O PODER DAS PALAVRAS: USANDO TEXTO PARA PREVISÕES ECONÔMICAS[1]

THE POWER OF WORDS: USING TEXT FOR ECONOMIC FORECASTS

Nathan Ramos de Almeida[2]
Hudson da Silva Torrent[3]

RESUMO

Este artigo estuda as novas ferramentas de Processamento de Linguagem Natural (NLP) e Análise de Sentimentos aplicadas à economia, com foco na utilização de textos para previsões macroeconômicas. O estudo constrói dois índices textuais: o Índice de Sentimento Econômico Textual (TESI) e o Índice de Incerteza de Política Econômica Textual (TEPU), aplicados a cinco categorias econômicas. Os resultados mostram que as variáveis textuais possuem um poder preditivo relevante para a taxa de crescimento do PIB no Brasil, especialmente em horizontes de curto prazo, superando benchmarks tradicionais e se revelando valiosas, principalmente na ausência de variáveis econômicas convencionais.[4]

ABSTRACT

This paper studies the new tools of Natural Language Processing (NLP) and Sentiment Analysis applied to economics, focusing on the use of texts for macroeconomic forecasting. The study constructs two textual indices: the Economic Sentiment Index (TESI) and the Economic Policy Uncertainty Index (TEPU), applied to five economic categories. The results show that textual variables have significant predictive power for Brazil's GDP growth rate, particularly in short-term horizons, outperforming traditional benchmarks and proving valuable, especially in the absence of conventional economic variables.

Keywords: Web Scraping, Text Mining, Sentiment, Economic Policy Uncertainty, Big Data, Machine Learning, Forecasting and Economics
J.E.L. Classifications: C52. C53. C80. E60.

[1]Trabalho de Conclusão de Curso apresentado, em 2024/1, ao Departamento de Ciências Econômicas e Relações Internacionais da Faculdade de Ciências Econômicas da Universidade Federal do Rio Grande do Sul (UFRGS), como requisito parcial para obtenção do título de Bacharel em Ciências Econômicas.

[2]Graduação em Economia pela Universidade Federal do Rio Grande do Sul. (nathanramosx5@gmail.com).
[3]Orientador. Doutor em Economia Aplicada pela Universidade Federal do Rio Grande do Sul. Professor do Departamento de Estatística da UFRGS. (hudson.torrent@ufrgs.br).

[4]O código e os dados desse artigo estão disponíveis em github.com/E30895/economics-forecast-using-text

1 INTRODUÇÃO

Com o avanço do mundo digital, cada vez mais informações são geradas na internet, e esses avanços provocaram mudanças em diversos paradigmas econômicos. Nos últimos anos, a macroeconomia tem recebido uma nova literatura relacionada ao processamento de dados de texto. Apesar de não serem a estrutura de dados mais convencional, a literatura recente reforça os benefícios de seu uso nas análises econômicas e financeiras (BAKER; BLOOM; DAVIS, 2016), (SHAPIRO; SUDHOF; WILSON, 2022) e (APRIGLIANO et al., 2023).

As pesquisas são inerentemente dispendiosas e frequentemente baseadas em amostras relativamente pequenas de indivíduos, o que pode levar a problemas de amostragem. Além disso, tendem a ser publicadas com frequência mensal e um atraso de duas ou mais semanas, reduzindo sua utilidade em momentos de viradas econômicas. Esse é o argumento de (SHAPIRO; SUDHOF; WILSON, 2022) para justificar seu estudo recente, que apresenta uma abordagem alternativa para medir sentimentos, focando no sentimento econômico contido nas notícias. Essa alternativa se contrapõe às medidas tradicionais baseadas em pesquisas, pois o índice proposto fundamenta-se no processamento do sentimento contido no texto dos artigos publicados em grandes jornais entre janeiro de 1980 e abril de 2015.

Em (SHAPIRO; SUDHOF; WILSON, 2022), consideram-se duas aplicações para os índices de sentimento baseados nos textos das notícias. A primeira considera até que ponto o sentimento diário das notícias pode prever os índices de sentimento do consumidor, concluindo que o sentimento das notícias é altamente preditivo, principalmente nos dias que antecedem os lançamentos do Índice de Sentimento do Consumidor de Michigan e do Índice de Confiança do Consumidor da Conference Board. Na segunda aplicação, a partir de modelos VAR(p), estimaram-se as respostas por impulso dos principais resultados macroeconômicos aos choques de sentimentos, em frequência mensal, onde constatou-se que choques de sentimentos positivos aumentam o consumo, a produção e as taxas de juros, ao mesmo tempo em que reduzem temporariamente a inflação.

Em um segundo momento, (APRIGLIANO et al., 2023) apresenta novas evidências sobre como as empresas e famílias italianas dedicam atenção às notícias dos jornais, com 60% da amostra citando esses veículos como uma das fontes mais importantes de informação para suas decisões econômicas. Apesar do recente aumento da mídia alternativa, os agentes continuam a prestar atenção às notícias econômicas relatadas nos jornais, sugerindo que as informações neles contidas são oportunas e significativas para suas escolhas econômicas.

