

**NOTÍCIAS FINANCEIRAS E O CUSTO DA INFORMAÇÃO: A RELAÇÃO ENTRE
COMPLEXIDADE TEXTUAL E A DIFUSÃO DE INFORMAÇÃO NO MERCADO
FINANCEIRO**

RUAN RODRIGO ARAÚJO DA COSTA

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO RIO GRANDE DO NORTE (IFRN)

ANDERSON LUIZ REZENDE MÓL

UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE (UFRN)

RODRIGO RAPOSO DA FONSECA

UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE (UFRN)

NOTÍCIAS FINANCEIRAS E O CUSTO DA INFORMAÇÃO: A RELAÇÃO ENTRE COMPLEXIDADE TEXTUAL E AS DIFUSÃO DE INFORMAÇÃO NO MERCADO FINANCEIRO

1 INTRODUÇÃO

Como avaliar o valor de uma informação no mercado financeiro? Sob a hipótese de mercados eficientes, o preço já reflete toda a informação disponível. O problema é isso desconsidera a existência de um processo no qual interagem agentes com diferentes níveis de informação. Esse processo ocorre por meio de transações não síncronas. As ordens de negociação, quando lançadas por investidores com menor nível de informação, geram um estoque de volume negociável que pode ser estudado no contexto de precificação da informação (FAMA, 1970; AVDIS, 2016).

Ainda sobre a questão de como a informação é transmitida para os preços, a “Hipótese do Sapo na Panela” questiona o papel da sensibilidade e a atenção dos investidores no processo de transmissão da informação. A hipótese elenca que a velocidade de assimilação de uma informação depende de outros choques de informação ocorridos no passado. Em comum, ambas as abordagens consideram a velocidade de propagação da informação depende de esforço dos investidores (DA; GURUN; WARACHKA, 2014).

Sendo assim, este ensaio tem por base as premissas: notícias são marcos do processo informacional e a dificuldade de leitura delas é *proxy* para o custo de aquisição da informação. Estudos considerando o papel das notícias no contexto macroeconômico e no contexto de mercados financeiros tornam a primeira premissa razoável, a segunda premissa é mais sensível, todavia, a internet reduziu o esforço necessário para buscar notícias, restando para os investidores o trabalho de ler e interpretar as notícias, que estaria relacionado com a complexidade textual (BIRZ; LOTT JR., 2011; LILLO *et al.*, 2015; BAI; PHILIPPON; SAVOV, 2016).

Portanto, o presente trabalho tem por objetivo testar se o custo para adquirir informações financeiras, a partir de notícias, pode explicar a velocidade de assimilação de informação no mercado financeiro.

1.1 REVISÃO DA LITERATURA E HIPÓTESES

1.1.1 A relação entre a oferta de ativos e o custo da informação

Partindo da premissa que o mercado financeiro é dinâmico, o preço do ativo reflete a expectativa média dos investidores. Nesse contexto, as mudanças de preços são produtos de interações entre investidores com níveis de expectativa diferentes. A expectativa individual é formada a partir de um conjunto de informação específicas para cada investidor. Todavia, a informação não é estática, mas um fluxo contínuo e imprevisível, afetando as expectativas individuais, bem como o consenso de preços de mercado (GREEN, 1973; GROSSMAN; STIGLITZ, 1980).

Ao considerar o fluxo de informação como causador da mudança de preços a “Teoria de Aquisição de Informação em Sistemas Dinâmicos” preconiza que o custo em adquirir informação está relacionado à oferta dos ativos financeiros. A nível de mercado, quanto maior o custo em adquirir uma determinada informação menor será sua velocidade de propagação entre os agentes. Já individualmente, o incentivo em arcar com o custo para conhecer antecipadamente uma informação é poder consumir ofertas de investidores menos informados (CAMPBELL;

KYLE, 1993; WANG, 1993; AVDIS, 2016).

As transações consomem a oferta estocástica de volume negociável. É por meio dessas transações que ocorre o ajuste de expectativas e transmissão de informações para o preço dos ativos. Assim, o comportamento da oferta estocástica de volume negociável sinalizaria a quantidade de investidores informados, bem como a velocidade de ingresso de novas informações. Em um sistema onde novas informações representam movimentações significativas, as variáveis envolvidas na formação do preço de consenso podem servir para otimizar a eficiência de mercado de forma geral (AVDIS, 2016; CAMPBELL; KYLE, 1993; WANG, 1993).

Considerando que notícias como transmissoras de informação, vale destacar como elas se tornaram amplamente acessíveis *on-line*. Neste caso o custo de acesso a informação estaria no esforço para processar a informação e interpretar as informações textuais contidas nos artigos jornalísticos. Essa relação entre custo da informação e complexidade textual já foi abordada em outros estudos dentro do contexto de informações financeiras. No entanto, sua abordagem no estudo das notícias financeiras ainda não foi explorada. Sendo assim, a presente pesquisa considera a complexidade textual de notícias financeiras como *proxy* para o custo de aquisição da informação em notícias financeiras (LI, 2008; DA; GURUN; WARACHKA, 2014; LILLO *et al.*, 2015; GUAY; SAMUELS; TAYLOR, 2016). Então, A hipótese 1 preconiza que quanto maior o custo de uma informação menor será sua velocidade de propagação:

Hipótese 1: *Notícias mais simples geram maiores volumes de transação*

Além disso, o estudo busca entender como o efeito da complexidade se relaciona ao efeito do sentimento textual no processo informativo das notícias financeiras. assumindo que a complexidade pode afetar a velocidade de propagação do mercado, foi observado que fatos complexos podem necessitar de vários artigos para serem explorados pela mídia. Isso indica que a complexidade textual pode afetar o mercado por dias ou meses, enquanto persistem questões sobre o fato que motivou a existência da notícia. Em relação ao sentimento textual, as pesquisas denotam que o sentimento parece ser importante no curto prazo, vale ressaltar que não existe consenso quanto a duração do efeito. Contudo, os estudos consideram que o efeito é mais intenso no curto prazo (GUAY; SAMUELS; TAYLOR, 2016; FOUCAULT; HOMBERT; ROŞU, 2016; CERCHIELLO; NICOLA, 2018; LEE *et al.*, 2020).

