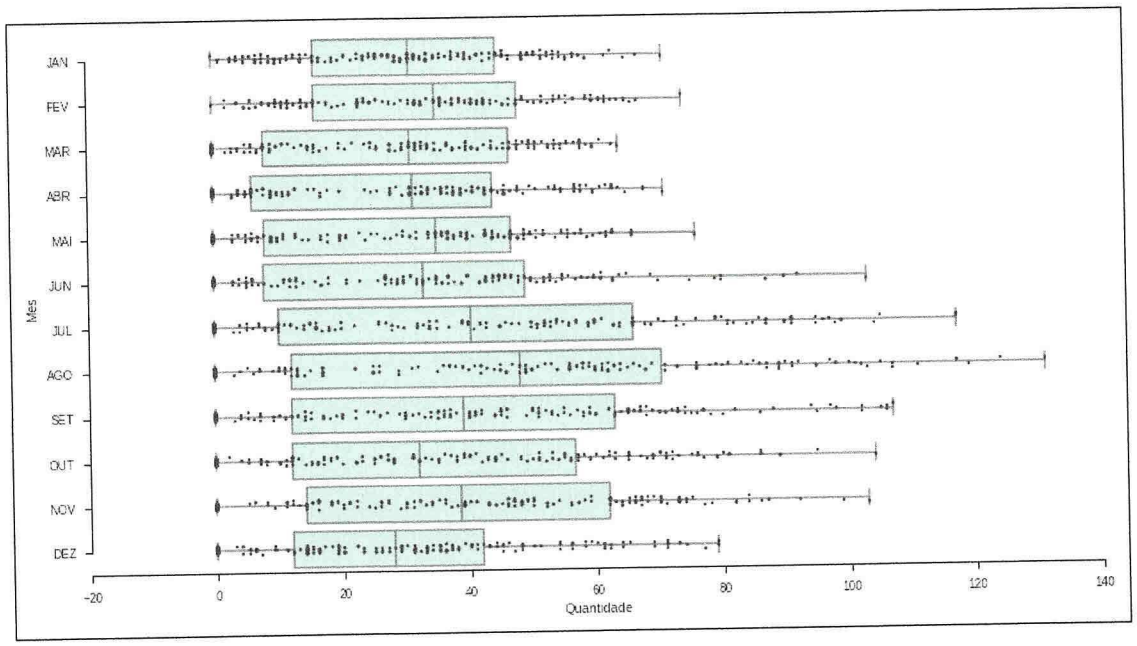

Figura C.1: Quantidade de notícias publicadas por mês para o jornal Folha de São Paulo.

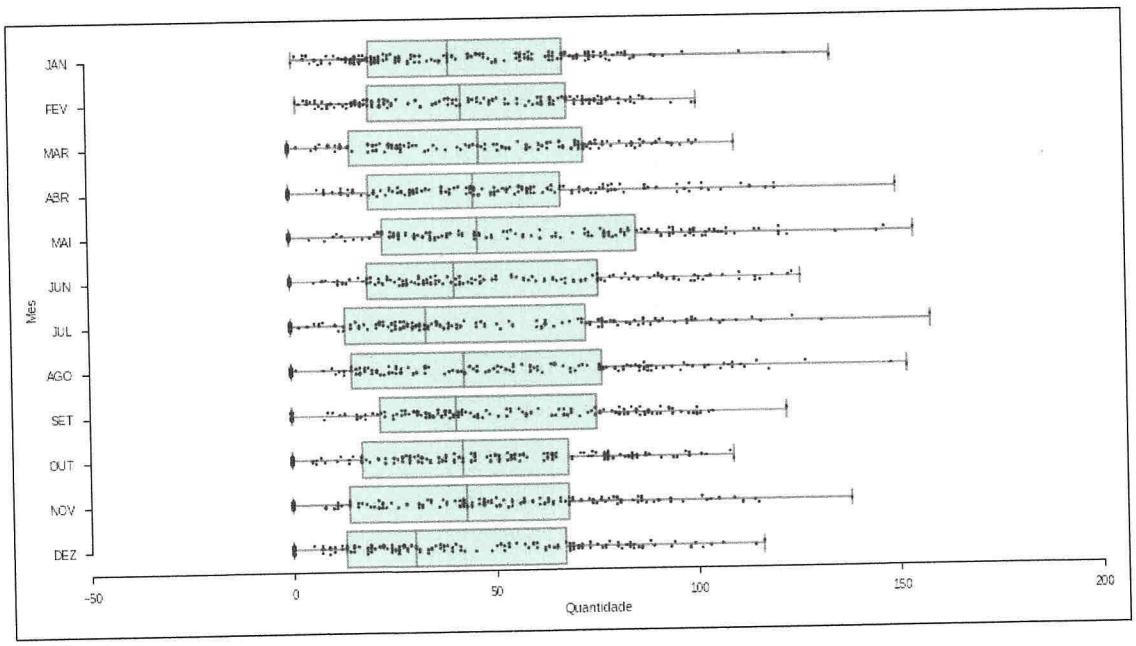

Figura C.2: Quantidade de notícias publicadas por mês para o jornal Estadão.

G1

A Figura C.3 apresenta a variabilidade do número de publicações ao longo dos meses do ano para o jornal G1. Algumas características notórias e únicas são a alta variabilidade da quantidade de publicações ao longo dos meses (falta de uniformidade) e a presença de *gaps* indicando uma descontinuidade em determinadas quantidades. Além disso, o número de *outliers* para o jornal G1 é superior a todos os outros jornais comparados.

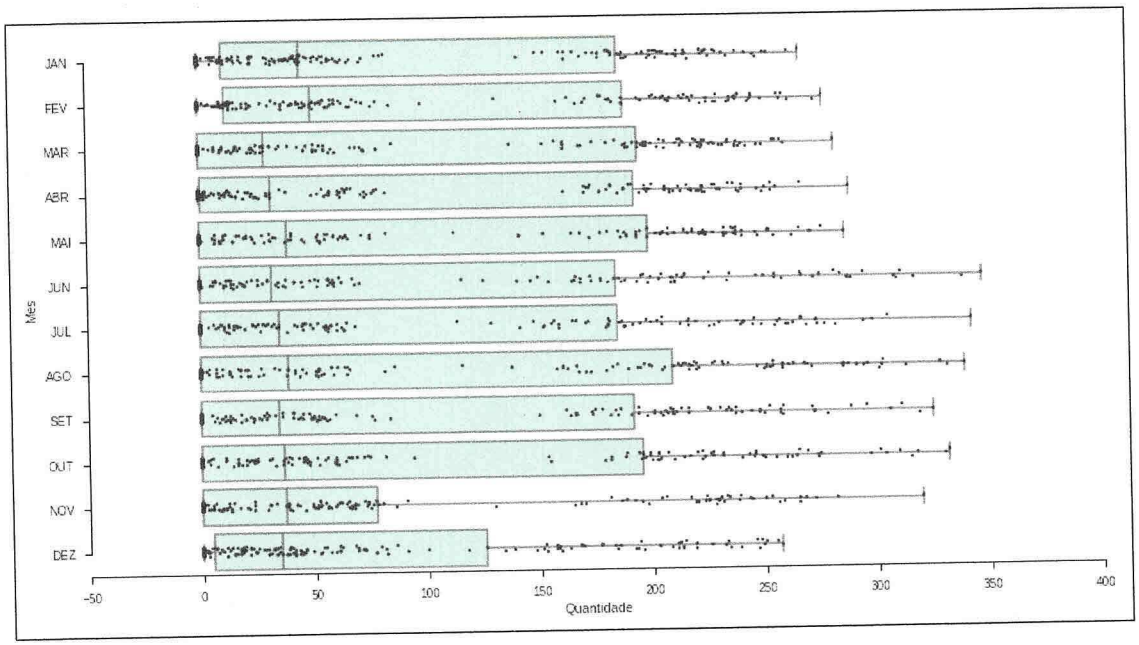

Figura C.3: Quantidade de notícias publicadas por mês para o jornal G1.

C.2 Dia de Semana

A seguir são apresentados detalhes da variabilidade para cada um dos jornais analisados por dia de semana separadamente:

Estadão

A Figura C.4 apresenta a variabilidade do número de publicações ao longo da semana para o jornal Estadão. Percebe-se um comportamento semelhante para todos os dias da semana excetuando-se sábados e domingos. As quartas-feiras apresenta uma sutil superioridade em relação aos demais dias.

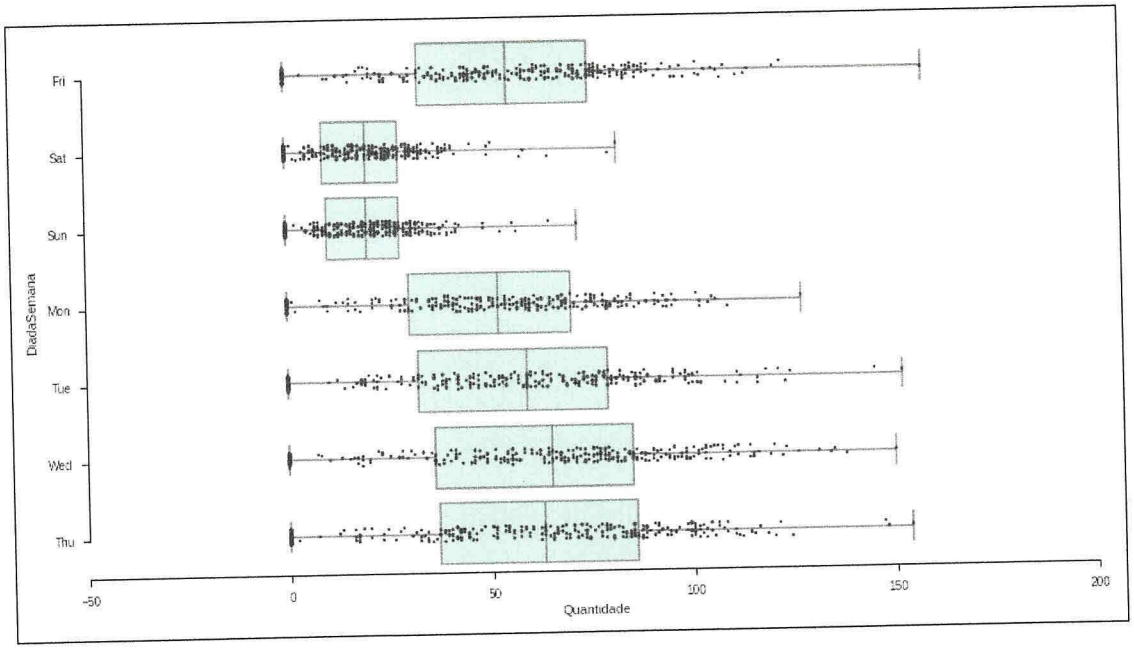

Figura C.4: Quantidade de notícias publicadas por dia da semana para o jornal Estadão.

Folha de São Paulo

A Figura C.5 apresenta a variabilidade do número de publicações ao longo da semana para o jornal Folha de São Paulo. Percebe-se um índice central aproximadamente uniforme para todos os dias da semana excetuando-se sábados e domingos.

G1

A Figura C.6 apresenta a variabilidade do número de publicações ao longo da semana para o jornal G1. É possível verificar bastante variabilidade em relação aos outros jornais e novamente os espaçamentos vazios (*gaps*) são percebidos para os dias da semana.

C.3 Repercussão

A seguir são apresentados detalhes da variabilidade em relação a repercussão para cada um dos jornais analisados separadamente.

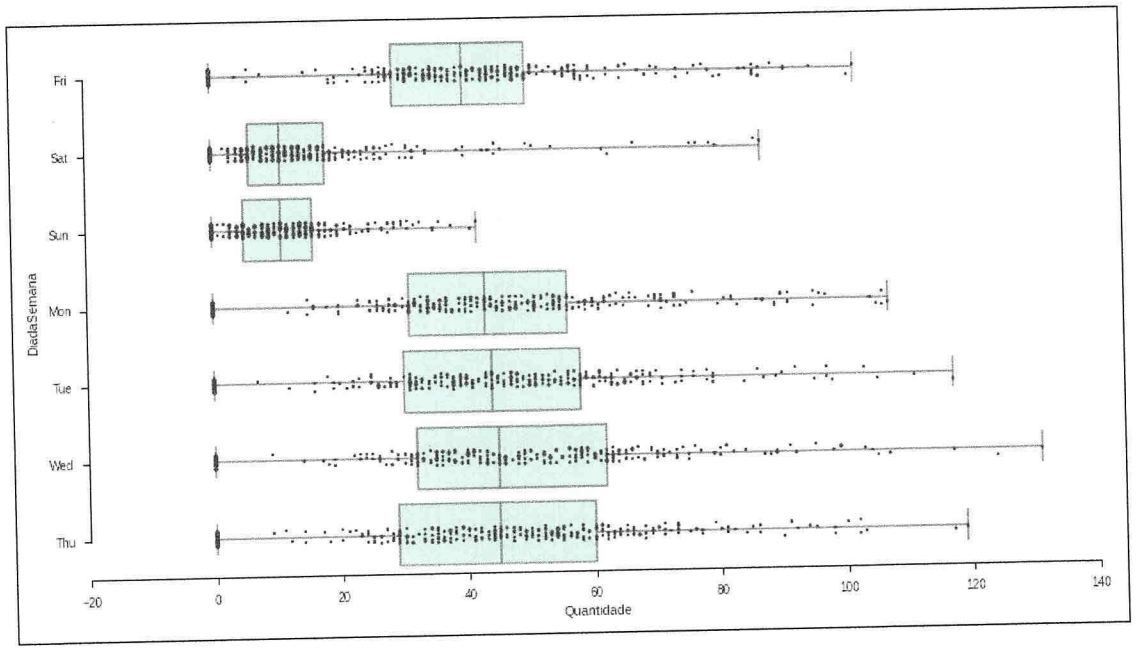

Figura C.5: Quantidade de notícias publicadas por dia da semana para o jornal Folha de São Paulo.