(APRIGLIANO et al., 2023) desenvolve sua pesquisa explorando como dados de origem textual podem ajudar na previsão da atividade econômica na Itália. Para isso, são propostos dois índices de texto: o Índice de Sentimento Econômico Textual (TESI), que representa a série de sentimentos sobre um determinado assunto, calculado conforme (ARDIA et al., 2021), e o Índice de Incerteza de Política Econômica Textual (TEPU), uma adaptação do EPU proposto por (BAKER; BLOOM; DAVIS, 2016). Para a construção do banco de dados, foram escolhidos 16 tópicos econômicos e setoriais, nos quais, para cada tópico, foi criada uma série TESI e TEPU, resultando em 36 séries baseadas em texto.

Como resultados, (APRIGLIANO et al., 2023) demonstram que os artigos de jornais são uma fonte relevante de informações para empresas e famílias em suas decisões econômicas e financeiras. Os indicadores apresentados mostram-se aptos para previsões econômicas de curto prazo e, quando combinados com variáveis econômicas, apresentam uma melhora na precisão das previsões durante períodos de recessão; porém, esse ganho de performance não é observado em toda a amostra.

Neste trabalho, serão apresentadas demonstrações sobre o uso de dados de texto em análises econômicas. Primeiramente, será construído um banco de dados para armazenar os textos das notícias disponibilizadas no G1. Em seguida, serão desenvolvidos os Índices de Sentimento Econômico Textual (TESI) e de Incerteza de Política Econômica Textual (TEPU) para cinco categorias e setores da economia, representando dez variáveis de origem textual. Por fim, será testado o poder preditivo dessas variáveis textuais, assim como o ganho de performance preditiva obtido ao combinar variáveis econômicas e textuais.

O trabalho está estruturado da seguinte forma: a seção 2 apresenta a revisão da literatura, abordando os tópicos de Processamento de Linguagem Natural (NLP) e Análise de Sentimentos; a seção 3 detalha a construção do banco de dados, da coleta ao tratamento dos dados, além da elaboração dos Índices de Sentimento Econômico (TESI) e de Incerteza de Política Econômica (TEPU), que serão empregados na geração das variáveis textuais; a seção 4 apresenta a estratégia adotada para a previsão, assim como os modelos empregados no processo; na seção 5, serão expostos os resultados da pesquisa; por fim, na seção 6, serão discutidos os resultados obtidos e apresentadas sugestões para futuras pesquisas.

2 REFERENCIAL TEÓRICO

Textos são extremamente abundantes, podendo ser encontrados em meios diversos. O uso de métodos computacionais para o processamento de texto não é uma novidade e vem sendo observado desde os anos 50. Historicamente, o desafio reside na aquisição de poder computacional, uma vez que o processo exige bastante capacidade de processamento, além de adaptar essa estrutura de dados particular, descrita como "menos estruturada" (GENTZKOW; KELLY; TADDY, 2019). Dito isso, veremos como a literatura tem abordado os temas de processamento de linguagem natural, análise de sentimentos e previsões baseadas em texto.

2.1 PROCESSAMENTO DE LINGUAGEM NATURAL (NLP)

Nascido da intersecção entre métodos linguísticos e computacionais, o Processamento de Linguagem Natural (NLP) é uma vertente do Machine Learning que tem ganhado espaço à medida que os recursos computacionais se sofisticam e o volume de textos subjetivos publicados na internet cresce de forma exponencial. Esse campo de estudo provê métodos para o processamento de dados tipicamente categóricos e não estruturados, como palavras, frases, parágrafos e textos em geral. Além disso, a chamada linguística computacional fornece recursos amplamente utilizados no dia a dia, como tradutores, corretores ortográficos e reconhecimento de fala (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011).