A hipótese 2 foi elaborada para abordar a lacuna teórica descrita anteriormente, ela o faz analisando de forma comparativa o efeito da complexidade textual e do sentimento para o processo de transmissão de informação ao mercado.

Hipótese 2: *A complexidade das notícias afeta o volume negociado por mais tempo do que o sentimento textual*

1.1.2 A transmissão do custo da informação para os preços

Segundo Avdis (2016), as negociações entre agentes informados e não informados são responsáveis pelas alterações no equilíbrio dinâmico do mercado, é por meio dessas interações que os preços oscilam e se tornam fonte de informação para os agentes menos informados. De acordo com a “Hipótese do Sapo na Panela”, uma das informações que o preço pode fornecer está no seu histórico recente de oscilações, que seria uma sinalização para o nível de sensibilidade e atenção do mercado no presente para ingressos futuros de informação (DA; GURUN; WARACHKA, 2014).

Se o histórico de preços absorve informação pela ação dos agentes informados, resta entender em qual momento as notícias se encaixam nesse processo. Ao entendermos notícias como ingressos de novas informações, a intensidade desse ingresso poderia ser explicada pelo custo de processar o conteúdo presente nas notícias. Os indícios de que notícias são marcos do ingresso de informação no mercado são dados pelo trabalho de [Engelberg, McLean e Pontiff \(2018\)](#), nessa pesquisa, o comportamento posterior dos investidores é parcialmente explicado pela ocorrência de notícias prévias.

Ainda sobre o poder das notícias na formação dos preços, ela não são vistas como eventos únicos. A publicação de um conjunto de notícias relacionadas e em sequência parece afetar a absorção da informação. Ou seja, notícias podem formar narrativas que são processadas pelo mercado como uma narrativa conjunta ([CERCHIELLO; NICOLA, 2018](#)).

Pelo que foi discutido até então, evidencia-se uma lacuna: Qual o efeito da complexidade textual na sensibilização do mercado para novas informações? ([DA; GURUN; WARACHKA, 2014](#); [FOUCAULT; HOMBERT; ROŞU, 2016](#)) A hipótese 3, então, foi construída para testar essa questão.

Hipótese 3: *A informatividade dos preços é inversamente relacionada ao custo em adquirir informação*

1.2 METODOLOGIA

1.2.1 Amostra, coleta e tratamento dos Dados

A amostra deste trabalho foi elaborada tomando como base as ações listadas no índice IBRx 100 em 01/02/2020. O índice foi selecionado pela sua representatividade. Na data, o índice era composto por 100 ativos que representavam 93 empresa diferentes, no caso de empresas com mais de uma representação, apenas as ações ordinárias foram utilizadas.

O recorte temporal foi de janeiro de 2012 até março de 2020, sendo definido pela disponibilidade de notícias nos arquivos dos portais. As notícias foram obtidas através de *web scrapping*, com o uso do software Python. Foram considerados três portais de notícias: Valor Econômico, Exame e Infomoney, que foram escolhidas por critérios de acesso e relevância. A partir dos artigos jornalísticos foram criadas variáveis que representam a complexidade textual, o sentimento textual e o número de artigos no período.

Para análise dos efeitos das notícias sobre o mercado, primeiro foi necessário vincular as notícias com as empresas da amostra. Nesse procedimento, foi definido que uma notícia era relacionada a uma ou mais empresas quando aquela notícia citava o nome da empresa, da marca ou de uma subsidiária, vale destacar que algumas notícias não puderam ser relacionadas com nenhuma empresa e foram descartadas. A partir das datas de publicação e da vinculação apresentada nesse parágrafo foi obtida a frequência de notícias por empresa, por período.

Cada notícia teve seu conteúdo textual analisado e recebeu uma classificação quanto ao sentimento textual. Esse índice varia entre 1 e -1 indo de totalmente positivo para totalmente negativo, conforme [Feuerriegel e Proelochs \(2019\)](#). O dicionário utilizado para classificar palavras como positivas e negativas foi o de [Silva \(2017\)](#). A equação 1 mostra como foi obtido o valor.

$$sentimento = \frac{(P - N)}{T} \quad (1)$$

Em que:

- *Sentimento* = é o índice de sentimento textual de cada texto;
- *P* = O número de palavras classificadas como positivas no texto;
- *N* = O número de palavras classificadas como negativas no texto;
- *T* = O número total de palavras;

Também foi calculada a complexidade textual, para tanto foi utilizado o índice de facilidade de leitura de [Flesch \(1948\)](#) calculado como descrito na equação 2. Textos com menores pontuações são mais complexos.

$$Flesch = 206,835 - (1,015 \times ASL) - \left(84,6 \times \left(\frac{N_{sy}}{N_w} \right) \right) \quad (2)$$

Em que:

- *ASL* = A média de palavras por sentença;
- *N_w* = O número de palavras;
- *N_{sy}* = O número total de sílabas;

Além das variáveis criadas a partir dos artigos jornalísticos, foram utilizadas variáveis financeiras e contábeis, obtidas a partir da plataforma de dados Thomson Reuters Eikon®.

Os dados foram organizados por dia e por empresa. As variáveis foram sincronizadas pelas datas de publicação das notícias com os dados de mercado, a exceção foi para os casos de notícias publicadas após do horário de pregão ou em dia não útil, nesses casos sempre foi considerado o próximo dia útil. Ocorreram casos que uma empresa era citada por mais de uma notícia no mesmo dia, nessas situações, a média simples diária das variáveis de complexidade e de sentimento foi considerada, tal escolha se justifica pois foi observado que grupos de notícias podem influenciar o mercado na forma de narrativas que são necessárias para relatar temas complexos aos investidores ([CERCHIELLO; NICOLA, 2018](#); [LEE et al., 2020](#)).

1.2.2 Descrição das Variáveis

A variável *volume* indica o consumo da oferta estocástica de volume negociável, prevista pela “Teoria de Aquisição de Informação em Sistemas Dinâmicos” e representa o consumo de ofertas durante o período de ingresso de novas informações. O volume transacionado sinaliza a demanda por negociação dado um novo ingresso de informação ([AVDIS, 2016](#); [BROMAN; SHUM, 2018](#)). As variáveis utilizadas no trabalho são mostradas na tabela 1.