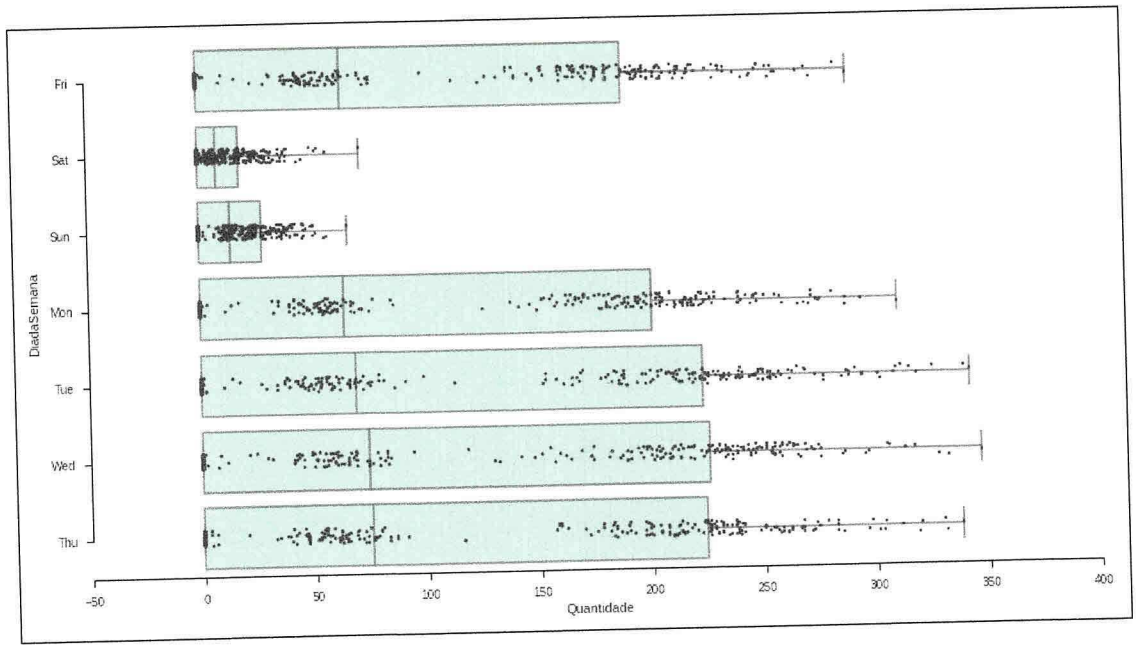

Figura C.6: Quantidade de notícias publicadas por dia da semana para o jornal G1.

C.3.1 Comentários

Ano

As Figuras C.7 e C.8 apresentam a variabilidade da quantidade de comentários por ano para os jornais Folha de São Paulo e Estadão e elucida que o comportamento do gráfico geral não é apenas fruto de *outliers* esporádicos e sim de interesse dos leitores em postar seus comentários. A partir de 2011 é possível perceber uma diminuição do número de comentários provavelmente devido ao crescimento de outras mídias como Twitter e Facebook.

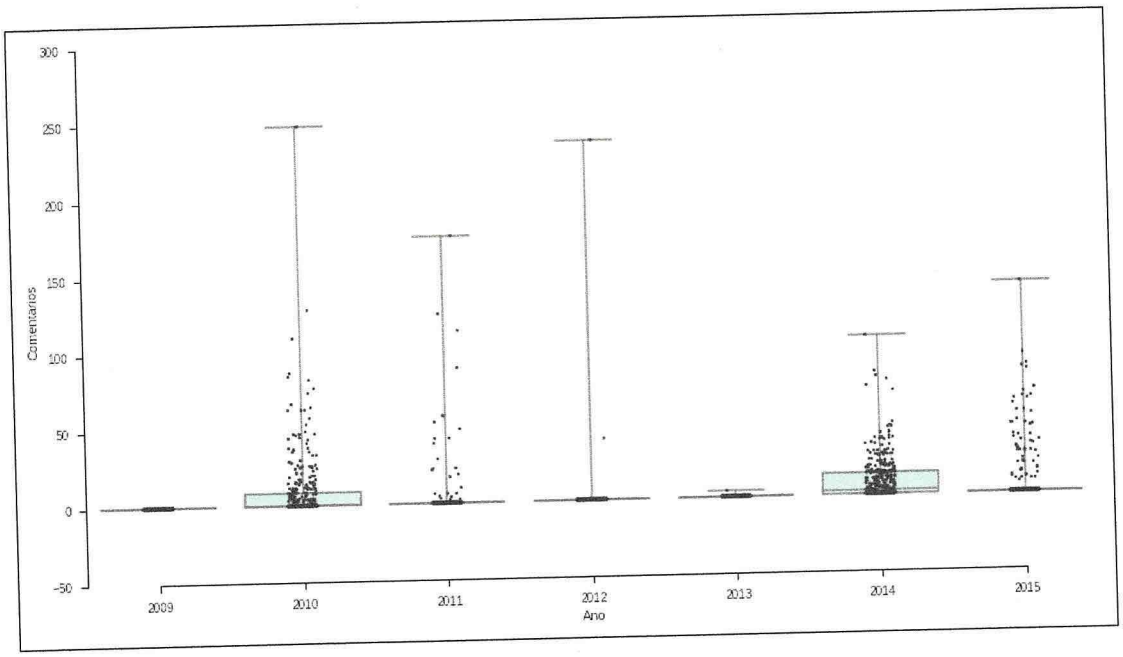

Figura C.7: Variabilidade dos comentários recebidos ao longo dos anos para o jornal Estadão.

Mes

As Figuras C.9 e C.10 que apresentam a variabilidade dos comentários ao longo do meses para cada jornal mostra que a tendência central é preservada em todos os meses. Setembro é o mês mais provável para o número de comentários.

Dia

As Figuras C.11 e C.12 apresentam a variabilidade da quantidade de comentários ao longo dos dias do mês. Para os jornais Estadão e Folha de São Paulo é notório que alguns dias

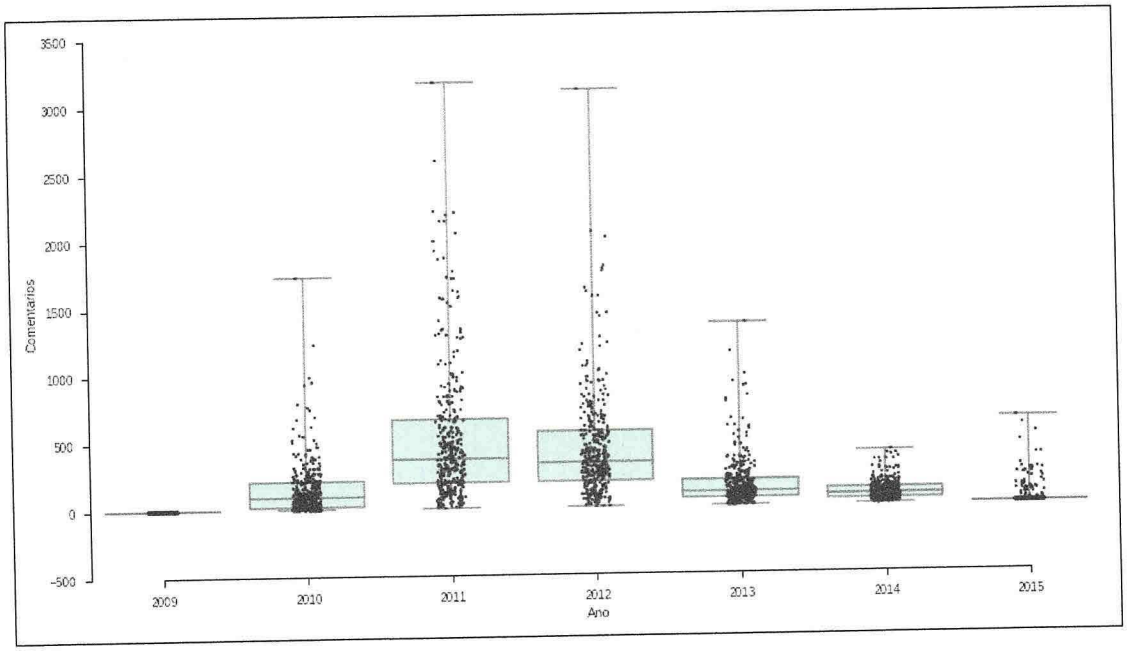

Figura C.8: Variabilidade dos comentários recebidos ao longo dos anos para o jornal Folha de São Paulo.

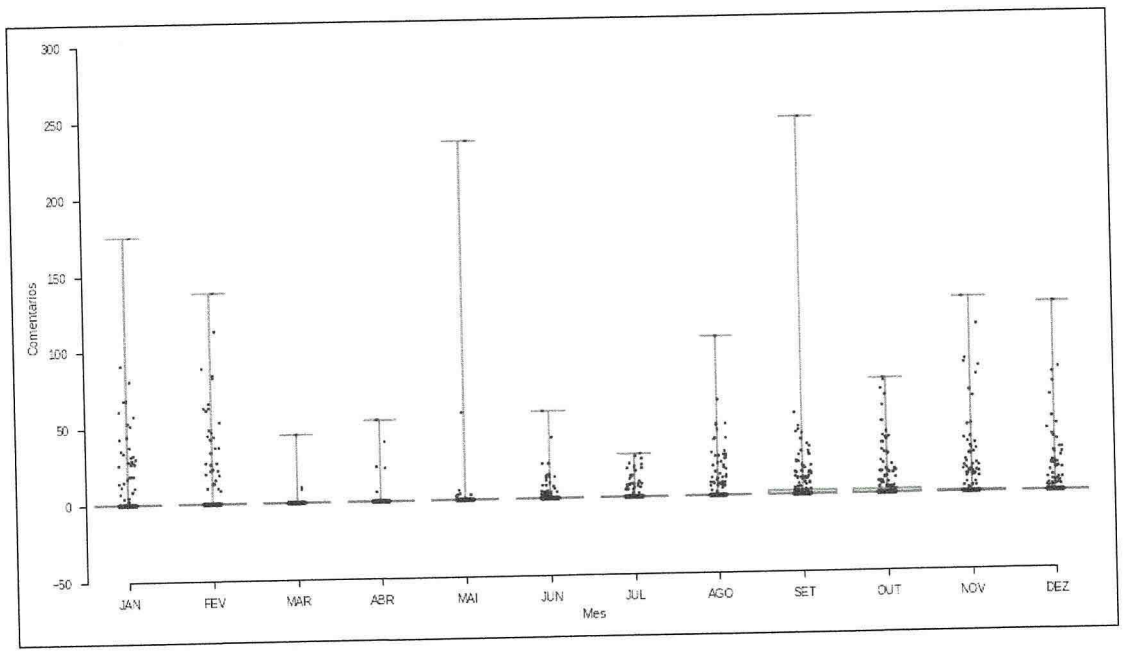

Figura C.9: Quantidade de comentários de notícias econômicas recebidos ao longo dos meses pelo jornal Estadão.

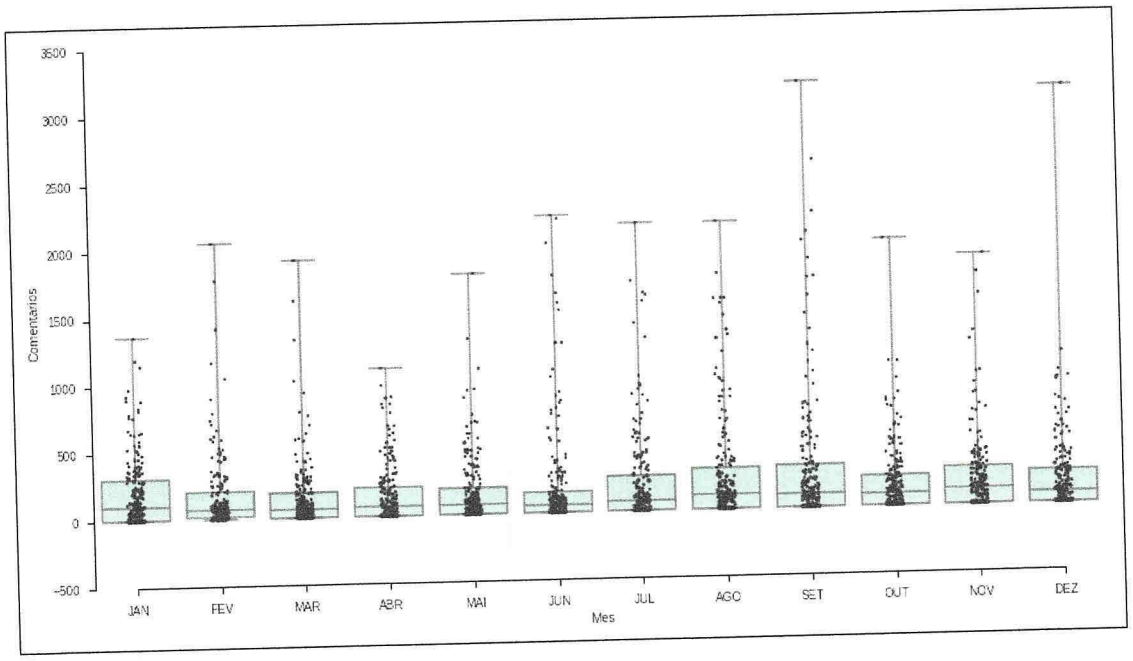

Figura C.10: Quantidade de comentários de notícias econômicas recebidos ao longo dos meses pelo jornal Folha de São Paulo.

de pico não são causados por *outliers* e que de fato representam um comportamento a ser explorado.

Dia de Semana

As Figuras C.13 e C.14 evidenciam um comportamento recorrente de predileção pela quarta-feira.

C.4 Twitter

Ano

As Figuras C.15, C.16 e C.17 apresentam a variabilidade da quantidade de tweets recebidos por cada notícia para os jornais Estadão, Folha de São Paulo e G1, respectivamente, apresentam um expressivo número de *outliers* para o ano de 2014, provavelmente devido as campanhas eleitorais. Apesar de haver um gradativo e crescente número de tweets para notícias econômicas o pico é artificial proporcionado por fatos atípicos.

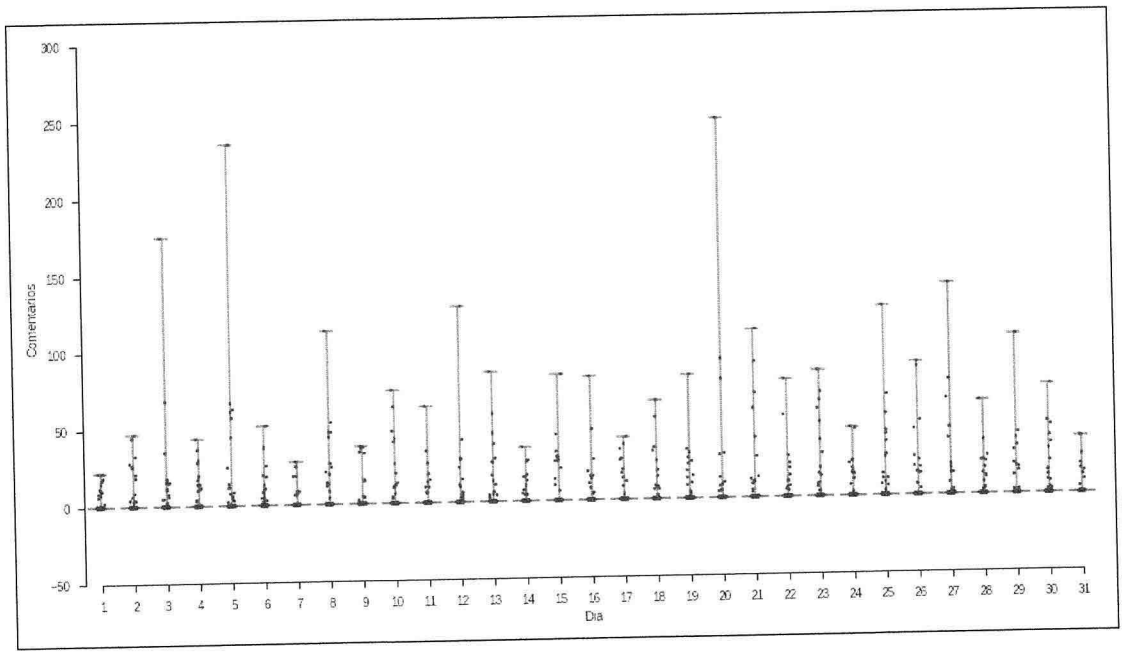

Figura C.11: Quantidade de comentários de notícias econômicas por dia do mês para o jornal Estadão.