2.2 ANÁLISE DE SENTIMENTOS

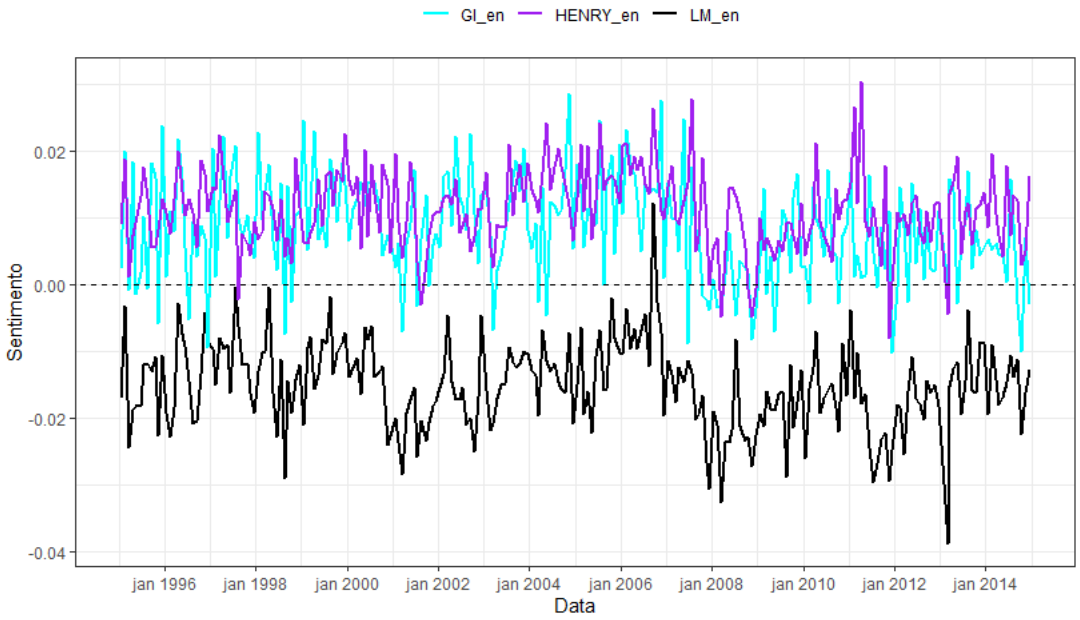
A análise de sentimentos, descrita por (NANLI et al., 2012) como um estudo computacional das opiniões, atitudes ou comportamentos de indivíduos e/ou entidades, é normalmente enquadrada como um problema de classificação binária (positiva ou negativa) ou ternária (positivo, negativo, neutro). Esse processo tem como finalidade encontrar opiniões e/ou identificar sentimentos expressos em elementos textuais, denominados Corpus, e, então, mensurá-los. A análise de sentimento é, portanto, uma medida do tom, da atitude ou da avaliação sobre um determinado tópico, independentemente da orientação emocional do próprio tópico (SHAPIRO; SUDHOF; WILSON, 2022).

Os métodos para análise de sentimentos, da forma que conhecemos, envolvem uma estrutura de dados chamada de dicionário. Introduzidos por (KIM; HOVY, 2004) e (HU; LIU, 2004), os dicionários são conjuntos de palavras escritas em determinado idioma, nos quais se inserem os sentimentos expressos em cada palavra. Eles são comumente apresentados em forma de tabelas, onde a primeira coluna contém a palavra e a segunda coluna o sentimento expresso, sendo muitas vezes representado por 1 para sentimentos positivos e -1 para sentimentos negativos.

O dicionário General Inquirer (GI), proposto inicialmente por (STONE; DUNPHY; SMITH, 1966) e mantido por Harvard, tem como objetivo ser um léxico geral para a língua inglesa, composto por palavras classificadas como positivas ou negativas. Porém, (LOUGHRAN; MCDONALD, 2011) argumentam que o dicionário General Inquirer (GI) considera muitas palavras como negativas, embora sejam neutras no contexto econômico e financeiro (como custos, despesas e risco). Nessas circunstâncias, é apresentado o Loughran-McDonald Master Dictionary (LM), um dicionário exclusivo para economia e finanças.[5]

Os dicionários foram comparados entre si utilizando o banco de dados "usnews", disponibilizado por (ARDIA et al., 2021), e apresentaram resultados distintos. Essa diferença ilustra a importância crucial do contexto em que o dicionário é aplicado, ressaltando que a eficácia da análise depende profundamente do contexto e da qualidade do dicionário empregado.

Figure 1 – Comparação entre dicionários para um mesmo conjunto de dados



Fonte: Elaboração Própria com dados de (ARDIA et al., 2021). O gráfico apresenta o efeito de diferentes dicionários para um mesmo conjunto de dados, destacando o quanto a análise é impactada pelo dicionário utilizado. Em azul, temos o dicionário General Inquirer; em roxo, o dicionário proposto por (HENRY, 2008), que foi construído a partir de análises de comunicações corporativas; e em preto, o dicionário Loughran McDonald.

[5] Uma característica do Loughran-McDonald Master Dictionary é sua construção e atualização contínua, baseada nas palavras predominantes dos relatórios 10-K publicados por companhias de capital aberto (LOUGHRAN; MCDONALD, 2011).

3 BASE E TRATAMENTO DOS DADOS

Essa seção apresenta e detalha o processo de construção dos bancos de dados textuais e econômicos que servirão de insumo para as análises e os modelos de Machine Learning. As bases terão caráter Open Source, disponíveis no repositório do Github.

3.1 A BASE DE DADOS TEXTUAL

O banco de dados textual é composto apenas por dados originados de textos. Na construção, foram coletadas mais de 200 mil notícias na plataforma do G1 entre os períodos de Mai/10 a Dez/23. As notícias estão distribuídas em 5 tópicos: (a) Agronegócio, (b) Indústria, (c) Mercado Financeiro, (d) Mercado de Trabalho e (e) Serviços. A escolha desses tópicos segue a metodologia de (APRIGLIANO et al., 2023), adaptada às Contas Nacionais Trimestrais, divulgadas pelo IBGE. A coleta dos dados envolveu o processo de Web Scraping, navegando pela maior quantidade de notícias possível para cada tópico entre as duas datas.