A variável id representa a informação discreta presente nos preços de um ativo em determinada data, o nível de informação discreta de um ativo está relacionado aos choques informacionais ocorridos no passado, quanto maior a volatilidade no ano anterior, maior será o id e mais sensível o preço do ativo para novas informações. id pode assumir valores que vão de -1 até 1, valores positivos denotam informação discreta e valores negativos denotam informação contínua, de baixa volatilidade e que possui pouca atenção do mercado (DA; GURUN; WARACHKA, 2014). Na equação 5 id foi calculada com frequência diária, já na equação 6 variável denota frequência trimestral, neste caso ela é calculada tomando o último dia do trimestre fiscal como data de referência, a variação idf também é usada em frequência trimestral. A equação é detalhada em 3.

$$id_{i,t} = sng(PRET) \times (\%NEG - \%POS) \quad (3)$$

Em que:

- $sng(PRET)$ = Sinal (sng) do retorno acumulado nos últimos 250 dias (12 meses), excluindo os últimos 22 dias (cerca de um mês) ($PRET$);
- $\%NEG$ = Percentual do número de dias com retornos negativos;
- $\%POS$ = Percentual do número de dias com retornos positivos.

A variável $nunot12m$ representa o número notícias por empresa durante um ano, tem a finalidade de servir de controle para os efeitos da atenção da mídia (DA; GURUN; WARACHKA, 2014).

$flesch_{it}$ são o conjunto de variáveis que indicam a complexidade textual, quanto maior o valor, mais complexo o texto. Foram criadas três variáveis que consideram o valor médio da complexidade textual das notícias no período ($fleschmd$, $fleschPD$ e $Mflesch12m$) e três que consideram o valor agregado da complexidade ($flesch12m$, $flesch6m$ e $flesch1m$), utilizar o valor agregado visa captar a ocorrência de eventos informacionais complexos, esses eventos seriam vistos pelos investidores como uma narrativa contada por uma série de artigos sobre um mesmo tema. Nesta pesquisa as variáveis baseadas no índice de complexidade são usados como indicação do custo de aquisição de informação pelos investidores (FLESCH, 1948; FEDYK, 2018; LEE *et al.*, 2020).

O grupo de variáveis que expressam o sentimento textual é formado por seis variáveis diferentes: $sentmd$ que representa o sentimento médio das notícias publicadas no mesmo dia; $sentpd$ que representa o efeito do sentimento médio das notícias no dia posterior a sua publicação; $sent12m$, $sent6m$ e $sent1m$ que expressam respectivamente o efeito do sentimento acumulado para 1 ano (252 dias de negociação), 6 meses (151 dia) e 1 mês (20 dias); e a variável $Msent12m$ representa o sentimento médio das notícias publicadas durante um ano (RANCO *et al.*, 2015; SILVA, 2017; LEE *et al.*, 2020).

As outras variáveis usadas representam dados financeiros e contábeis, elas são controles na equação 6 do modelo econométrico utilizado. A variável $\Delta turnover$ é a variação no giro das ações entre dois trimestre consecutivos; a variável $\Delta analyst$ é a variação do número de analistas *sell-side* atuantes na avaliação da empresa e indica a atenção dedicada por investidores

Tabela 1 – Tabela com o símbolo da variável, a descrição, a fonte e a frequência das variáveis utilizadas

Variável	Descrição	Fonte	Frequência
volume	Quantidade de ações negociadas durante um dia de negociação	Thomson Reuters	Diária
id	Índice de informatividade do histórico recente de preços, com valores entre -1 e +1, o índice é diretamente relacionado com a informatividade dos preços sobre a atenção do mercado	Equação 3	Diária*
idf	Variável <i>id</i> considerando a data referente ao próximo trimestre fiscal	Equação 3	Trimestral
fleshmd	Valor médio do índice de Flesch (1948) obtido a partir das notícias publicadas no dia t	Notícias	Diária
fleshpd	Valor médio do índice de Flesch (1948) obtido a partir das notícias publicadas no dia t_{-1}	Notícias	Diária
flesh1m	Valor acumulado do índice de Flesch (1948) das notícias publicadas entre os dias t e t_{-19}	Notícias	Mensal
flesh6m	Valor acumulado do índice de Flesch (1948) das notícias publicadas entre os dias t e t_{-150}	Notícias	Semestral
flesh12m	Valor acumulado do índice de Flesch (1948) das notícias publicadas entre os dias t e t_{-251}	Notícias	anual
Mflesh12m	Valor médio do índice de Flesch (1948) das notícias publicadas entre os dias t e t_{-251}	Notícias	anual
sentmd	Valor médio do índice de sentimento textual das notícias publicadas no dia t	Notícias	Diária
sentpd	Valor médio do índice de sentimento textual das notícias publicadas no dia t_{-1}	Notícias	Diária
sent1m	Valor acumulado do índice de sentimento textual das notícias publicadas entre os dias t e t_{-19}	Notícias	Mensal
sent6m	Valor acumulado do índice de sentimento textual das notícias publicadas entre os dias t e t_{-150}	Notícias	Semestral
sent12m	Valor acumulado do índice de sentimento textual das notícias publicadas entre os dias t e t_{-251}	Notícias	Anual
Msent12m	Valor médio do índice de sentimento textual das notícias publicadas entre os dias t e t_{-251}	Notícias	Anual
nunot12m	Número de notícias publicadas no período de um ano sobre uma mesma empresa	Notícias	Anual
Δ nunot12m	Variação do número de notícias publicadas no período de um ano sobre uma mesma empresa	Notícias	Anual
Δ Turnover	Valor obtido pela diferença do <i>turnover</i> médio em t menos o <i>turnover</i> médio em $t - 1$	Thomson Reuters	Trimestral
$ SUE $	O módulo do resultado inesperado padronizado do lucro líquido trimestral	Thomson Reuters	Trimestral
Δ Analyst	Variação no número de analistas <i>sell-side</i> responsáveis pela cobertura da empresa entre dois trimestres consecutivos	Thomson Reuters	Trimestral

Fonte: Elaboração própria (2021)

profissionais; já a variável $|SUE|$ é o módulo do resultado inesperado padronizado do lucro líquido trimestral, $|SUE|$ sinaliza a assimetria de informação em relação ao resultado da empresa. $|SUE|$ é obtida pela equação 4:

$$|SUE| = \left| \frac{(L.L. \text{ reportado} - L.L. \text{ esperado})}{\text{Desvio pad. L.L.}} \right| \quad (4)$$

Em que:

- $|SUE|$ = módulo do resultado inesperado padronizado do lucro líquido
- *L.L. reportado* = valor contábil do lucro líquido reportado para o período;
- *L.L. esperado* = valor contábil do lucro líquido esperado para o período;
- *Desvio pad. L.L.* = O desvio padrão da série histórica do lucro líquido da empresa.