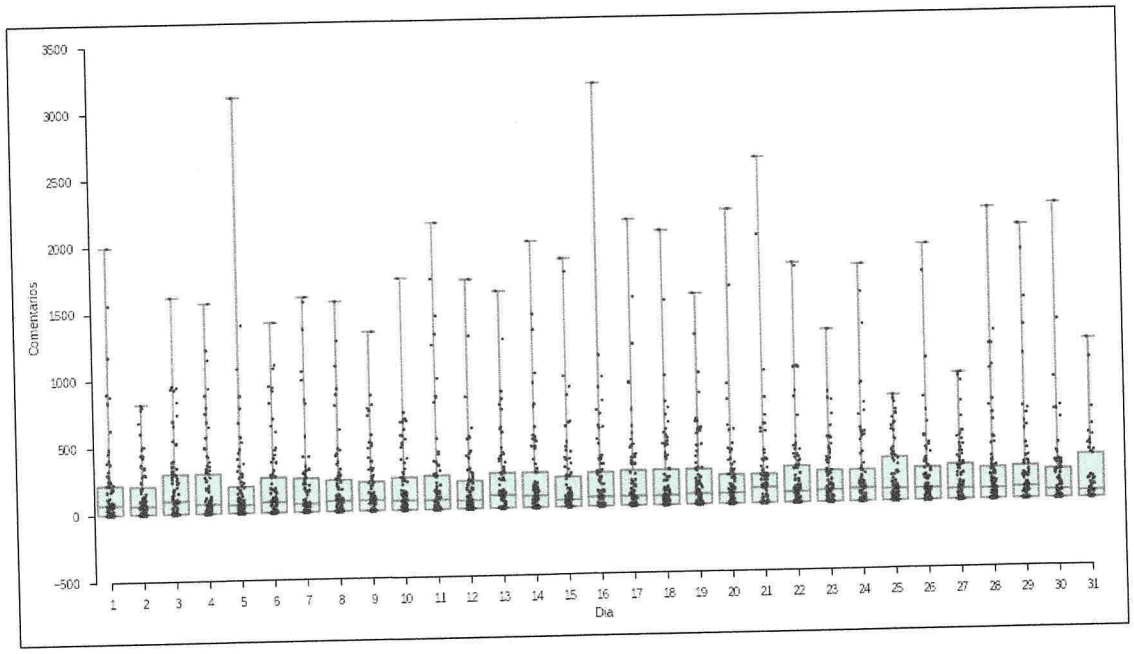

Figura C.12: Quantidade de comentários de notícias econômicas por dia do mês para o jornal Folha de São Paulo.

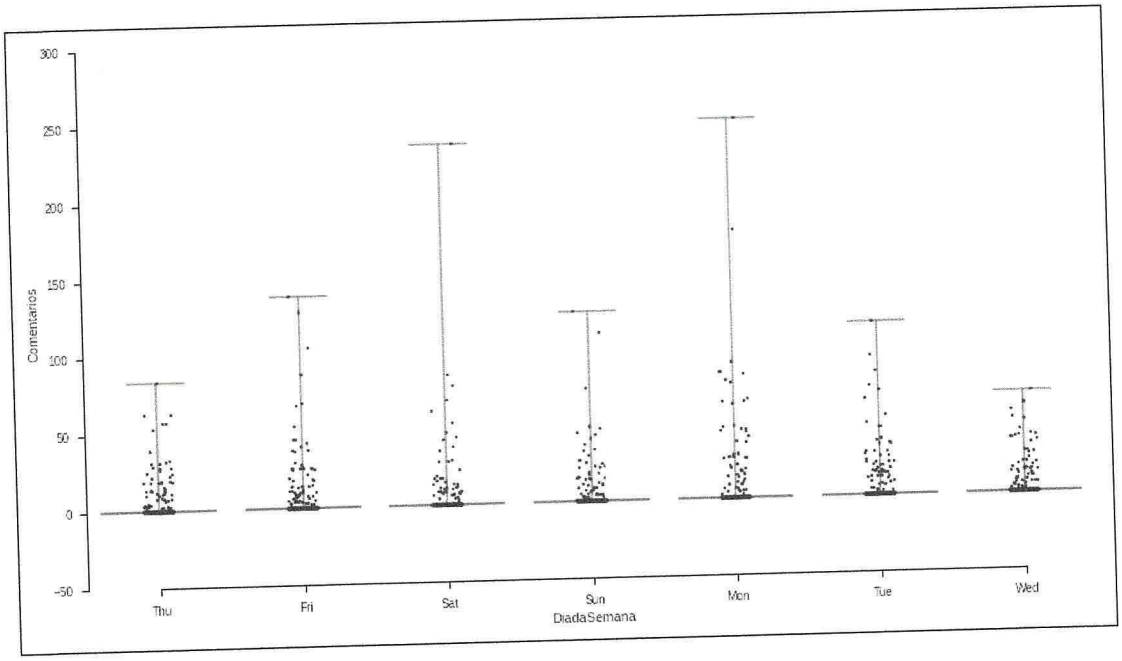

Figura C.13: Quantidade de comentários de notícias econômicas por dia da semana para o jornal Estadão.

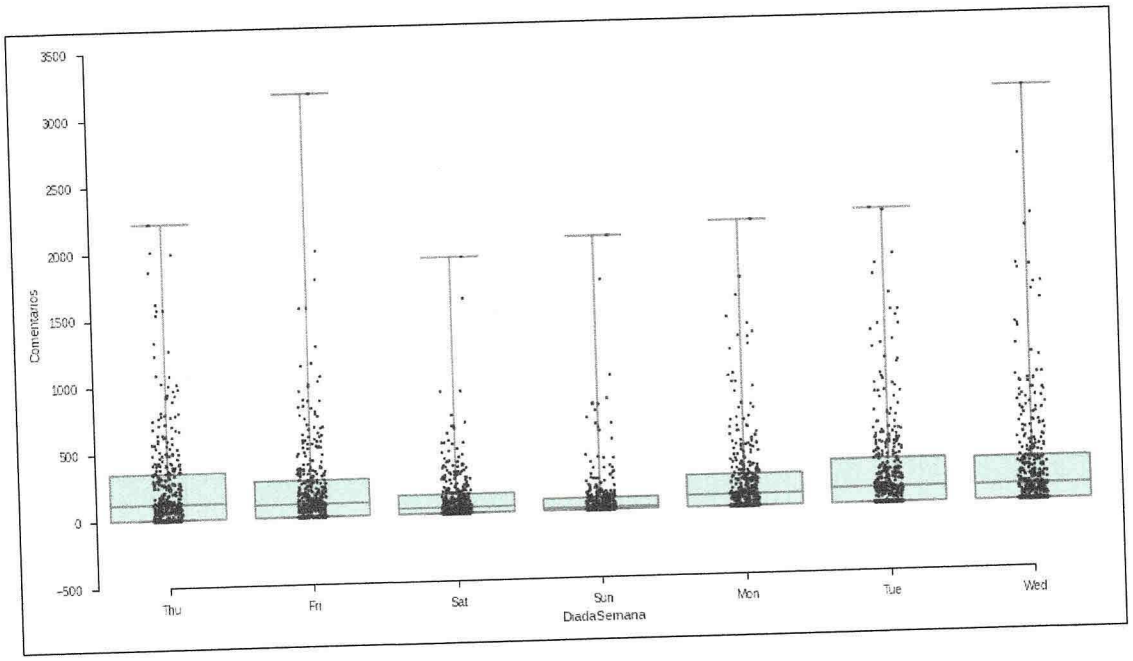

Figura C.14: Quantidade de comentários de notícias econômicas por dia da semana para o jornal Folha de São Paulo.

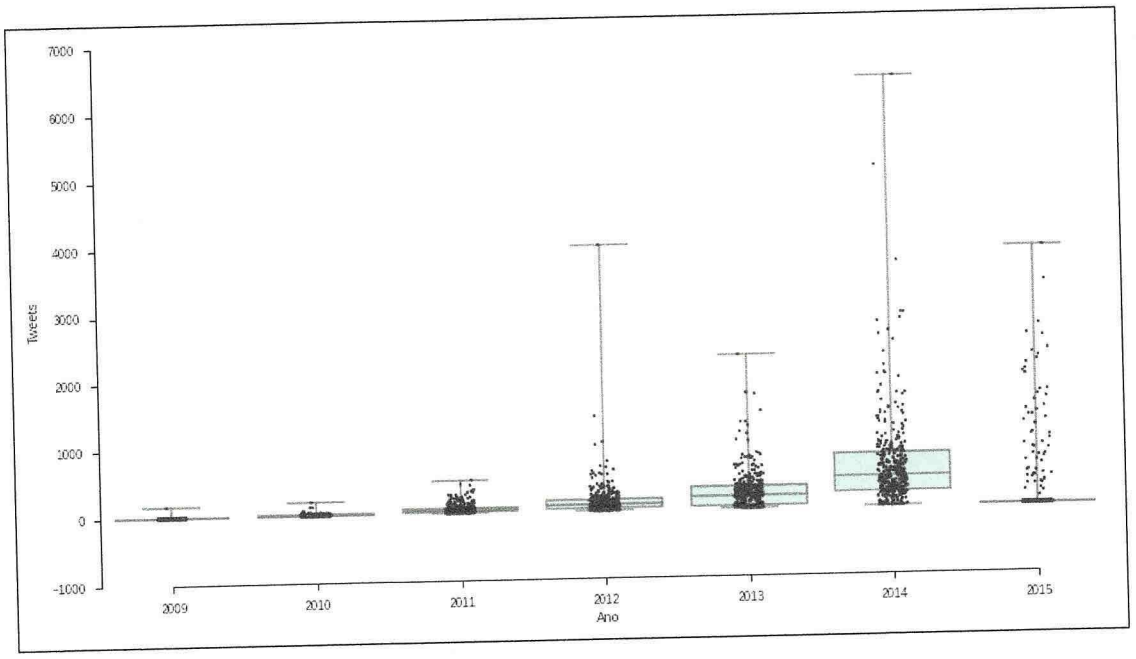

Figura C.15: Quantidade de compartilhamentos de notícias econômicas do jornal Estadão via Twitter ano a ano.

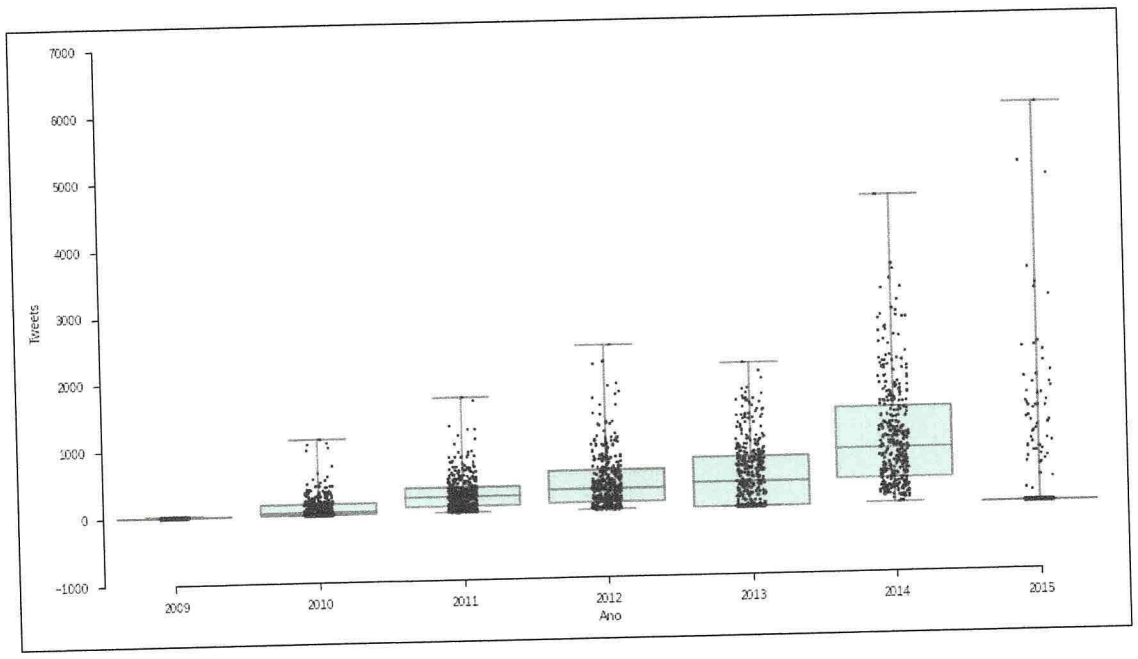

Figura C.16: Quantidade de compartilhamentos de notícias econômicas do jornal Folha de São Paulo via Twitter ano a ano.