3.1.1 ÍNDICE DE SENTIMENTO ECONÔMICO (TESI)

O Índice de Sentimento Econômico Textual (TESI) é um método para transformar texto em índice. Ele é a implementação da Análise de Sentimentos, representando o sentimento expresso em um determinado texto. O TESI será aplicado aos 5 tópicos mencionados, resultando em 5 séries temporais que refletem o sentimento relacionado ao tópico.

O cálculo do TESI apoiou-se no *Framework Sentometrics* (ARDIA et al., 2021), uma abordagem que integra métodos quantitativos e qualitativos para análise de dados textuais, facilitando o processamento dos dados de texto. O processo ocorre em duas etapas, na qual a primeira etapa consiste em computar o sentimento individual de cada notícia, representado por:

$$SS_j = \frac{\sum_i^{Nwords} Polarity_i \times Shifter(i)}{Nwords} \tag{1}$$

Onde $Nwords$ é o número total de palavras do texto. $Polarity$ é uma classe que pode assumir um de dois valores: 1 para palavras positivas e -1 para palavras negativas, de tal modo que ambos se cancelem no somatório. O $Shifter$ representa o fator de deslocamento associado à palavra, usado para ajustar a polaridade da palavra atual de acordo com a palavra antecessora. De forma similar à *polaridade*, ele pode assumir -1 ou 1, dependendo da polaridade referente à palavra anterior.[6]

A segunda etapa é a agregação, na qual a pontuação final de sentimento em um determinado dia$_t$ é derivada pela combinação das séries individuais de cada jornal (SS_j). Cada série para um jornal específico é calculada a partir da média das pontuações dos artigos individuais. O TESI $_t$ é então obtido como uma média ponderada, em que os pesos são determinados pelo número de artigos publicados em um mês específico (APRIGLIANO et al., 2023).

[6] O Shifter foi concebido para remediar uma lacuna no método "tradicional" de análise de sentimentos. Nesse método, a palavra imediatamente anterior, frequentemente influenciando o contexto da palavra atual, não é considerada, resultando em índices sub-ótimos. Ao abordar essa limitação, o Shifter aprimora a capacidade de capturar nuances contextuais, promovendo uma análise mais precisa e abrangente dos sentimentos expressos no texto.

3.1.2 ÍNDICE DE INCERTEZA DE POLÍTICA ECONÔMICA (TEPU)

O Índice de Incerteza de Política Econômica Textual (TEPU) também é um método para transformar textos em índices. Ele mede a incerteza política e econômica através de palavras-chave e do processamento dos textos. Semelhante ao TESI, o TEPU será aplicado aos 5 tópicos mencionados, resultando em 5 séries temporais que refletem o sentimento de incerteza política e econômica. Pela natureza do índice, quanto maior for a incerteza piores são as expectativas.

O TEPU (APRIGLIANO et al., 2023) é uma adaptação do EPU (BAKER; BLOOM; DAVIS, 2016). O EPU é definido como qualquer texto que inclua pelo menos um termo associado à incerteza, política e economia, refletindo o nível de incerteza econômica e política no contexto específico (APRIGLIANO et al., 2023). A formulação do índice é representada pela seguinte equação:

$$TEPU_t = \frac{\sum_i^{N_t} 1[EPUArticle]}{N_t} \tag{2}$$

Nesta equação, N_t representa o número total de artigos coletados no período t, enquanto $1[.]_i$ é uma variável binária que identifica se um artigo específico aborda incerteza política e econômica (EPU). Assim, o TEPU reflete a proporção de textos classificados como EPU em relação ao total de artigos analisados no período, pois atendem simultaneamente aos critérios de incerteza, política e economia.

3.2 A BASE DE DADOS ECONÔMICA

O banco de dados econômico é destinado às variáveis de natureza econômica, como taxas de juros e inflação. Foi adotada a mesma base de dados do Bloomberg's Key Economic Indicators, apresentada por (BANTIS; CLEMENTS; URQUHART, 2023) no exercício de previsão da atividade econômica para o Brasil. Foram selecionadas 88 variáveis, distribuídas entre Atividade Econômica, Setor Externo, Governo, Mercado Imobiliário, Mercado de Trabalho, Monetária e Preços. A variável resposta do exercício de previsão será *BZEAMOM%*, uma proxy para a variação da Atividade Econômica Mensal.

3.3 A ESCOLHA DOS DADOS

A escolha do G1 como fonte de dados textuais se baseia em três considerações: em primeiro lugar, sua longa trajetória como plataforma de notícias no Brasil, proporcionando uma ampla amostragem. Em segundo lugar, sua acessibilidade para técnicas de Web Scraping, possibilitando a extração automatizada de dados. Em terceiro lugar, destaca-se a natureza pública da plataforma, garantindo que os índices desenvolvidos possam ser replicados e verificados por qualquer interessado.[7] Os dados econômicos foram escolhidos de acordo com (BANTIS; CLEMENTS; URQUHART, 2023).