1.2.3 Modelos Econométricos

Utilizamos o Método dos Momentos Generalizados Sistêmico (GMM-Sys) para verificar as Hipóteses 1, 2 e 3. O GMM-Sys permite contornar a exogeneidade restrita e garantir estimadores não viesados com o uso de defasagens na variável dependente, o uso de *lags* da variável dependente como parte das variáveis explicativas permite evitar a má especificação da forma funcional. Os testes para validação das condições de cada modelo são apresentados conjuntamente aos resultados encontrados nos tópicos a seguir (ARELLANO; BOND, 1991).

Visando verificar as Hipóteses 1 e 2 propomos o modelo presente na equação 5. O modelo se propõe a estimar a relação entre o custo de aquisição da informação e o consumo do estoque estocástico de ofertas dos ativos de risco, esse consumo seria menor quanto mais custoso fosse processar a informação (FOUCAULT; HOMBERT; ROȘU, 2016; AVDIS, 2016).

$$\begin{aligned} volume_{it} = & \alpha + L1.volume_{it-1}\delta_1 + L2.volume_{it-2}\delta_2 \\ & + sent_{it}\beta_1 + flesch_{it}\beta_2 + nunot12m_{it}\beta_3 + ID_{it}\beta_4 + \mu_{it} \end{aligned} \quad (5)$$

O modelo foi estruturado da seguinte forma: o volume de transações no dia t do ativo i (variável $volume_{it}$) é explicado pelas variáveis obtidas a partir das notícias: o número de notícias que citam a empresa nos últimos 12 meses ($nunot12m_{it}$), o sentimento textual das notícias ($sent_{it}$) e o indicador de complexidade textual ($flesch_{it}$). O número das notícias e o sentimento textual são as variáveis de controle relacionadas com a atenção da mídia e a direção do preço no ingresso de novas informações (DA; GURUN; WARACHKA, 2014; AVDIS, 2016; FOUCAULT; HOMBERT; ROȘU, 2016).

Também foram usadas a variável id_{it} que indica a sensibilidade do preço para absorver informações novas e as variáveis de defasagem temporal $L1volume_{it-1}$ e $L2volume_{it-2}$. A sensibilidade do mercado remonta o trabalho de Da, Gurun e Warachka (2014), seu uso visa controlar a atenção dos investidores, já as variáveis de defasagens temporal do volume transacionado são relacionadas ao modelo GMM-Sys, devido a estrutura dos dados e ao teste de Arellano e Bond (1991) foram adotados dois *lags* temporais, especificações com mais defasagens não foram estimadas por limitação dos dados.

Para testar a relação proposta na Hipótese 3, foi estimado o modelo apresentado na equação 6. A relação preconiza que a informatividade dos preços estaria inversamente relacionada com o custo da informação pois, segundo a “Hipótese do Sapo na Panela”, informação discreta estaria relacionada com mais atenção da mídia (DA; GURUN; WARACHKA, 2014). Tal atenção é expressa pela quantidade de artigos publicados e tem relação com o sentimento dos artigos escritos, todavia não foi observada se a atenção da mídia se reflete em menor custo de acesso a informação. Nesse sentido, a variável do tipo $flesch_{it}$ tenta emular a dificuldade dos investidores em transformar atenção da mídia em preço. A variável id_{it} foi utilizada como *proxy* da discricionariedade da informação presente nos preços e as variáveis explicativas como controle para atenção da mídia, exceto pelas as variáveis de complexidade que foram adicionadas ao modelo de Da, Gurun e Warachka (2014) visando acrescentar o custo de aquisição da informação na equação.

$$id_{it} = \alpha + L1.id_{it-1}\delta_1 + L2.id_{it-2}\delta_2 + \Delta turnover\beta_1 + |SUE|_{it}\beta_2 + nunot_{it}\beta_3 + \Delta Analyst_{it}\beta_4 + sent_{it}\beta_5 + flesch_{it}\beta_6 + \mu_{it} \quad (6)$$

O modelo apresentado na equação 6 tem com variável dependente a informatividade dos preços (id_{it}), sendo explicada pela variação no giro das ações $\Delta turnover$, pelo módulo da surpresa padronizada do lucro líquido (*standardized unexpected earnings*) $|SUE|_{it}$, pelo número de notícias sobre a empresa no ano anterior ($nunot_{it}$), pela variação na cobertura de analista *sell-side* ($\Delta Analyst_{it}$), pelo sentimento das notícias sobre a empresa ($sent_{it}$) e pela complexidade textual dos artigos jornalísticos sobre a empresa ($flesch_{it}$).

1.3 DISCUSSÃO DOS RESULTADOS

1.3.1 Análise descritiva dos dados

Para realizar a pesquisa, foram coletadas 21086 notícias, organizadas de acordo com as empresas citadas e as datas de publicação. Após o tratamento dos dados, foram obtidos, em média, pouco menos do que 32 artigos por ano por empresa. Para testar as hipóteses, a amostra foi dividida em duas frequências temporais distintas. A Tabela 2 apresenta as estatísticas descritivas para os dados diários e os trimestrais.