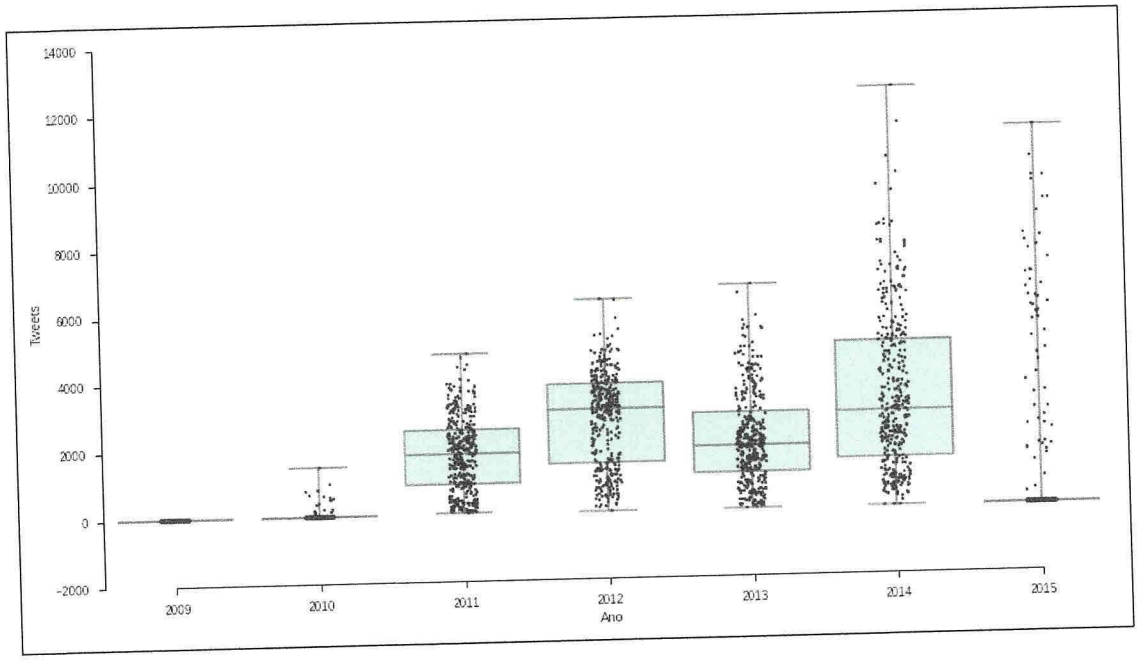

Figura C.17: Quantidade de compartilhamentos de notícias econômicas do jornal G1 via Twitter ano a ano.

C.4.1 Mês

As Figuras C.18, C.19 e C.20 apresentam a variabilidade de repercussão de notícias econômicas via Twitter ao longo dos meses do ano e confirmam em grande medida a afirmação anterior em especial para o jornal Folha de São Paulo onde há um gradual aumento das medianas. Por fim, o jornal G1 possui uma mediana de $\sim$ 1000 tweets por notícia o que é $\sim$10 vezes maior que a repercussão média dos outros jornais.

C.4.2 Dia de Semana

As Figuras C.21, C.22 e C.23 apresentam os gráficos de *bloxplot* para os jornais Estadão, Folha de São Paulo e G1 respectivamente. Duas características ficam evidentes: a mediana de todos os jornais mantidas em um mesmo patamar de Segunda à Sexta e uma diminuição nos finais de semana.

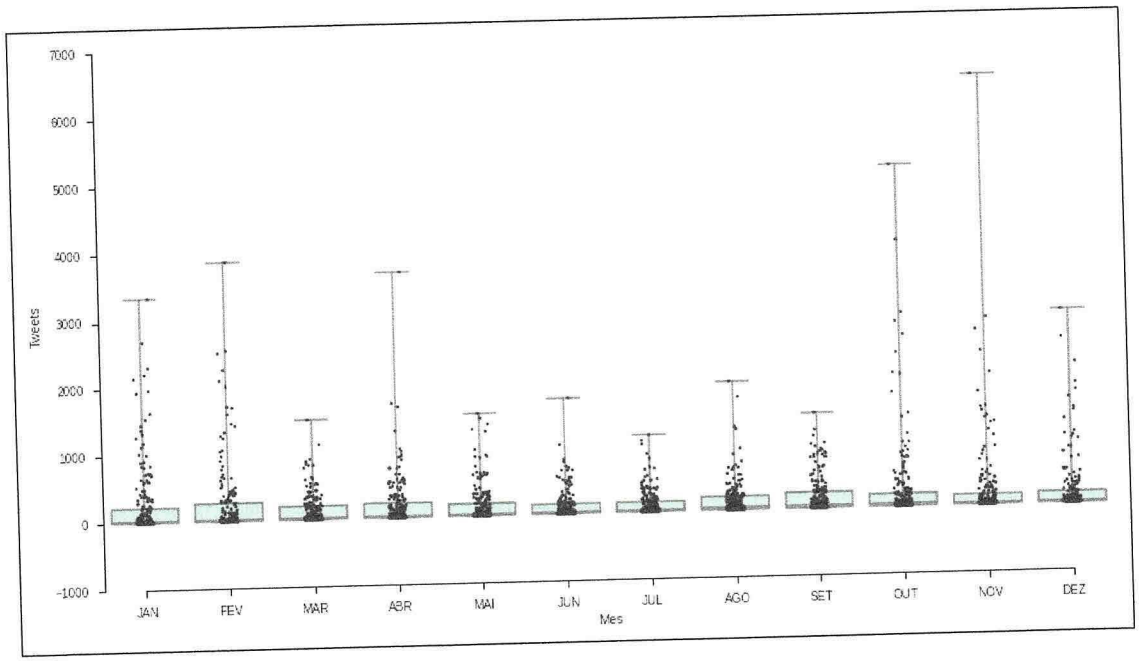

Figura C.18: Quantidade de compartilhamentos de notícias econômicas do jornal Estadão via Twitter mês a mês.

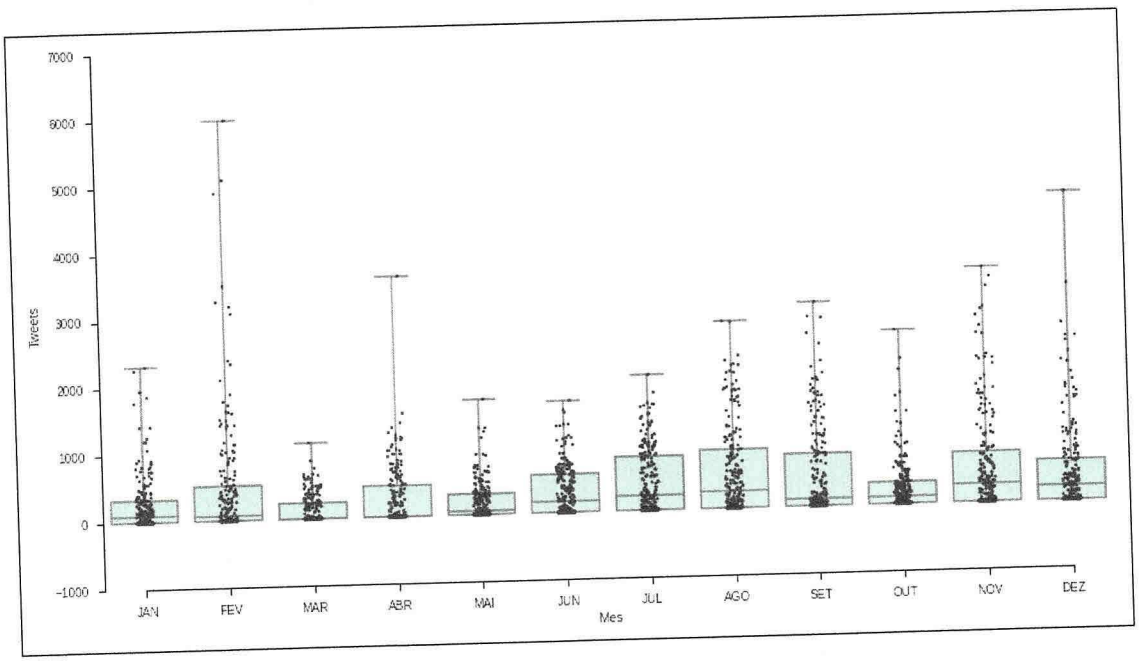

Figura C.19: Quantidade de compartilhamentos de notícias econômicas do jornal Folha de São Paulo mês a mês.

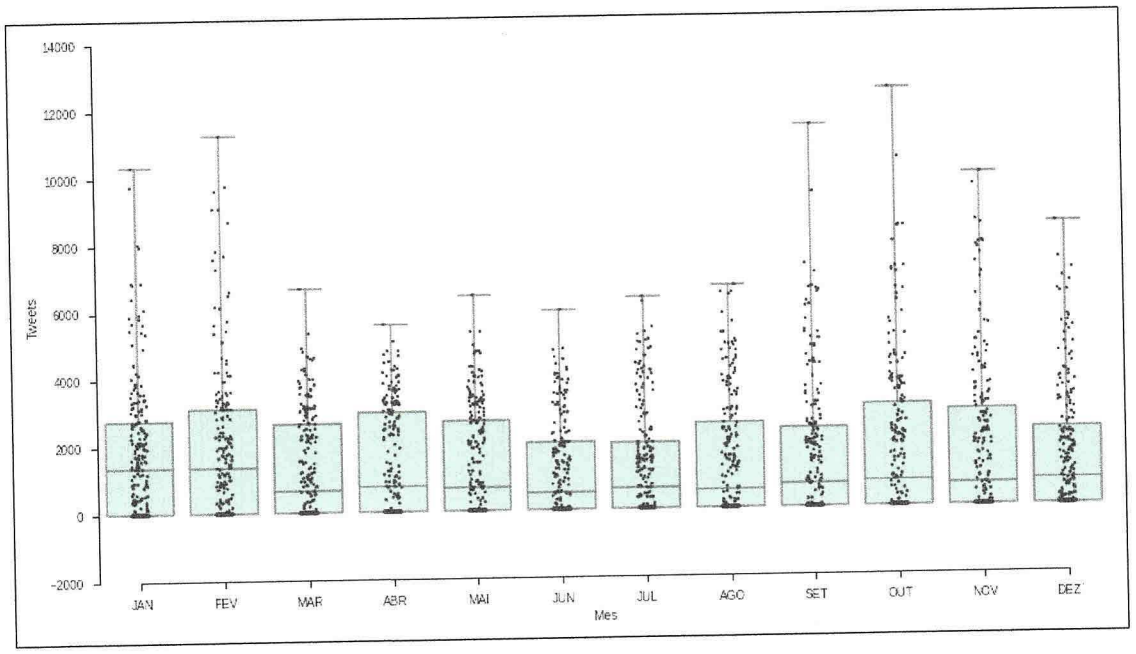

Figura C.20: Quantidade de compartilhamentos de notícias econômicas do jornal G1 via Twitter mês a mês.

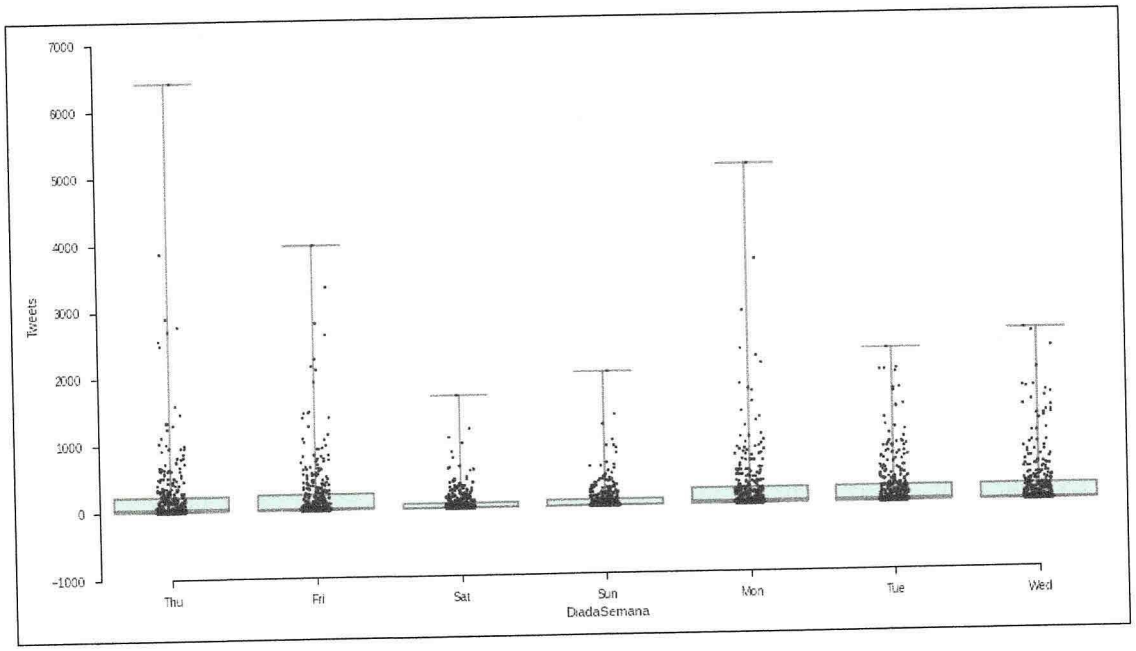

Figura C.21: Quantidade de compartilhamentos de notícias econômicas do jornal Estadão durante os dias da semana.

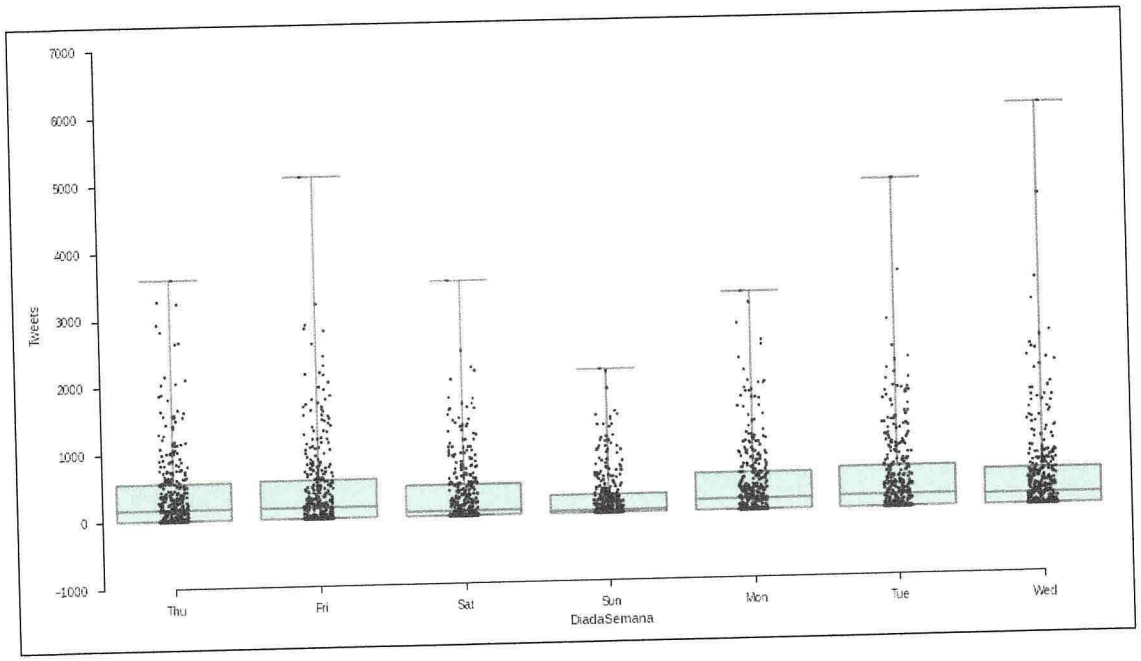

Figura C.22: Quantidade de compartilhamentos de notícias econômicas do jornal Folha de São Paulo durante os dias da semana.