[7](APRIGLIANO et al., 2023) extrai as notícias de 4 grandes jornais. Para o escopo desse trabalho, devido a limitação de tempo e orçamento não há possibilidade de replicar tal fidelidade ao trabalho, tendo então simplificar para tornar factível.

3.4 TRATAMENTO DOS DADOS

O banco de dados textual demanda um tratamento especial, que foi dividido em 7 etapas: (1) Remoção dos textos sem relação com o tópico[8], (2) Remoção dos textos duplicados, (3) Remoção das expressões regulares, (4) Tradução dos textos para a língua inglesa[9], (5) Remoção das palavras sem valor semântico, (6) Remoção das pontuações textuais e (7) Remoção de números. A nível de replicabilidade, todo o processo de coleta e tratamento foi implementado na linguagem de programação Python.

3.5 ESTACIONARIEDADE E DEFASAGENS

Após o tratamento especial do banco de dados textual, foram aplicados os índices TESI e TEPU aos textos para transformá-los em índices. Para os bancos de dados (tanto econômico quanto textual), foram realizados testes de Raiz Unitária a fim de garantir a estacionariedade das séries. Os dados foram agrupados em uma única base, e foram adicionadas 4 defasagens para cada variável.[10]

4 MÉTODOS DE PREVISÃO

Esta seção detalha os modelos adotados para a condução do exercício de previsão. O objetivo central é investigar a utilização de notícias como ferramenta de previsão econômica no contexto brasileiro. A estrutura metodológica é fundamentada em pesquisas recentes, mais precisamente nas contribuições de (APRIGLIANO et al., 2023), somadas aos trabalhos apresentados por (BANTIS; CLEMENTS; URQUHART, 2023).

4.1 MODELAGEM

Este estudo foca na aplicação prática das técnicas de previsão, evitando debates teóricos aprofundados. A escolha dos modelos LASSO, ENET e Boosting foi motivada por sua capacidade de realizar uma seleção automática das variáveis mais relevantes, sem a necessidade de aderir a uma teoria econômica específica. Esse aspecto é particularmente crucial devido ao grande volume de variáveis envolvidas: 88 variáveis econômicas e suas respectivas defasagens, além de 10 variáveis textuais e suas defasagens. A complexidade gerada por esse vasto conjunto de dados exige modelos que possam lidar eficientemente com a alta dimensionalidade, identificando as variáveis mais impactantes para a previsão. Para avaliação comparativa, o modelo SARIMA foi utilizado como benchmark, devido à sua reconhecida eficácia em análises de séries temporais. A nível de replicabilidade, todo o processo de modelagem e previsão foi implementado na linguagem estatística R.

[8] Visando identificar e remover os textos sem relação com os tópicos, foi criada uma lista de palavras-chave para cada tópico. Para que o texto avançasse na análise, ele precisava conter pelo menos uma palavra-chave. Os que não atenderam a esse critério foram descartados. A lista de palavras-chave encontra-se no repositório.

[9] Na data em que este trabalho foi escrito, ainda não havia um dicionário para economia e finanças em português. (APRIGLIANO et al., 2023) utilizou os textos coletados para propor um dicionário para o italiano, mas essa ação foge ao escopo do trabalho. Como alternativa, utilizou-se a API do Google através da biblioteca TextBlob (0.17.1) em Python para aplicar as traduções dos textos para a língua inglesa e usar o dicionário LM.

[10] Considerando que a amostra conta com apenas 164 observações, a inclusão de um grande número de defasagens poderia comprometer significativamente a qualidade das previsões, devido à perda de informações que isso acarreta. No contexto específico deste estudo, os modelos selecionam, em média, apenas as quatro primeiras defasagens.

4.1.1 SARIMA

O modelo Autorregressivo Integrado de Médias Móveis Sazonal (SARIMA) é uma extensão do modelo Autorregressivo Integrado de Médias Móveis (ARIMA). Nesse modelo, são preservadas as características do ARIMA e incluídos termos sazonais, permitindo a captura de padrões sazonais nas séries temporais, sendo considerado um avanço em relação ao seu antecessor (BOX et al., 2015)[11]. O modelo é dado por:

$$\text{ARIMA} \quad \underbrace{(p,d,q)}_{\text{Componente não sazonal}} \quad \times \quad \underbrace{(P,D,Q)}_{\text{Componente Sazonal}} \tag{3}$$

onde Componente não Sazonal

$$\phi_p(B)(1-B^d)Z_t = \theta_q(B)a_t \tag{4}$$

e Componente Sazonal:

$$\Phi_p(B)(1-B^d)Z_t = \Theta_q(B)a_t \tag{5}$$

combinando 4 e 5 obtém-se o ARIMA Sazonal:

$$\Phi_p(B^s)\phi_p(B)(1-B)^d(1-B^s)^D\dot{Z}_t = \theta_q(B)\Theta_Q(B^s)a_p, \tag{6}$$

onde, pelo lado não sazonal: $\phi(p)$ representa o componente autorregressivo (AR), obtido pela função de autocorrelação parcial (PACF); d representa a ordem de integração, obtida pelo teste de raiz unitária (Dickey-Fuller Aumentado, para nosso caso); $\theta(q)$ representa o componente da média móvel (MA), obtido pela função de autocorrelação (ACF). A parte sazonal do modelo apresenta termos semelhantes aos componentes não sazonais; porém, ela envolve os retrocessos apresentados no período sazonal, onde os componentes $\Phi(P)$ e $\Theta(Q)$ serão vistos nos atrasos sazonais das funções PACF e ACF (BOX et al., 2015).