Em relação a variação do número de notícias anuais ($\Delta nunot12m$), observou-se uma variação positiva de pouco mais de 2 artigos por período. Este aumento está em linha com o aumento da disponibilidade de informação financeira na internet. Por dia, o sentimento médio foi de 0,024 e 0,031 indicando que na média os textos são mais positivos que negativos, como mostram $sentmd$ e $sentpd$. O valor médio da variável de interesse $fleschmd$ denota baixa complexidade das notícias financeiras, considerar que notícias financeiras são textos simples é possível considerar que os artigos facilitariam o acesso de informação para os investidores e poderiam estar relacionados com o aumento da informatividade dos preços dos ativos (BAI; PHILIPPON; SAVOV, 2016).

A partir da estatística descritiva foi possível identificar que a variável $fleschmd$ assumiu valores mínimos extremos de -630.66, este valor representa um *outlier* fruto do procedimento automático de coleta e processamento, o algoritmo é sensível a partes de códigos perdidos no meio do texto para processamento textual do índice de Flesch (1948). Outros doze casos nos quais códigos foram coletados como texto foram identificados e retirados da amostra. O problema relatado tem menor prevalência para o resultado do processamento do sentimento textual,

Tabela 2 – A tabela apresenta os resultados para estatística descritiva das variáveis, são mostrados os resultados para número total de observações, média simples, desvio padrão e valores de mínimo e máximo.

Variáveis em frequência diária					
Variáveis	Observações	Média simples	Desvio padrão	Mínimo	Máximo
volume	175740	3618885	7265840	0	5.39E+08
nunot12m	139380	31.48	56.04	0	530
id	149568	-0.0295496	0.0548443	-0.312	0.264
sentmd	12909	4.04E-03	0.024	-0.239	0.263
fleschmd	12909	3.810431	19.61	-630.67	60.19
sentpd	12909	5.9E-03	0.031	-0.266	0.341
fleschpd	12909	5.426	22.186	-630.67	102.71
sent1m	99749	0.016	0.0620	-0.595	0.672
flesch1m	99749	14.961	44.241	-713.228	289.732
sent6m	143225	0.0621	0.1731	-0.61	1.4744
flesch6m	143225	61.644	134.894	-1315.935	824.363
sent12m	150705	0.101	0.278	-0.63	2.141
flesch12m	150705	109.349	232.371	-1477.734	1227.128
Variáveis em frequência trimestral					
id	2456	-0.03023	0.0548	-0.256	0.236
idf	2393	-0.03017	0.0549	-0.256	0.236
Δ Turnover	2571	6.74E+06	4.54E+07	-2.33E+08	9.43E+08
SUE	1913	7.591	97.356	0.00393	3286.4
Δ nunot12m	2943	2.121	7.538	-50	66
Δ Analist	3035	0.003	5.429	-34	46
Mflesch12m	2485	2.807	8.551	-1.2495	29.898
Msent12m	2485	0.0033	0.0096	-0.095	0.058
flesch12m	3035	85.876	202.097	-1473.843	1167.45
sent12m	3035	8.67E-02	0.2540	-2.627	1.931

Fonte: Resultados da Pesquisa (2021).

uma vez que os trechos com caracteres não textuais não classificados, apenas contabilizados, reduzindo a magnitude do resultado.

1.3.2 O Conteúdo Informacional das Notícias e a Velocidade de Propagação das Informações Financeiras

Nesta seção serão apresentados os resultados relacionados com as hipóteses 1 e 2, elas foram verificadas a partir do modelo econométrico descrito na Equação 5. Os resultados estão expostos na Tabela 3 com os testes de adequação na Tabela 4. Os testes aplicados são de autocorrelação serial de [Arellano e Bond \(1991\)](#), que verifica a existência da autocorrelação de primeira ordem e ausência em segunda ordem e o teste de [Sargan \(1958\)](#) para a validade dos instrumentos.

As variáveis de complexidade textual *fleshpd*, *flesh1m*, *flesh6m*, *flesh12m*, foram direta e significativamente relacionadas com o volume transacionado, o resultado segue em linha com a Hipótese 1, ou seja, eles indicam que notícias mais simples estão relacionadas com

Tabela 3 – Nesta tabela todos os modelos usam como variável dependente $volume_{i,t}$, a diferença consiste nas variáveis de sentimento e complexidade textual que são formadas tomando por base o dia que a notícia saiu (*md*), o próximo dia (*pd*), o acumulado mensal (*1m*), semestral (*6m*) ou anual (*12m*).

	(dia)	(+1 dia)	(1 mês)	(6 meses)	(1 ano)
<i>Lag 1 volume</i>	0.481*** (93.94)	0.415*** (68.65)	0.462*** (53.07)	0.404*** (104.84)	0.384*** (66.10)
<i>Lag 2 volume</i>	0.285*** (41.28)	0.301*** (82.58)	0.209*** (30.27)	0.250*** (71.31)	0.215*** (39.31)
nunot12m	9688.2*** (12.02)	12015.8*** (19.80)	15756.0*** (5.15)	-3133.3 (-0.20)	-4424.7 (-0.22)
id	-1.16E+07*** (-5.90)	-1.43E+07*** (-4.76)	-1.41E+07** (-2.19)	-1.80E+07** (-3.18)	-2.52E+07*** (-6.72)
sentmd	-9087228** (-2.55)				
fleschmd	-8804.9 (-1.16)				
sentpd		5979988 (1.00)			
fleschpd		14106.1 * (1.70)			
sent1m			-3632461 (-0.68)		
flesch1m			15589.6 * (1.86)		
sent6m				-1926534 (-0.69)	
flesch6m				17394.8*** (6.96)	
sent12m					-3450798 (-1.24)
flesch12m					13619.8*** (6.19)
constante	1151755*** (6.24)	428256.7** (2.55)	-37383.4 (-0.03)	37698.9 (0.03)	-194649.2 (-0.21)
Observações	6895	6701	38844	53385	55435

Os erros padrão estão entre parênteses, significância estatística: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Tabela 4 – Resultados dos testes de Autocorrelação serial de [Arellano e Bond \(1991\)](#) e de [Sargan \(1958\)](#), respectivamente, para autocorrelação serial em primeira e segunda ordem e para verificar validade dos instrumentos utilizados em decorrência da metodologia GMM-Sys utilizada na tabela 3.