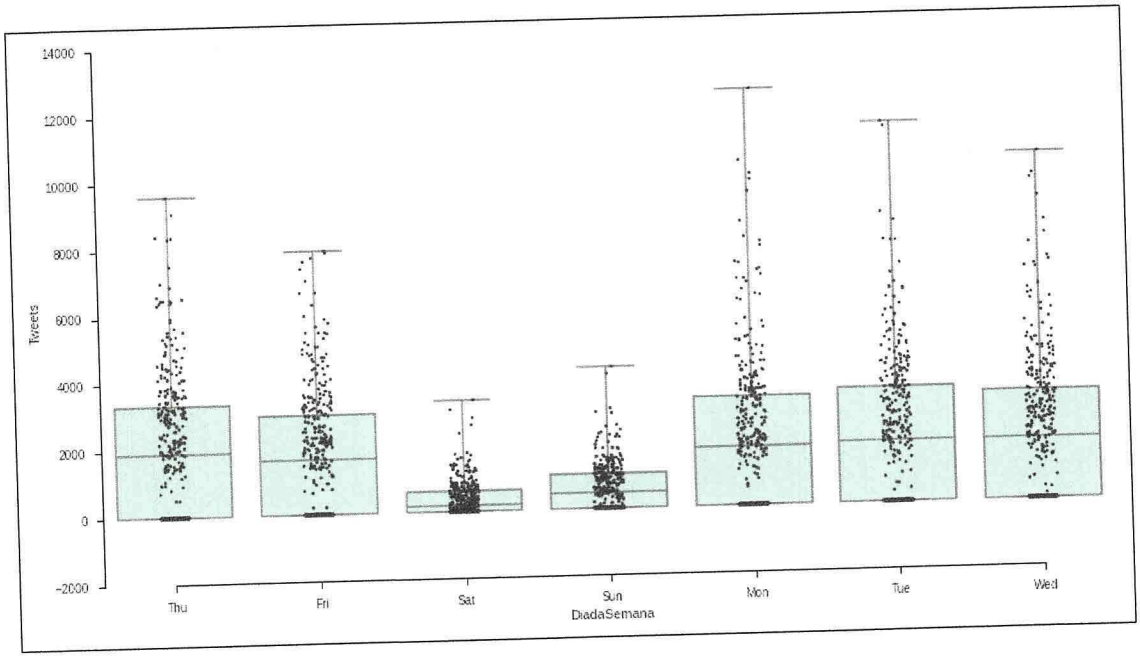

Figura C.23: Quantidade de compartilhamentos de notícias econômicas do jornal G1 durante os dias da semana.

C.5 Facebook

C.5.1 Ano

As Figuras C.24, C.25 e C.26 apresentam os gráficos de *bloxplot* para os jornais Estadão, Folha de São Paulo e G1 respectivamente. Em todos os gráficos percebe-se que algumas notícias atingem entre 350.000 e 450.000 compartilhamentos (para cada jornal há ao menos uma notícia com essa notoriedade), também é importante verificar que o crescimento do gráfico da série temporal não é dado em função de *outliers* e sim de uma participação constante dos leitores de notícias econômicas que muito provavelmente também são usuários do Facebook.

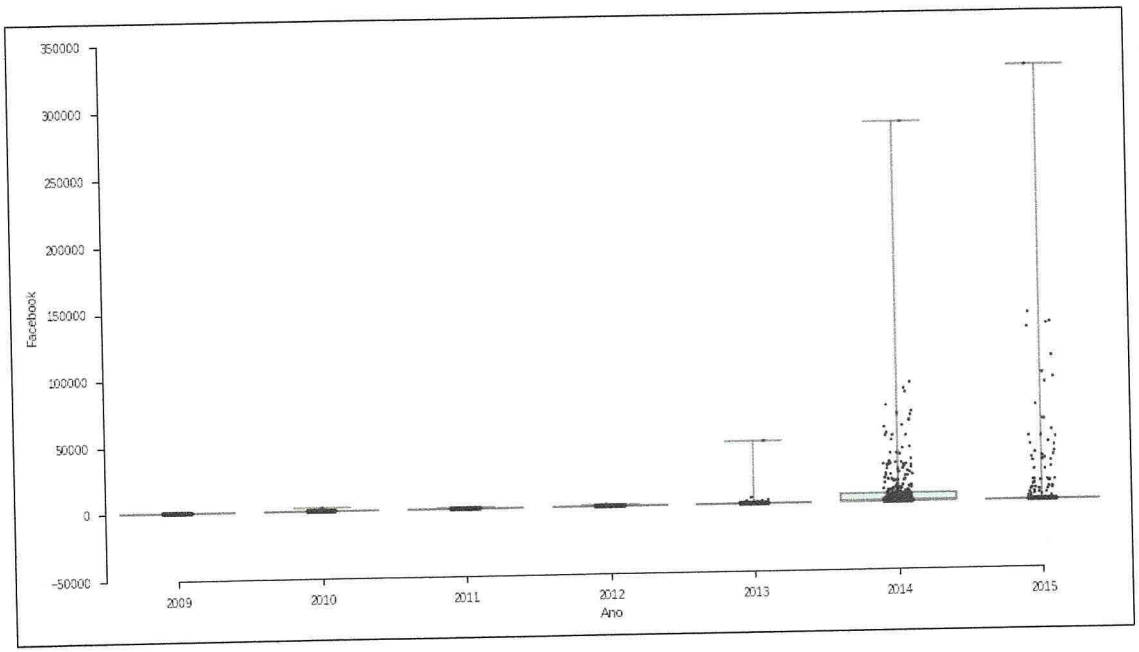

Figura C.24: Quantidade de publicações de notícias econômicas do jornal Estadão via Facebook ano a ano.

C.5.2 Mês

As Figuras C.27, C.28 e C.29 não apresentam *outliers* que expliquem essa preferência, sendo assim, é possível considerar esses picos como reflexos do comportamento dos leitores de notícias econômicas que também são usuários do Facebook.

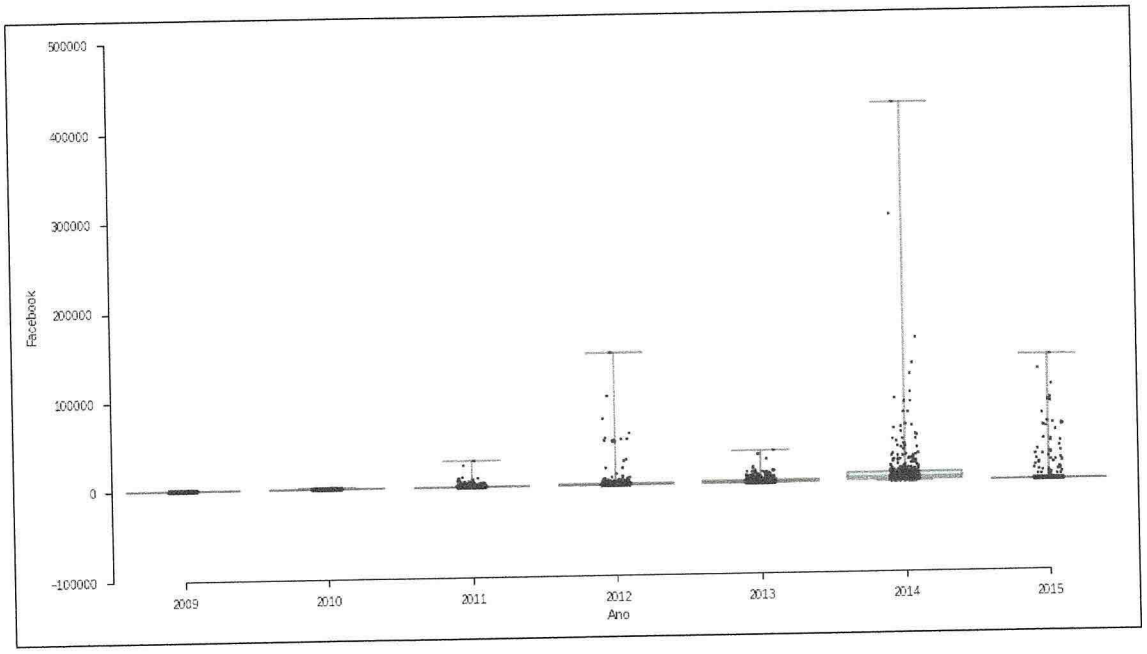

Figura C.25: Quantidade de publicações de notícias econômicas do jornal Folha de São Paulo via Facebook ano a ano.

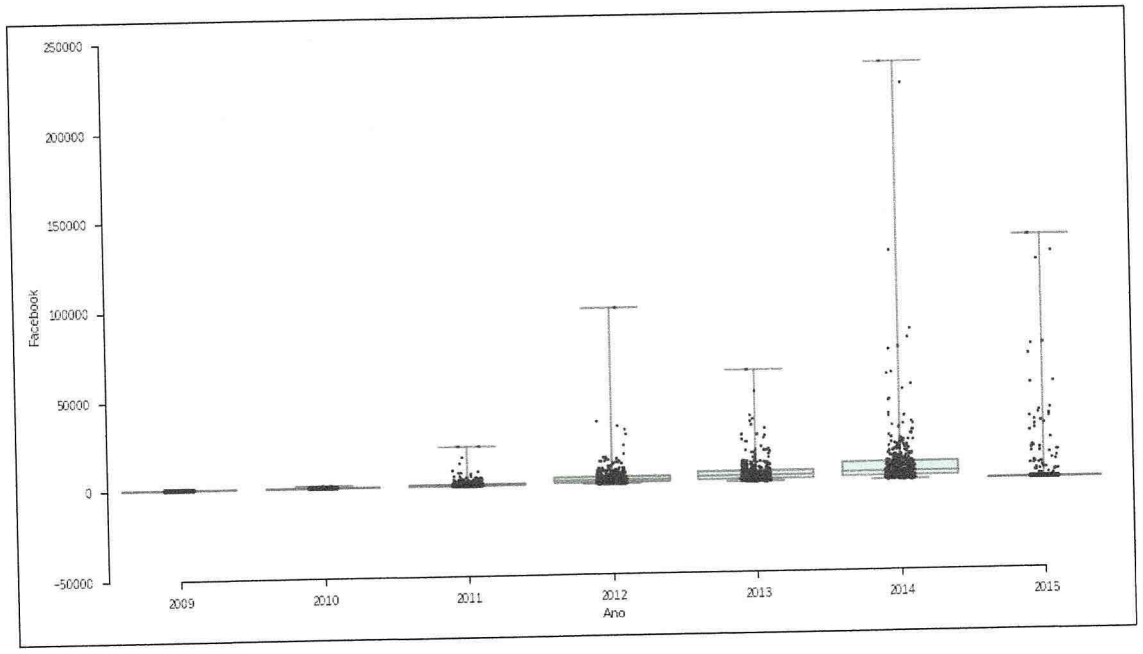

Figura C.26: Quantidade de publicações de notícias econômicas do jornal G1 via Facebook ano a ano.

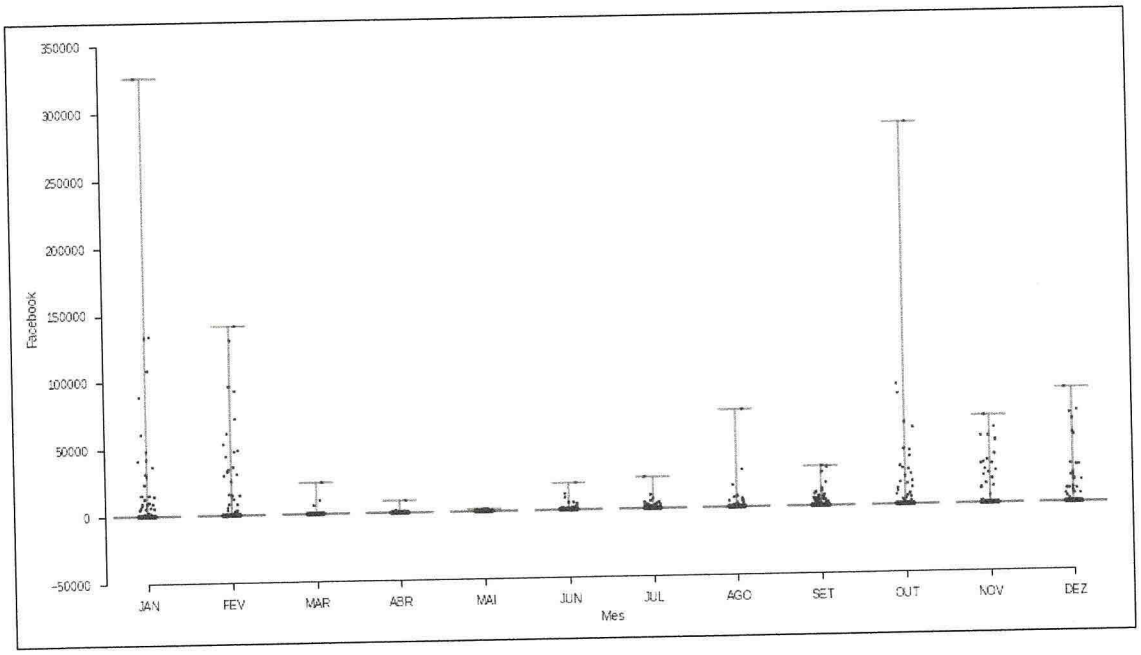

Figura C.27: Quantidade de publicações de notícias econômicas do jornal Estadão via Facebook mês a mês.