4.1.2 LASSO

Conforme (BANTIS; CLEMENTS; URQUHART, 2023), o operador de seleção e redução absoluta mínima (LASSO) foi introduzido inicialmente por (TIBSHIRANI, 1996). Esse método de regressão aplica simultaneamente a regularização de variáveis, potencialmente melhorando a acurácia e a interpretação dos resultados. O LASSO minimiza a soma dos quadrados dos resíduos, impondo uma restrição na soma dos valores absolutos dos coeficientes, para que seja menor que uma constante:

$$\hat{\beta}_{\text{lasso}} = \arg\min_{\beta} \sum_{t=1}^{T}(y_t - \beta_0 - \sum_{i=1}^{n} x_{ti}\beta_i)^2 \quad \text{sujeito a} \quad \sum_{i=1}^{n}|\beta_i| \leq s \tag{7}$$

onde y_t é a t-ésima observação da variável alvo, β_0 é um intercepto, x_{ti} é a t-ésima observação do i-ésimo preditor, β_i é o coeficiente correspondente, $\sum_{i=1}^{n}|\beta_i|$ denota a penalidade L1, e s representa o parâmetro de ajuste. Uma forma equivalente de escrever o estimador LASSO na forma lagrangiana é:

[11] Como estratégia metodológica para implementação em larga escala, adotou-se a função auto.arima() da biblioteca Forecast no R. Os parâmetros foram ajustados para que o algoritmo encontre o melhor modelo entre as classes não sazonal e sazonal, usando o método de Ljung-Box e o critério de informação BIC.

$$\hat{\beta}_{\text{lasso}} = \arg\min_{\beta} \left(\frac{1}{2} \sum_{t=1}^{T} (y_t - \beta_0 - \sum_{i=1}^{n} x_{ti}\beta_i)^2 + \lambda \sum_{i=1}^{n} |\beta_i| \right) \tag{8}$$

O multiplicador de Lagrange λ, conhecido como parâmetro de regularização do LASSO, determina a quantidade de redução: quando $\lambda = 0$, o estimador corresponde ao OLS, enquanto $\lambda \to \infty$ resulta na eliminação de todos os coeficientes. Essa modificação indica que alguns coeficientes são ajustados exatamente para zero, o que é uma característica vital, especialmente quando o conjunto de preditores possui uma estrutura de big data. Além disso, deve-se ressaltar que todas as variáveis foram padronizadas para que a estimativa não dependesse das unidades de medida. Existem duas abordagens para calcular o valor de λ: um esquema de validação cruzada em série temporal e o uso de critérios de informação, como AIC ou BIC. Para este trabalho, optou-se pela validação cruzada.

4.1.3 ELASTIC NET

O LASSO apresentado anteriormente pode ser de grande utilidade quando existem muitos coeficientes zero ou próximos de zero no modelo verdadeiro. Porém, ele apresenta limitações. (ZOU; HASTIE, 2005) identificam dois problemas em relação ao estimador LASSO: primeiro, quando a dimensão transversal excede o número de observações, $(p > n)$, o LASSO pode escolher apenas até n variáveis, o que representa uma desvantagem significativa para um método de seleção de variáveis. Em segundo lugar, quando há preditores com coeficientes parciais elevados, o LASSO escolhe apenas um entre esses preditores e não se importa com qual deles é escolhido. Dito isso, (ZOU; HASTIE, 2005) introduziram a penalidade do elastic net (ENET):

$$\hat{\beta}_{\text{enet}} = \arg\min_{\beta} \left(\sum_{t=1}^{T} (y_t - \beta_0 - \sum_{i=1}^{n} x_{ti}\beta_i)^2 + \lambda_1 \sum_{i=1}^{n} |\beta_i| + \lambda_2 \sum_{i=1}^{n} \beta_i^2 \right) \tag{9}$$

De forma semelhante ao estimador LASSO, o ENET realiza simultaneamente a regularização e a seleção de variáveis. No entanto, ele possui a vantagem adicional de poder selecionar grupos de preditores correlacionados.