Painéis	(dia)	(+1 dia)	(1 mês)	(6 meses)	(1 ano)
Teste Arellano e Bond (1991) para autocorrelação serial:					
1 lag	0.0238	0.1250	0.0044	0.0294	0.0204
2 lags
Teste de Sargan (1958) para sobreidentificação:					
chi2	54.23	55.86	52.40	50.86	54.74
Prob > chi2	1.00	1.00	1.00	1.00	1.00
nº de instrumentos	2.1e+03	2.0e+03	5.0e+03	5.2e+03	5.2e+03

Fonte: Resultados da Pesquisa (2021).

os dias de maior quantidade de ativos negociados. Partindo da premissa que o índice de [Flesch \(1948\)](#) é *proxy* para o custo no acesso a informação, o resultado está de acordo com a perspectiva de que informações com menor custo de aquisição tendem a gerar maior volume negociado pois o conteúdo informacional destas notícias seria absorvido mais rápido. Também encontra suporte teórico na “Teoria de Aquisição de Informação em Sistemas Dinâmicos” ([AVDIS, 2016](#)).

nunot12m indica que o número de notícias nos 12 meses anteriores é direta e significativamente relacionado com o volume transacionado no dia de publicação de novas notícias. Como a variável *nunot12m* serve de *proxy* para a atenção da mídia, está em linha com o trabalho de [Zhang et al. \(2016\)](#), que é relatada que uma maior atenção da mídia acelera a propagação da informação. Também é possível que a atenção da mídia provoque a atenção dos investidores, neste caso os resultados estão em linha com estudo de [Da, Gurun e Warachka \(2014\)](#), onde a atenção dos investidores pode acelerar o processo de propagação da informação.

A variável *id* foi significativa, mas inversamente relacionado com o volume transacionado. Indicando que quanto menos volátil é o histórico recente de preços, maior será o volume transacionado no dia em que notícias são publicadas. O resultado contraria o que seria esperado pela nossa interpretação da “Hipótese do Sapo na Panela”. Mesmo estando em linha com os resultados observados na tabela 5, o resultado não é conclusivo. Vale destacar que *id* é *proxy* indireta para atenção do mercado, mas é produzida a partir das oscilações de preço, é possível que ao controlar para o custo de aquisição de informação, para o sentimento e para atenção da mídia essas oscilações de preço estejam relacionadas com o prêmio pela precificação de informações passadas ([DA; GURUN; WARACHKA, 2014; ZHANG et al., 2016](#)).

A variável *sentmd* aponta que o efeito do sentimento é significativo para explicar o volume transacionado no dia de publicação de uma nova notícia, ou seja, nos dias que notícias negativas foram publicadas, o mercado negociou mais as empresas citadas que nos dias com notícias positivas. Tal assimetria encontra respaldo na teoria da perspectiva, na medida que agentes seriam mais aversos às perdas que aos ganhos. Outra justificativa possível é o fato do sentimento médio na amostra ser preponderantemente positivo. Logo, é possível que o resultado seja decorrente do período avaliado ([KAHNEMAN; TVERSKY, 1979; LAAKKONEN; LANNE, 2008](#)).

Verificar a hipótese 2 exigia que os resultados das variáveis de complexidade e sentimento textual fossem analisados de maneira conjunta. A hipótese preconiza que o efeito da

complexidade é mais persistente que o efeito do sentimento textual dada a ocorrência de eventos informacionais. O resultado encontrado nas variáveis *flesh12pd*, *flesh1m*, *flesh6m*, *flesh12m* e *sentmd* está consoante com essa hipótese, ou seja, o efeito do custo da informação parece persistir por mais tempo do que o efeito sentimento. Tradicionalmente, os relatos sobre o efeito das notícias no mercado financeiros focam na variação dos preços e retornos. Enquanto isso, a duração dos eventos informacionais não encontra consenso na literatura, ao considerar o custo de acesso a informação, o estudo indica que existem efeitos de curto e de médio prazos, mas esses são decorrentes de componentes informacionais diferentes, o efeito de curto prazo está no sentimento textual e o efeito de mais longo prazo pode estar relacionado com o esforço de processamento das informações (TETLOCK, 2007; LILLO *et al.*, 2015; FOUCAULT; HOMBERT; ROȘU, 2016; CERCHIELLO; NICOLA, 2018).

Um resultado inesperado, mas consistente com a hipótese 2, foi que o sentimento textual parece só estar relacionado com o volume transacionado no dia de publicação das notícias, diferente da complexidade que tem efeitos persistentes por até um ano. Tal resultado, se observado em conjunto com o trabalho de Manela e Moreira (2017) e Foucault, Hombert e Roșu (2016), evidencia a relevância da velocidade para estratégias baseadas em sentimento.

Também não era esperado que a variável *nunot12m* deixasse de ser significativa em modelos que consideram o efeito das notícias em prazos maiores, tal fato ocorreu nos modelos com sentimento e complexidade textual acumulada por seis meses e um ano. A variável *nunot12m* não sofre alteração nestes modelos, a única diferença decorre da interação com as outras variáveis. É possível que o resultado tenha relação com o comportamento dos investidores frente a uma sequência de notícias, nestes casos foi relatado que a atenção seria direcionada aos artigos com menor custo de leitura e relacionados com as informações já adquiridas anteriormente (FEDYK, 2018; LEE *et al.*, 2020).

As variáveis *Lag1volume* e *Lag2volume* foram significativas. O resultado demonstra que o volume de um e de dois dias passados pode explicar o volume no dia atual, apontando para a utilização do modelo GMM-Sys. Devido à estrutura e à quantidade de observações não foi possível verificar a dependência do volume para três ou mais dias.

1.3.3 O Conteúdo Informacional das Notícias como Estímulo para Atenção dos Investidores.

Este tópico foi dedicado a hipótese 3. O objetivo é verificar como a complexidade textual das notícias e o ruído produzido pelo custo de aquisição da informação estão relacionados com a informatividade decorrente de variações intensas nos preços dos ativos. A relação esperada é que quanto mais simples forem as notícias no período em que *id* é calculado mais informação discreta estará presente nos preços, ou seja, produzindo *id* mais próximo de +1 e mais atenção dos investidores (DA; GURUN; WARACHKA, 2014).