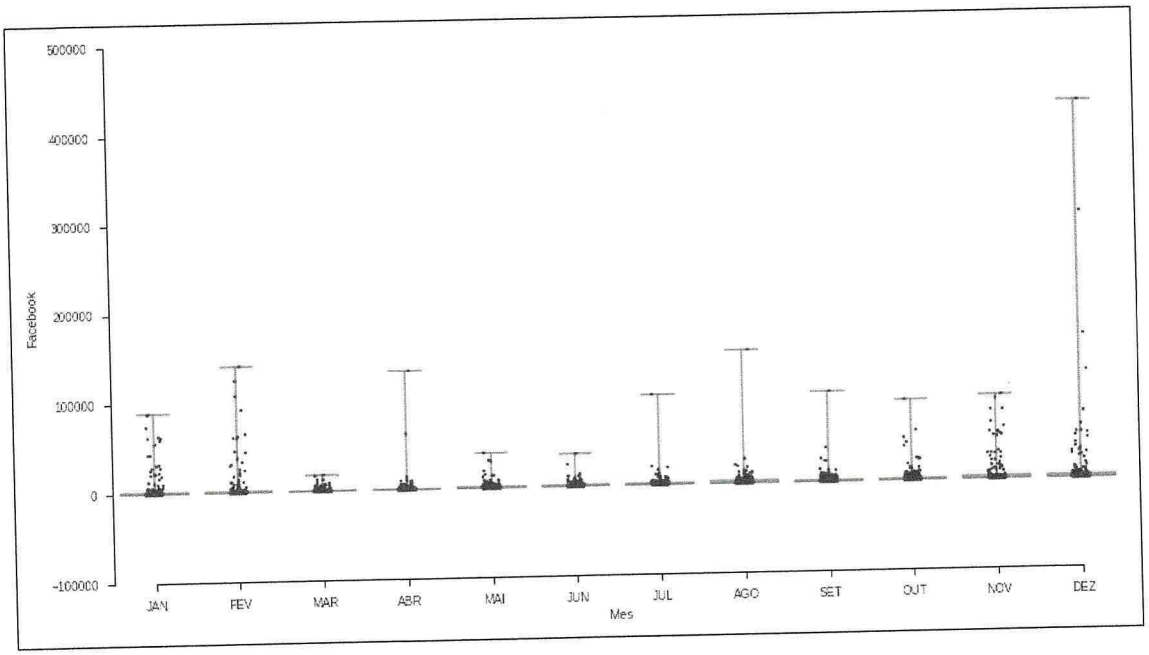

Figura C.28: Quantidade de publicações de notícias econômicas do jornal Folha de São Paulo via Facebook mês a mês.

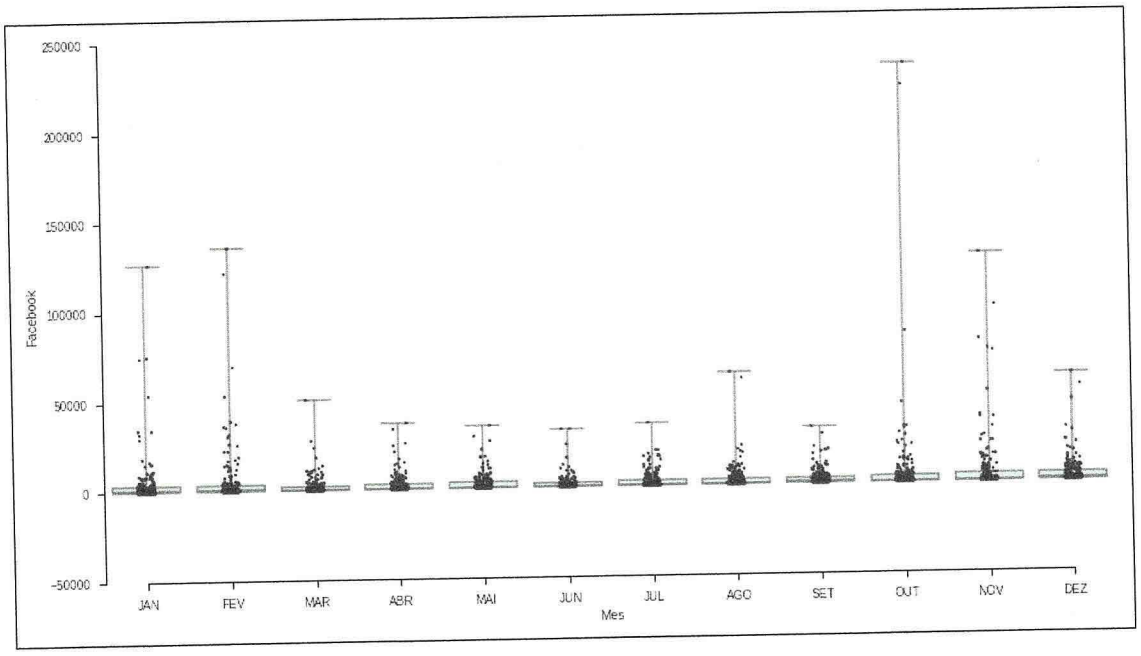

Figura C.29: Quantidade de publicações de notícias econômicas do jornal G1 via Facebook mês a mês.

C.5.3 Dia de Semana

As Figuras C.30, C.31 e C.32 apresentam a variabilidade do número de compartilhamentos de notícias econômicas via Facebook para os jornais Estadão, Folha de São Paulo e G1 respectivamente ao longo da semana e enfatiza que apesar de existirem alguns pontos bastante distantes da mediana, como a terça-feira do Estadão, a sexta-feira da Folha de São Paulo ou a quarta-feira do G1 eles não são suficientes para anular o efeito geral e confirmam o comportamento descrito anteriormente.

C.6 LinkedIn

C.6.1 Ano

As Figuras C.33, C.34 e C.35 apresentam a variabilidade para os jornais Estadão, Folha e G1 respectivamente. Para dois dos jornais, percebe-se que o aumento gradativo não é fruto de *outliers* e sim uma tendência natural de utilização da mídia.

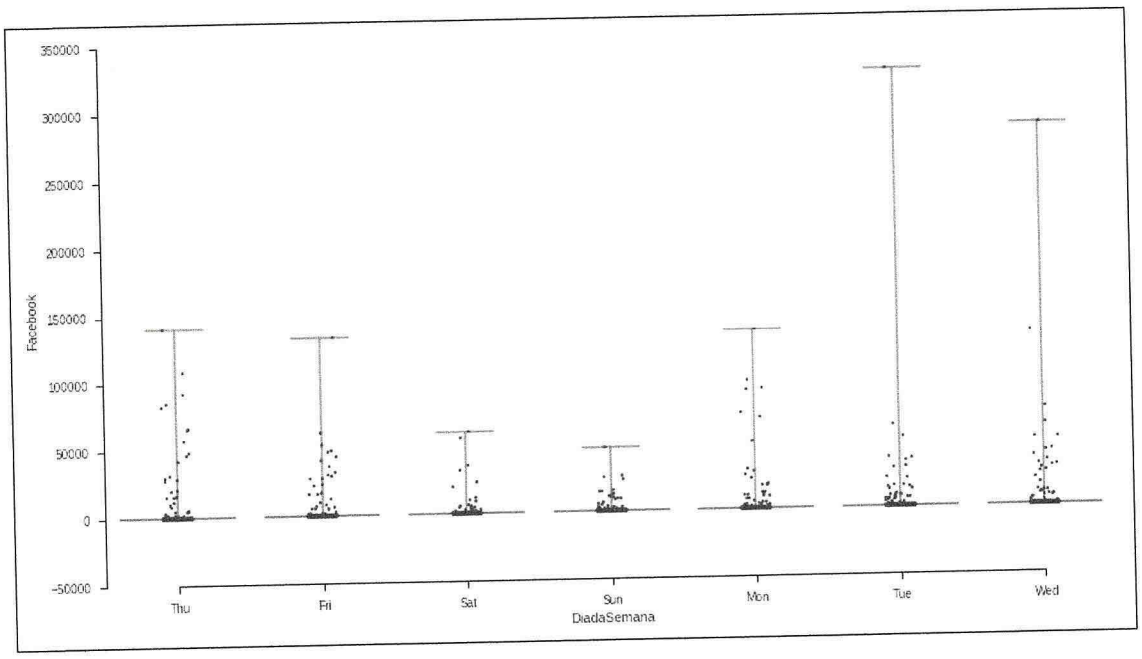

Figura C.30: Quantidade de publicações de notícias econômicas do jornal Estadão via Facebook ao longo dos dias da semana.

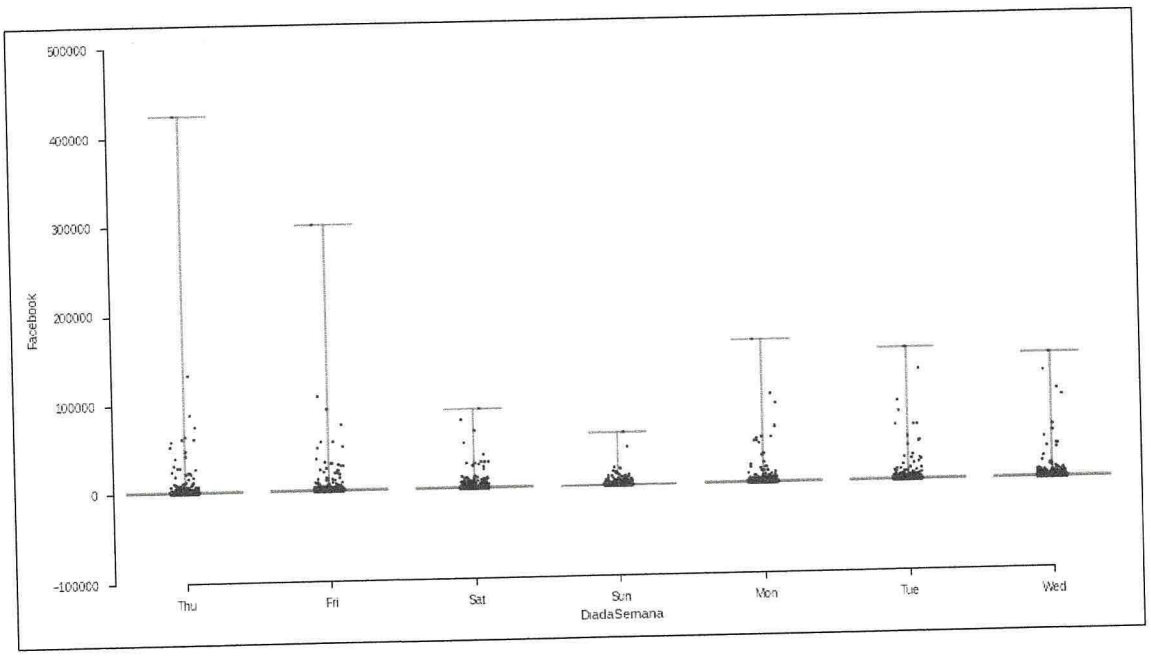

Figura C.31: Quantidade de publicações de notícias econômicas do jornal Folha de São Paulo via Facebook ao longo dos dias da semana.

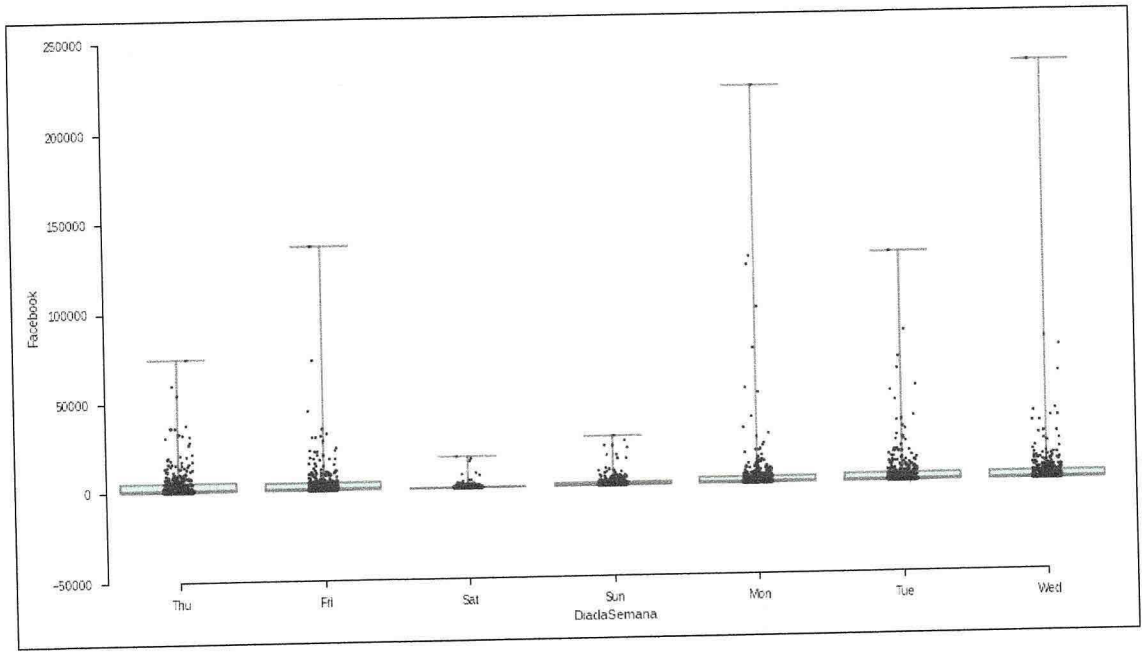

Figura C.32: Quantidade de publicações de notícias econômicas do jornal G1 via Facebook ao longo dos dias da semana.

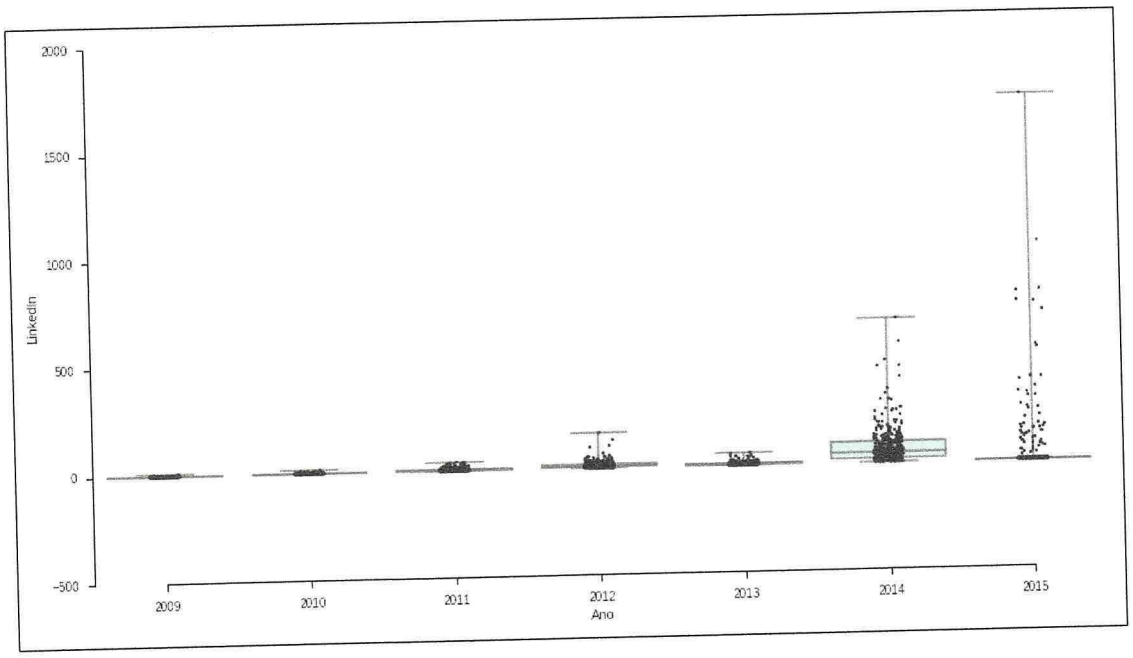

Figura C.33: Quantidade de publicações de notícias econômicas do jornal Estadão via LinkedIn ano a ano.