4.1.4 BOOSTING

Para essa seção, foi adotado o mesmo procedimento empregado em (LINDEN-MEYER; SKORIN; TORRENT, 2021)[12]. O Boosting é um algoritmo de Machine Learning que tem como objetivo produzir um modelo, seja ele linear ou não linear, de forma iterativa e adaptativa às co-variáveis. Seja y_t uma série temporal genérica, x_t o vetor de regressores, sendo $x_t(p)$ as defasagens de x_t até p, onde $p = 1, 2, \ldots, 4$, e M o critério de finalização das iterações. Então, o modelo é o resultado da soma de todas as M componentes distintas, sendo expresso por:

$$\hat{f}\left(x_t\right) = \hat{f}^{(0)} + \nu \sum_{m=1}^{M} \hat{g}^{(m)} \tag{10}$$

onde v é um parâmetro de aprendizado $0 < v \leq 1$ (FRIEDMAN, 2001) e $\hat{g}^{(m)}(x_t; \hat{\beta}_m)$ é o aprendiz, com $\hat{\beta}_m$ sendo o conjunto de coeficientes obtidos a partir de um procedimento de ajuste e uma função de perda, para um dado m:

[12] O artigo apresenta o uso do Boosting para dezenas de regressores e suas defasagens. Esse procedimento vai de encontro com a proposta do trabalho, habilitando seu uso.

$$\hat{g}^{(m)}(x_t; \hat{\beta}_m) = \arg\min_{\hat{h}(\cdot)} L(y_t, \hat{h}(x_t)), \tag{11}$$

$L(\cdot)$ é a função de perda escolhida e $\hat{h}$ é o procedimento de ajuste. O critério de parada M pode ser escolhido de forma arbitrária ou determinado por procedimentos estatísticos, como AIC, BIC ou Validação Cruzada. O procedimento de Validação Cruzada foi escolhido para encontrar o parâmetro M adequado e adotou-se o Boosting Linear. O algorítimo implementado é descrito pelas etapas a seguir:

Passo 1. Inicia-se com $m = 0$, onde é definido $\hat{f}^{(0)} = \bar{y}_t$, sendo $\bar{y}_t$ a média.
Passo 2. Para $m = 1$ até M:

1. Calcula-se os resíduos, definidos como $\varepsilon_t = y_t - \hat{f}^{(m-1)}$.
2. É feita a regressão dos resíduos ε_t em cada preditor $x(p)$, com $p = 1, 2, \ldots, 4$, e calcula-se a soma dos quadrados dos resíduos (SSR).
3. Seleciona-se o preditor $x(p^*)$ que tem a menor SSR.
4. Define-se $\hat{g}^{(m)} = \hat{\beta}^{(p^*)} x(p^*)$.
5. Por fim, a estimativa é atualizada $\hat{f}^{(m)} = \hat{f}^{(m-1)} + v\hat{g}^{(m)}$.

Assim como apresentado na Seção 4.1.2, há penalização ℓ_2. Conforme descrito por (PARK; LEE; HA, 2009), o ℓ_2 Boosting consiste no ajuste repetido de mínimos quadrados dos resíduos.

4.2 MÉTRICAS DE PERFORMANCE

Nesta subseção, serão apresentadas as métricas utilizadas para avaliar e comparar os resultados dos modelos preditivos discutidos anteriormente. Para alcançar esse objetivo, será utilizada a métrica do Erro Quadrático Médio (RMSE), que mede a dispersão dos erros residuais, ou seja, a diferença entre os valores previstos pelo modelo e os valores reais observados nos dados, representada pela formulação:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{12}$$

onde n é o número de observações, y_i é o valor real observado e $\hat{y}_i$ é o valor previsto pelo modelo. Quanto menor o RMSE, melhor a precisão do modelo em relação aos dados observados. Em essência, ele nos ajuda a entender o quão bem nosso modelo está fazendo previsões e quão concentrados esses valores estão em torno da linha de melhor ajuste. Pensando na otimização do processo comparativo entre o modelo de referência (benchmark) e o modelo de interesse, será trabalhado com o RMSE Relativo (rRMSE):

$$rRMSE_h^{modelo} = \frac{RMSE_h^{modelo}}{RMSE_h^{SARIMA}} \tag{13}$$

Se rRMSE > 1, o modelo analisado apresenta desempenho preditivo inferior ao modelo de referência. Por outro lado, se rRMSE < 1, o desempenho preditivo do modelo analisado é superior ao do modelo de referência. Com isso, percebe-se que ele apresenta uma maior facilidade para a interpretação dos resultados.

Adicionalmente, conforme (CALDEIRA; MOURA; SANTOS, 2016), as métricas convencionais de performance, como o RMSE (Root Mean Square Error), apresentam uma desvantagem significativa: por se tratarem de estatísticas únicas, elas resumem os erros de previsão individuais ao longo de uma amostra inteira. Isso implica que o RMSE não fornece uma visão detalhada sobre onde, na amostra, um modelo específico comete seus maiores e menores erros de previsão. Essa limitação é particularmente preocupante em contextos onde a distribuição dos erros pode ter implicações importantes na interpretação dos resultados.