Os resultados dispostos na tabela 5 são produzidos tomando como base o modelo econométrico referenciado na equação 6, foram produzidos quatro painéis: dois deles (*id A* e *id M*) consideram que as notícias ocorridas em sincronia ao período de formação de *id* afetam a variável dependente, outros dois (*id_{i,t+1} A* e *id_{i,t+1} M*) desconsideram o trimestre mais recente que é substituído pelo último trimestre anterior ao período de formação de *id*. Dois tipos de efeitos foram considerados: o efeito médio do sentimento e da complexidade textual e o efeito acumulado do sentimento e da complexidade textual dos artigos jornalísticos.

Na tabela 6 estão resultados para os testes de adequação da metodologia estatística GMM-Sys, são eles: os testes de autocorrelação serial de Arellano e Bond (1991) e os testes de

Sargan (1958) de validade dos instrumentos.

Tabela 5 – Nesta tabela são apresentados os resultados referentes ao modelo econométrico descrito na equação 6. Quatro painéis foram elaborados como o objetivo de testar a hipótese 3. O método estatístico utilizado para obtenção dos painéis foi a metodologia GMM Sys.

	(id A)	(id M)	(id_{t+1} A)	(id_{t+1} M)
<i>Lag 1</i>	0.355*** (0.053)	0.334*** (0.046)	0.330*** (0.044)	0.321*** (0.040)
<i>Lag 2</i>	-0.055 (0.038)	-0.059* (0.034)	-0.055 (0.049)	-0.046 (0.047)
$\Delta Turnover$	-2.79e-11 (1.96e-11)	-2.89e-11 (2.15e-11)	1.84e-11 (3.78e-11)	-1.76e-11 (3.93e-11)
<i>SUE</i>	2.18e-5 (1.36e-5)	2.13e-5 (1.33e-5)	-4.82e-06 (8.84e-06)	-4.40e-06 (9.30e-06)
$\Delta nunot12m$	4.33e-04 (3.71e-04)	4.73e-04 (365e-04)	-4.06e-04 (3.22e-04)	-4.93e-04 (3.29e-04)
<i>flesch12m</i>	4.62e-06 (1.31e-05)		3.07e-06 (1.83e-05)	
<i>sent12m</i>	0.010 (0.018)		-0.025 (0.017)	
$\Delta Analyst$	1.09e-03 (9.97e-04)	1.26e-03 (1.16e-03)	-5.00e-03 (9.38e-03)	-3.34e-04 (9.81e-04)
<i>Mflesch12m</i>		6.28e-05 (1.94e-04)		-6.25e-04** (2.72e-04)
<i>Msent12m</i>		-0.040 (0.313)		-0.456** (0.190)
Constante	-0.019*** (0.004)	-0.020*** (0.003)	-0.017*** (0.004)	-0.017*** (0.003)
Observações	1643	1531	1594	1481

Erros padrão entre parênteses, significância estatística: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Fonte: Resultados da Pesquisa (2021).

A Hipótese 3 prevê que a ocorrência de notícias mais simples poderia estar relacionada com maior informatividade nos preços. O resultado para a variável *Mflesch12m* no painel ($id_{i,t}$) M foi significativo, mas apresentou comportamento oposto ao esperado, a ocorrência de notícias mais complexas no período foi relacionado com índices de informatividade maiores, ou seja, notícias mais complexas aguçariam a atenção dos investidores em relação as mais simples. É possível que o esforço para processar informações complexas ou o prêmio relacionado ao valor

Tabela 6 – Resultados dos testes de Autocorrelação serial de [Arellano e Bond \(1991\)](#) e de [Sargan \(1958\)](#), respectivamente, para autocorrelação serial em primeira e segunda ordem e para verificar validade dos instrumentos utilizados em decorrência da metodologia GMM-Sys utilizada na tabela 5.

painéis	(id M)	(id A)	(id_{t+1} M)	(id_{t+1} A)
Teste Arellano e Bond (1991) para autocorrelação serial:				
1 lag	0.0000	0.0000	0.0000	0.0000
2 lags	0.2304	0.2707	0.1457	0.1473
Teste de Sargan (1958) para sobreidentificação:				
chi2	54.23	55.86	52.40	50.8
Prob > chi2	1.00	1.00	1.00	1.00
n° de instrumentos	2.1e+03	5.0e+03	5.2e+03	5.2e+03

Fonte: Resultados da Pesquisa (2021).

da informação estimulem a atenção dos investidores, outra possibilidade é que informações complexas produzam mais oscilação e ruído nos preços ([DA; GURUN; WARACHKA, 2014](#)).

Também merece destaque o resultado de $Msent12m$ em relação a id_{t+1} . Foi encontrado que as notícias negativas estariam relacionadas com maiores níveis de informatividade discreta nos preços. Isso implica na transmissão de informação para os preços com maior intensidade e preços mais sensíveis a novas informações ([HENS; VLCEK, 2011](#)).

1.4 CONCLUSÕES

Para a amostra e os métodos aplicados neste estudo, notícias complexas foram responsáveis por elevar a atenção dos investidores contrariando a hipótese 3. Esse resultado é interessante pois permite questionar como o esforço necessário para processar e adquirir informação financeira afeta a percepção do investidor no médio e longo prazo. Estudos posteriores podem verificar o esforço na prospecção de informação na atenção do investidor.

Outra conclusão está relacionada com a hipótese 2. Considerando que a informatividade presente nas notícias assume duas formas, o sentimento e a complexidade, foi observado que cada uma delas apresentam prevalências temporais específicas: no curtíssimo prazo o sentimento seria responsável por nortear a volatilidade nas transações e, após o primeiro momento, o custo da informação teria papel de nortear o ingresso de informação. Faz sentido ponderar que o sentimento seria responsável pelo sentido do mercado e o custo da informação por determinar a velocidade de propagação da informação.

Ao analisar especificamente a relação entre a complexidade textual e a velocidade de transferência da informação para o mercado, foi possível confirmar a hipótese 1, para qual a complexidade textual está inversamente relacionada com maior velocidade de propagação de uma informação. Também é destaque o papel da complexidade textual como *proxy* para o custo de aquisição de informação em eventos noticiados, aparentemente esta variável se mostrou passível de ser utilizada em outros estudos.