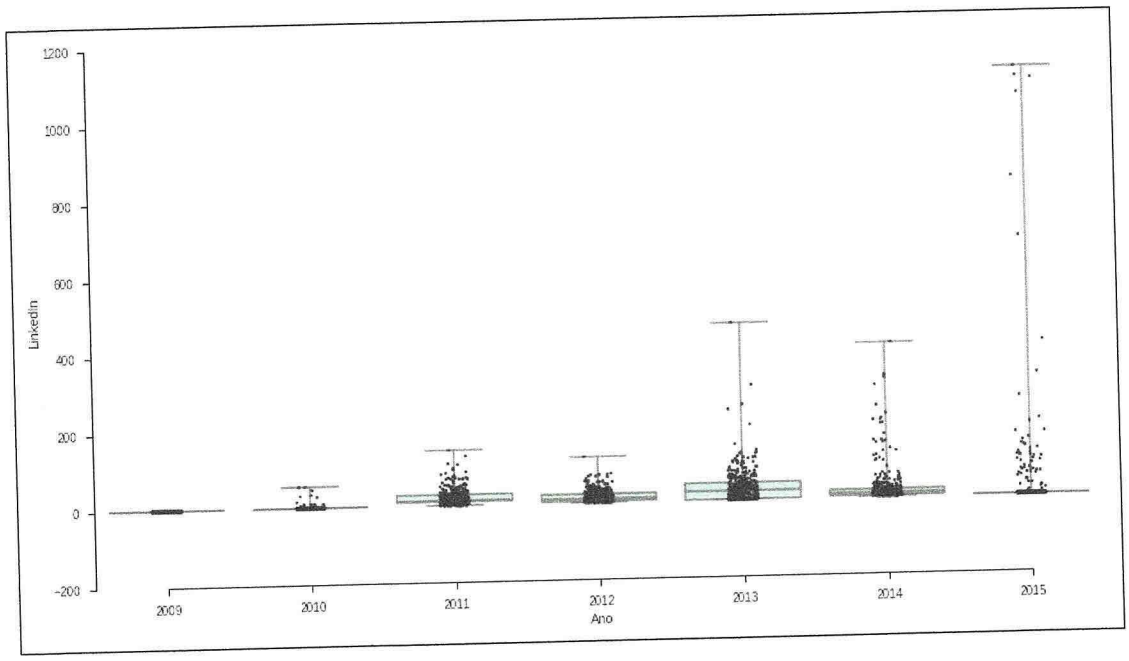

Figura C.34: Quantidade de publicações de notícias econômicas do jornal Folha de São Paulo via LinkedIn ano a ano.

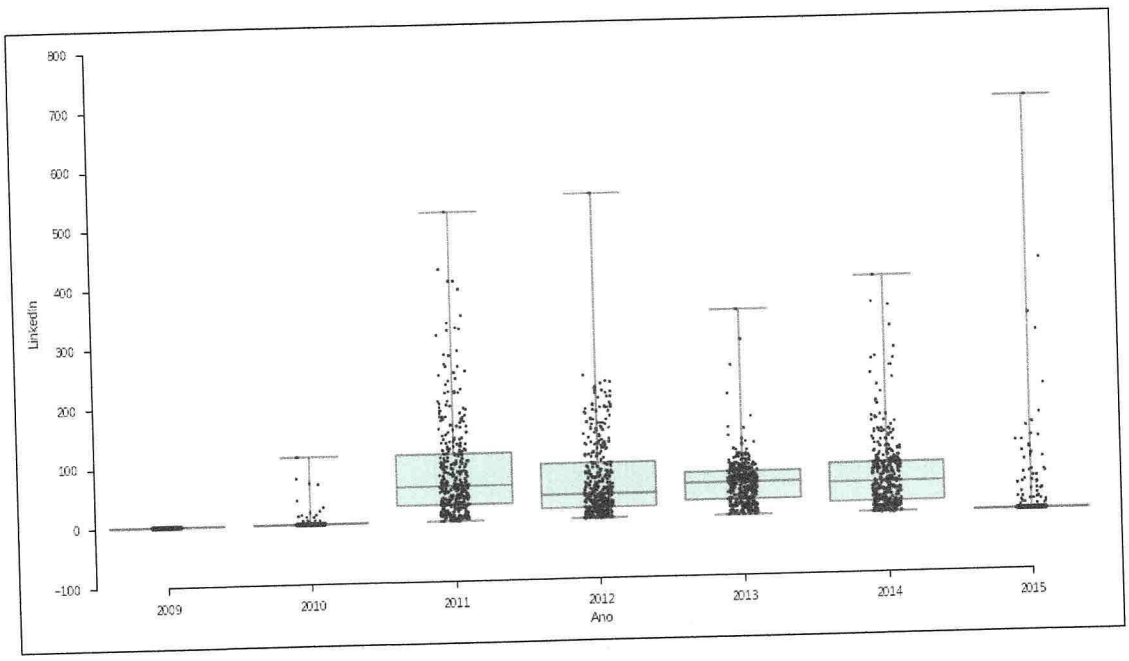

Figura C.35: Quantidade de publicações de notícias econômicas do jornal G1 via LinkedIn ano a ano.

C.6.2 Mês

As Figuras C.36, C.37 e C.38 apresentam a variabilidade do número de compartilhamentos de notícias econômicas via LinkedIn para os jornais Estadão, Folha de São Paulo e G1 respectivamente. Por estes gráficos percebe-se que Maio é o mês onde há maior variabilidade de publicações e Fevereiro a maior mediana o que também foi evidenciado para as outras mídias analisadas.

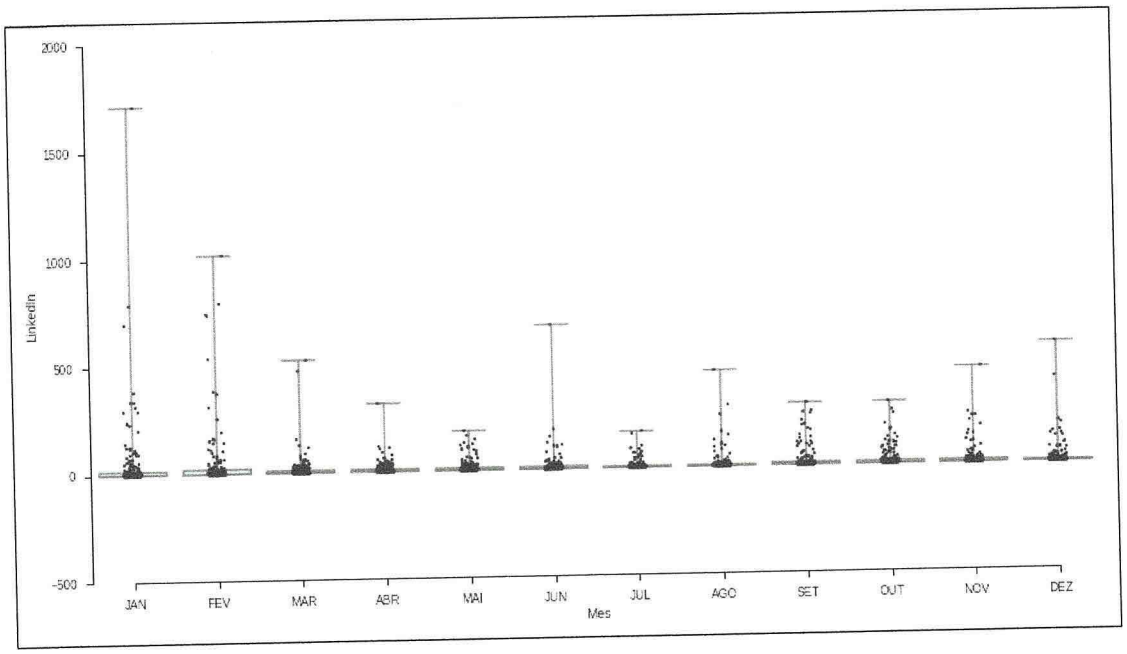

Figura C.36: Quantidade de publicações de notícias econômicas do jornal Estadão via LinkedIn mês a mês.

C.6.3 Dia

As Figuras C.39, C.40 e C.41 mostram a variabilidade no número de publicações econômicas via LinkedIn ao longo dos dias do mês para os jornais Estadão, Folha de São Paulo e G1 respectivamente. Com exceção do dia 27 para o jornal G1 há bastante regularidade quanto ao compartilhamento por dia do mês para todos os jornais.

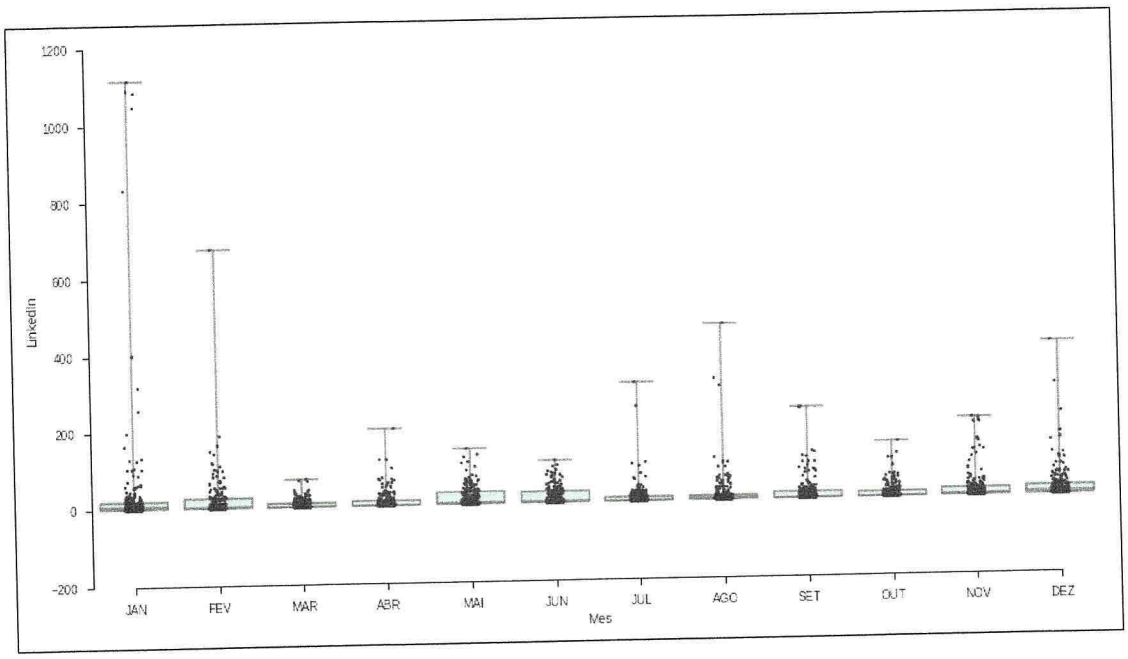

Figura C.37: Quantidade de publicações de notícias econômicas do jornal Folha de São Paulo via LinkedIn mês a mês.

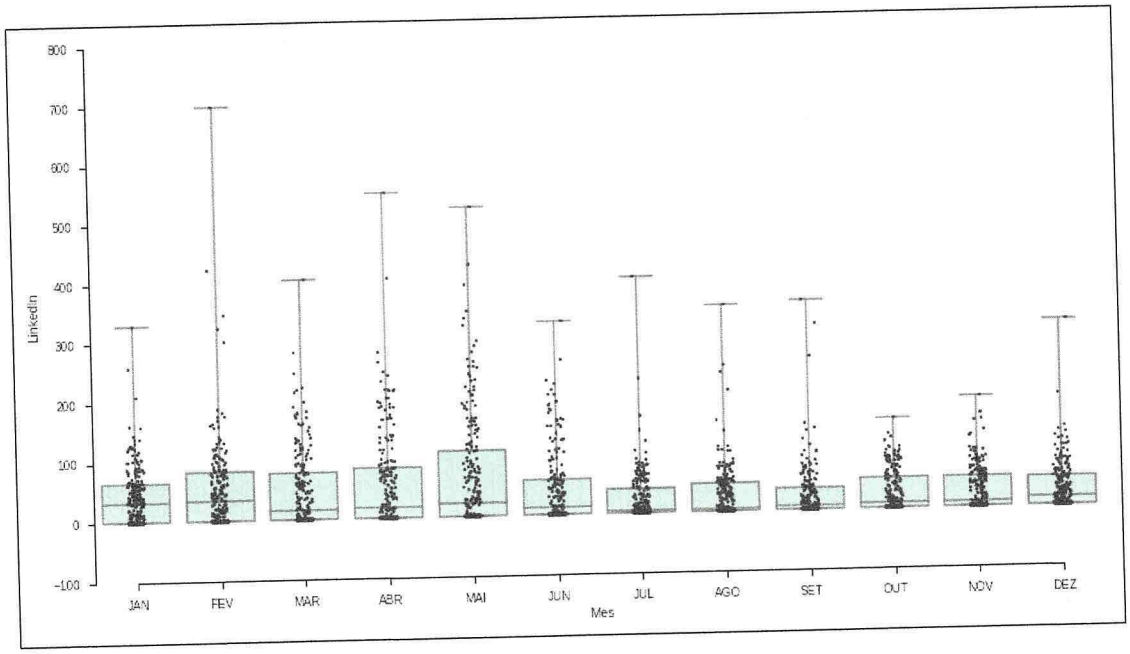

Figura C.38: Quantidade de publicações de notícias econômicas do jornal G1 via LinkedIn mês a mês.

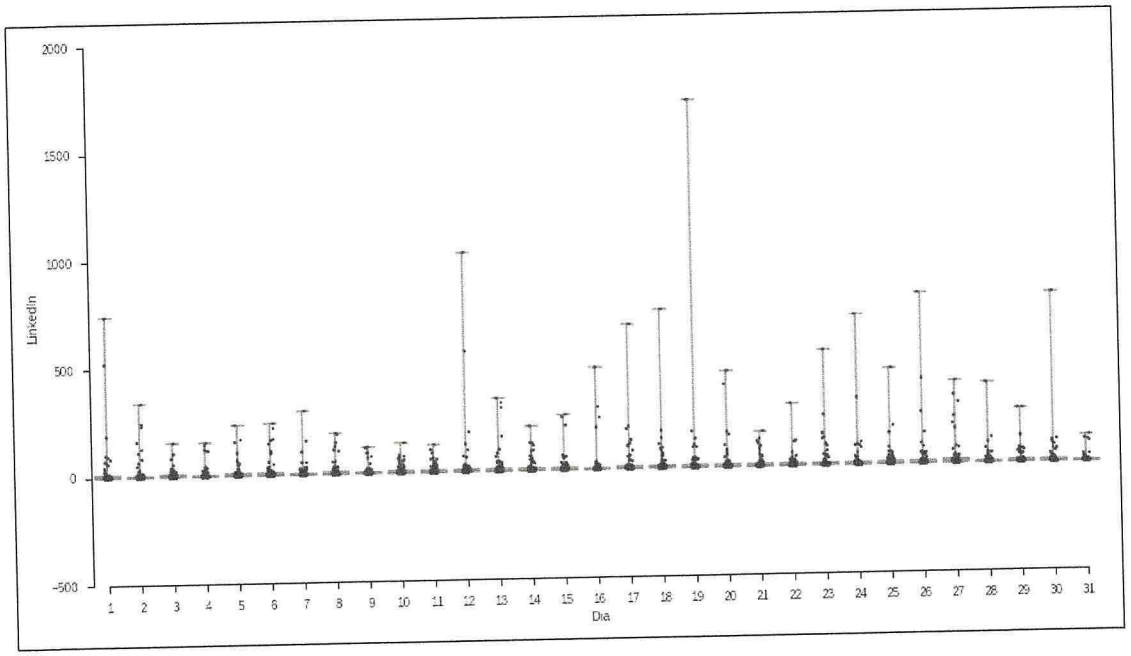

Figura C.39: Quantidade de publicações de notícias econômicas do jornal Estadão via LinkedIn ao longo dos dias do mês.

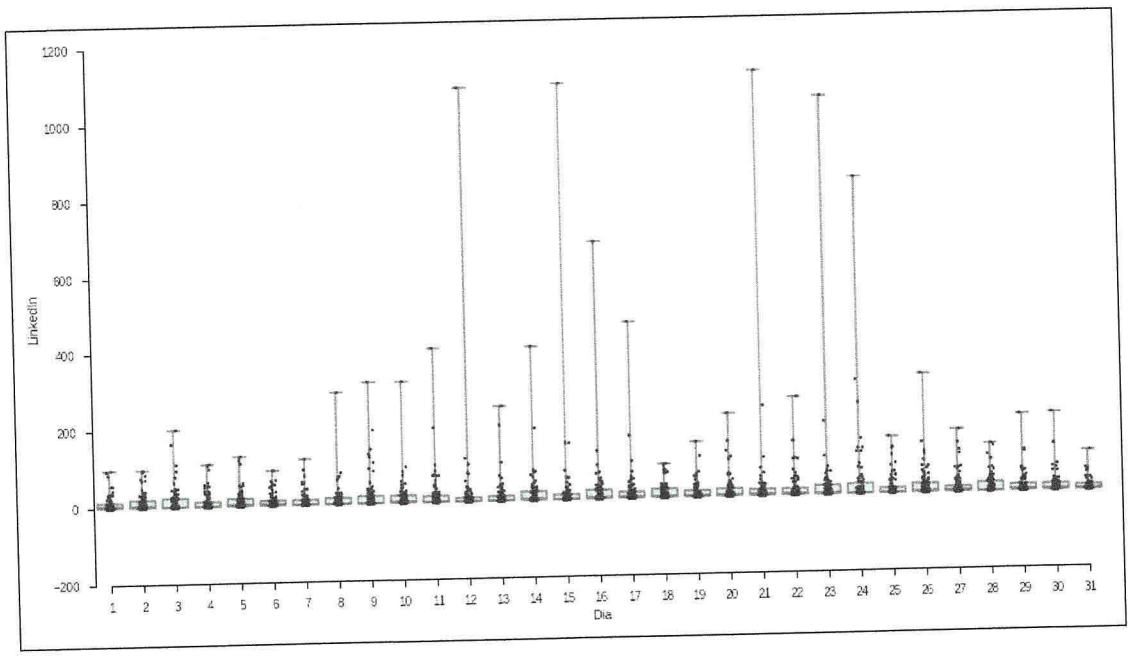

Figura C.40: Quantidade de publicações de notícias econômicas do jornal Folha de São Paulo via LinkedIn ao longo dos dias do mês.

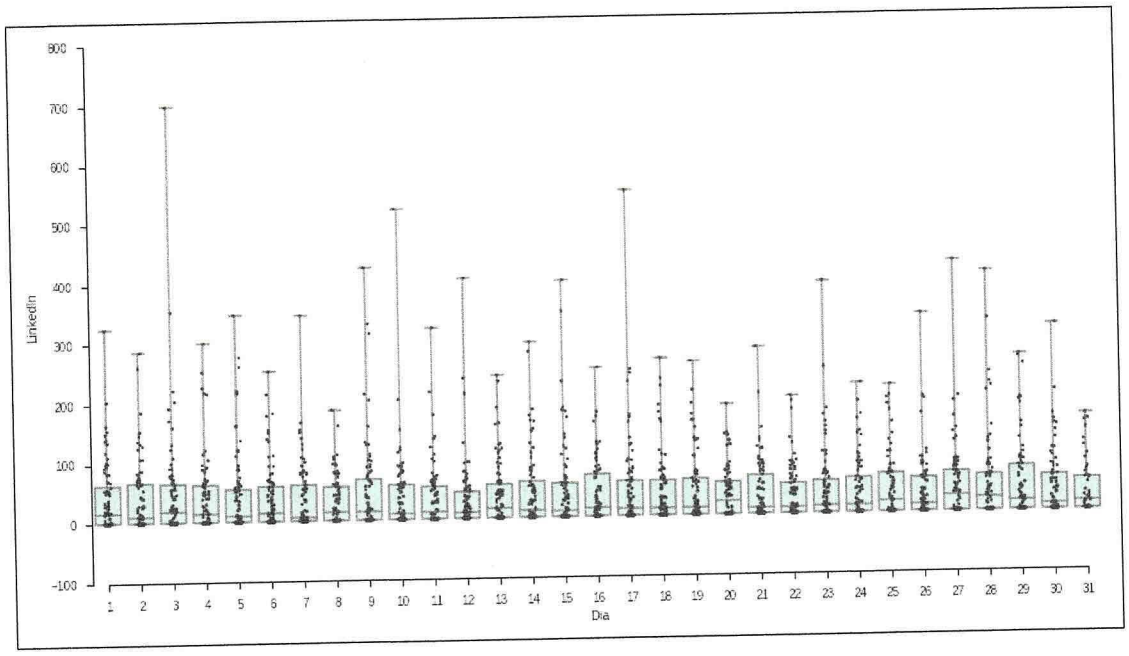

Figura C.41: Quantidade de publicações de notícias econômicas do jornal G1 via LinkedIn ao longo dos dias do mês.

C.6.4 Dia de Semana

As Figuras C.42, C.43 e C.44 apresentam a variabilidade dessas quantidades para os jornais Estadão, Folha de São Paulo e G1 respectivamente. Apesar dos *outliers* estarem na Segunda e Terça-feiras em todos os jornais a mediana é bastante regular entre os dias da semana.

C.7 Google Plus

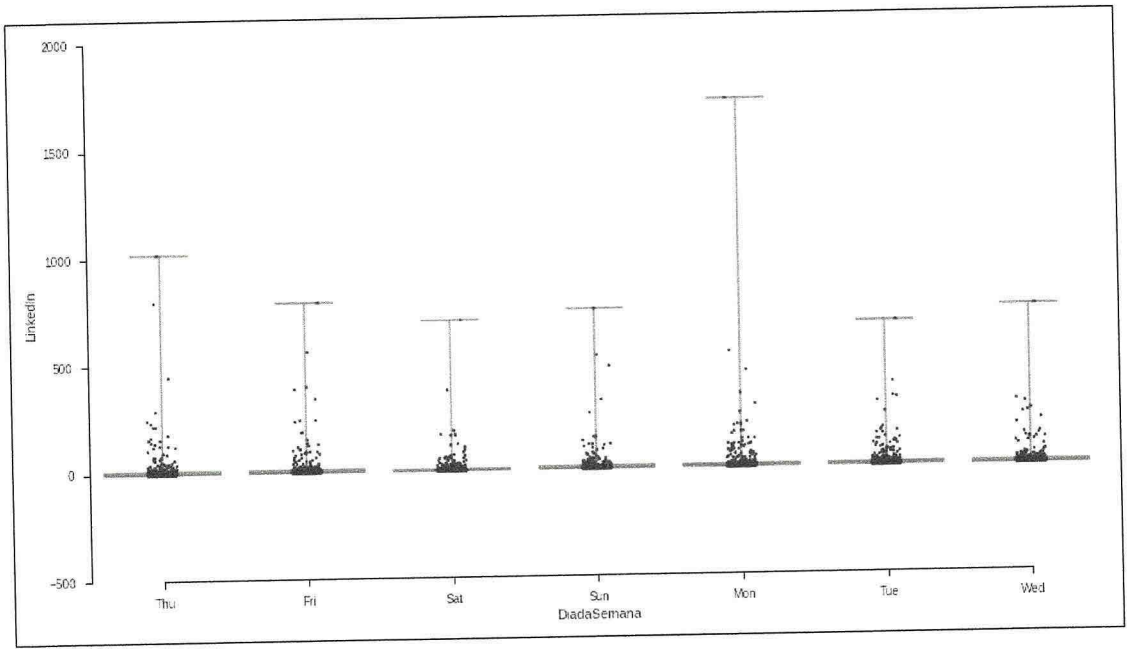

Figura C.42: Quantidade de publicações de notícias econômicas do jornal Estadão via LinkedIn ao longo dos dias da semana.

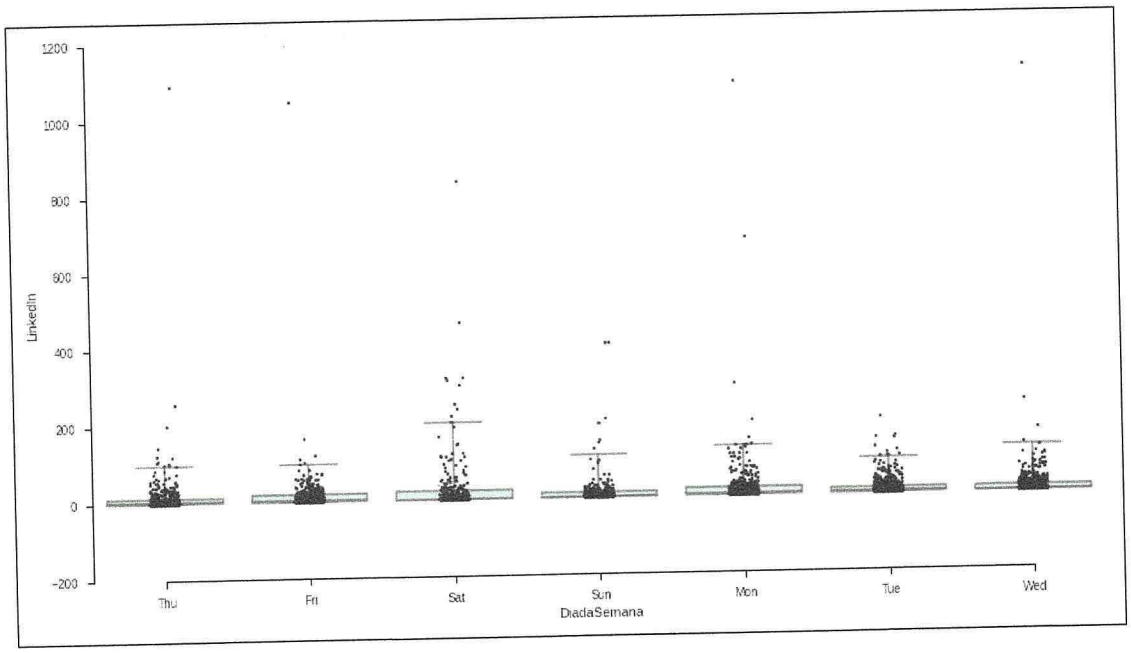

Figura C.43: Quantidade de publicações de notícias econômicas do jornal Folha de São Paulo via LinkedIn ao longo dos dias da semana.

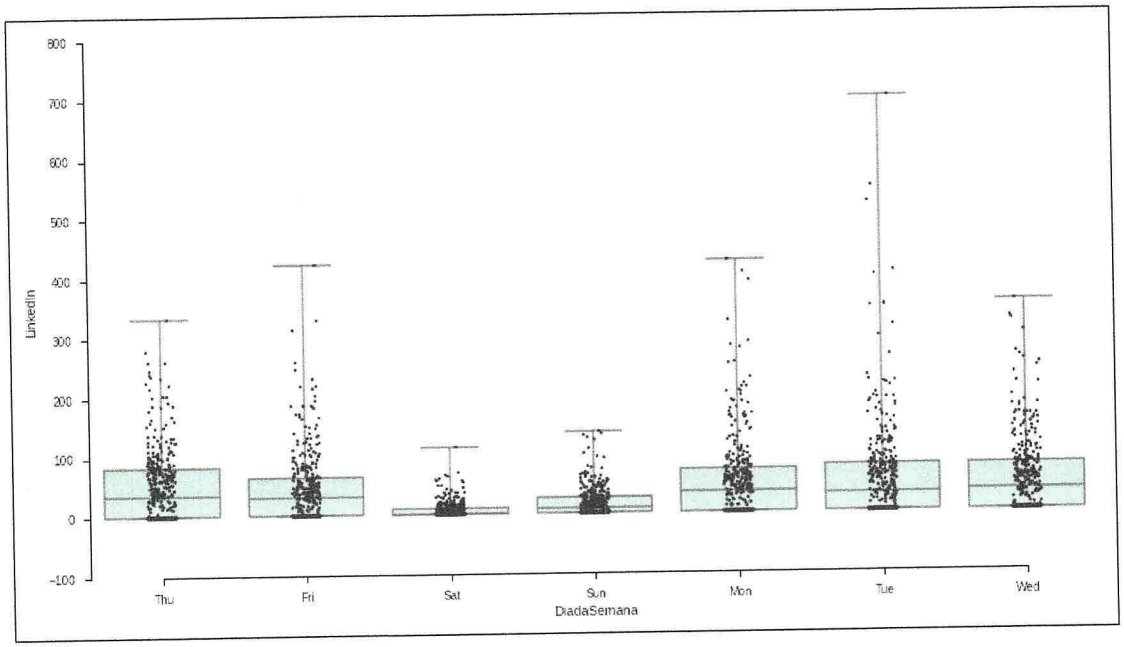

Figura C.44: Quantidade de publicações de notícias econômicas do jornal G1 via LinkedIn ao longo dos dias da semana.