REFERÊNCIAS

ARELLANO, M.; BOND, S. Some tests of specification for panel data: Monte carlo evidence

and an application to employment equations. *The review of economic studies*, Wiley-Blackwell, v. 58, n. 2, p. 277–297, 1991.

AVDIS, E. Information tradeoffs in dynamic financial markets. *Journal of Financial Economics*, Elsevier, v. 122, n. 3, p. 568–584, 2016.

BAI, J.; PHILIPPON, T.; SAVOV, A. Have financial markets become more informative? *Journal of Financial Economics*, Elsevier, v. 122, n. 3, p. 625–654, 2016.

BIRZ, G.; LOTT JR., J. R. The effect of macroeconomic news on stock returns: New evidence from newspaper coverage. *Journal of Banking & Finance*, Elsevier, v. 35, n. 11, p. 2791–2800, 2011.

BROMAN, M. S.; SHUM, P. Relative liquidity, fund flows and short-term demand: evidence from exchange-traded funds. *Financial Review*, Wiley Online Library, v. 53, n. 1, p. 87–115, 2018.

CAMPBELL, J. Y.; KYLE, A. S. Smart money, noise trading and stock price behaviour. *The Review of Economic Studies*, Wiley-Blackwell, v. 60, n. 1, p. 1–34, 1993.

CERCHIELLO, P.; NICOLA, G. Assessing news contagion in finance. *Econometrics*, Multidisciplinary Digital Publishing Institute, v. 6, n. 1, p. 5, 2018.

DA, Z.; GURUN, U. G.; WARACHKA, M. Frog in the pan: Continuous information and momentum. *Review of Financial Studies*, v. 27, n. 7, p. 2171–2218, 2014.

ENGELBERG, J.; MCLEAN, R. D.; PONTIFF, J. Anomalies and news. *The Journal of Finance*, Wiley Online Library, v. 73, n. 5, p. 1971–2001, 2018.

FAMA, E. F. Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, v. 25, n. 2, p. 383–417, 1970.

FEDYK, A. *Front page news: The effect of news positioning on financial markets*. [S.l.], 2018. Working paper. Disponível em: <https://scholar.harvard.edu/files/fedyk/files/afedyk_frontpagenews.pdf>. Acesso em: 2020-09-14.

FEUERRIEGEL, S.; PROELLOCHS, N. *SentimentAnalysis: Dictionary-Based Sentiment Analysis*. [S.l.], 2019. R package version 1.3-3. Disponível em: <<https://CRAN.R-project.org/package=SentimentAnalysis>>. Acesso em: 2018-05-03.

FLESCHE, R. A new readability yardstick. *Journal of applied psychology*, American Psychological Association, v. 32, n. 3, p. 221, 1948.

FOUCAULT, T.; HOMBERT, J.; ROŞU, I. News trading and speed. *The Journal of Finance*, Wiley Online Library, v. 71, n. 1, p. 335–382, 2016.

GREEN, J. R. Information, efficiency and equilibrium. Harvard Institute of Economic Research Discussion Paper No. 284. 1973.

GROSSMAN, S. J.; STIGLITZ, J. E. On the impossibility of informationally efficient markets. *The American economic review*, JSTOR, v. 70, n. 3, p. 393–408, 1980.

GUAY, W.; SAMUELS, D.; TAYLOR, D. Guiding through the fog: Financial statement complexity and voluntary disclosure. *Journal of Accounting and Economics*, Elsevier, v. 62, n. 2-3, p. 234–269, 2016.

- HENS, T.; VLCEK, M. Does prospect theory explain the disposition effect? *Journal of Behavioral Finance*, v. 12, n. 3, p. 141–157, 2011. Disponível em: <<http://www.tandfonline.com/doi/abs/10.1080/15427560.2011.601976>>. Acesso em: 2021-04-12.
- KAHNEMAN, D.; TVERSKY, A. Prospect theory: An analysis of decision under risk. *Econometrica*, v. 47, n. 2, p. 363–391, 1979.
- LAAKKONEN, H.; LANGE, M. Asymmetric news effects on volatility: good vs. bad news in good vs. bad times. 2008.
- LEE, Y.-H.; HU, P. J.-H.; ZHU, H.; CHEN, H.-W. Discovering event episodes from sequences of online news articles: A time-adjoint frequent itemset-based clustering method. *Information & Management*, Elsevier, v. 57, n. 7, p. 103348, 2020.
- LI, F. Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, Elsevier, v. 45, n. 2-3, p. 221–247, 2008.
- LILLO, F.; MICCICHÈ, S.; TUMMINELLO, M.; PILO, J.; MANTEGNA, R. N. How news affects the trading behaviour of different categories of investors in a financial market. *Quantitative Finance*, Taylor & Francis, v. 15, n. 2, p. 213–229, 2015.
- MANELA, A.; MOREIRA, A. News implied volatility and disaster concerns. *Journal of Financial Economics*, Elsevier, v. 123, n. 1, p. 137–162, 2017.
- RANCO, G.; ALEKSOVSKI, D.; CALDARELLI, G.; GRČAR, M.; MOZETIČ, I. The effects of Twitter sentiment on stock price returns. *PloS one*, v. 10, n. 9, p. e0138441, 2015.
- SARGAN, J. D. The estimation of economic relationships using instrumental variables. *Econometrica: Journal of the Econometric Society*, JSTOR, v. 26, n. 3, p. 393–415, 1958.
- SILVA, M. D. O. P. *O efeito do sentimento das notícias sobre o comportamento dos preços no mercado acionário brasileiro*. Tese (Doutorado) — Universidade de Brasília, 2017. Disponível em: <<https://repositorio.unb.br/handle/10482/32178>>. Acesso em: 2019-12-26.
- TETLOCK, P. C. Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance*, Wiley Online Library, v. 62, n. 3, p. 1139–1168, 2007.
- WANG, J. A model of intertemporal asset prices under asymmetric information. *The Review of Economic Studies*, Wiley-Blackwell, v. 60, n. 2, p. 249–282, 1993.
- ZHANG, Y.; SONG, W.; SHEN, D.; ZHANG, W. Market reaction to internet news: Information diffusion and price pressure. *Economic Modelling*, Elsevier, v. 56, p. 43–49, 2016.