Com o objetivo de superar essa desvantagem, adotou-se o método proposto por (WELCH; GOYAL, 2008), que analisa graficamente os erros de previsão acumulados ao quadrado (CSFE - Cumulative Squared Forecast Error). O CSFE é representado pela seguinte equação:

$$\text{CSFE}_{m,T} = \sum_{t=1}^{T} \left[\left(\hat{y}_{t+h|t,\text{benchmark}} - y_{t+h} \right)^2 - \left(\hat{y}_{t+h|t,m} - y_{t+h} \right)^2 \right] \tag{14}$$

Essa abordagem permite uma análise mais profunda e clara do desempenho dos modelos. Ao observar os valores do CSFE ao longo do tempo, é possível identificar períodos específicos em que um modelo se destaca em relação ao benchmark, bem como aqueles em que ele falha. Se a série de CSFE for crescente, isso indica que o modelo em análise está superando o benchmark. Por outro lado, se a série for decrescente, isso sugere que o benchmark está apresentando um desempenho superior. Essa visualização não apenas facilita a identificação de tendências nos erros de previsão, mas também oferece uma ferramenta valiosa para a avaliação comparativa de diferentes modelos em cenários variados.

4.3 TESTE DE SIGNIFICÂNCIA

A aplicação de um teste de significância se faz necessária para esclarecer a aleatoriedade da performance preditiva. Em outras palavras, há o interesse de entender se existem evidências de superioridade preditiva ou apenas um mero acaso (DIEBOLD, 2015). O teste Diebold-Mariano compara o termo de erro de dois modelos sob a hipótese nula de que ambos tenham a mesma precisão, e a hipótese alternativa de que eles tenham níveis diferentes de precisão. Valores-p abaixo de 10% são considerados significativos. Portanto, um bom ajuste tem tanto RMSE relativo menor que um quanto um valor-p menor que 10%.

4.4 FORECASTING

Para o exercício de previsão, foram definidos os horizontes ($h = 1$, 2, 3, 4, 5, 6, 9 e 12) e segmentaram-se os dados em três grupos: variáveis textuais, variáveis econômicas e um grupo combinado. Essa estratégia, inspirada por (BANTIS; CLEMENTS; URQUHART, 2023), visa comparar o desempenho das previsões entre diferentes conjuntos de dados. A segmentação permitiu uma análise detalhada do desempenho de cada categoria. Avaliou-se o desempenho das previsões usando apenas dados textuais, especificamente os índices TESI e TEPU, além da eficácia do modelo ao integrar variáveis textuais e econômicas. Essa abordagem facilitou a identificação das variáveis com maior poder preditivo em distintos horizontes temporais. A comparação dos resultados entre os grupos de dados revelou informações valiosas sobre a eficácia das variáveis textuais nas previsões econômicas.

5 RESULTADOS

O uso de técnicas automatizadas de coleta de informações por meio de *Web Scraping* resultou em um volume considerável de dados. No entanto, constatou-se que menos da metade apresentava relevância direta com os tópicos de interesse. Assim, foram removidas as notícias com baixa relação, garantindo que apenas aquelas estritamente relacionadas aos temas integrassem os índices TEPU e TESI. Embora representem um grande volume para este estudo, isso corresponde a cerca de 10% das notícias processadas por (APRIGLIANO et al., 2023).

Table 1 – Volume de Notícias Pré e Pós processamento.

Tópico	Pré-Processamento	Pós-Processamento
Agronegócio	15.134	14.222
Indústria	53.894	25.272
Mercado de Trabalho	75.000	21.644
Mercado Financeiro	43.680	16.026
Serviços	36.017	28.789
Total	223.725	105.953

Fonte: Elaboração Própria. A tabela mostra a quantidade de notícias antes e depois do processo de tratamento de dados, apresentado na seção 3.4.

Não obstante, a pesquisa também identificou um crescimento significativo no volume de notícias e tokens ao longo do período analisado. Este aumento evidencia uma intensificação na frequência de publicações de notícias, bem como uma expansão na extensão abrangida por essas publicações.

Table 2 – Resumo das Notícias Processadas

Ano	$\sum$ **Notícias**	$\sum$ **Tokens**	μ **Tokens**	min **Tokens**	max **Tokens**
2010	3.442	631.107	183,3547	20	2330
2011	4.444	835.865	188,0884	21	2485
2012	4.920	961.152	195,3561	15	2137
2013	5.210	1.040.488	199,7098	17	2436
2014	7.280	1.443.789	198,3227	17	2445
2015	7.492	1.670.644	222,9904	19	2694
2016	6.973	1.628.620	233,5609	20	2987
2017	7.134	2.078.631	291,3696	30	2751
2018	8.007	2.535.745	316,6910	28	3062
2019	8.983	3.042.895	338,7393	34	2950
2020	9.143	3.315.205	362,5949	47	3513
2021	10.108	3.984.219	394,1649	27	2991
2022	10.727	4.412.018	411,3003	50	3343
2023	11.418	4.775.056	418,2042	65	2986
$\sum$	**105.953**	**32.658.422**	**4405,322**	**515**	**41.966**

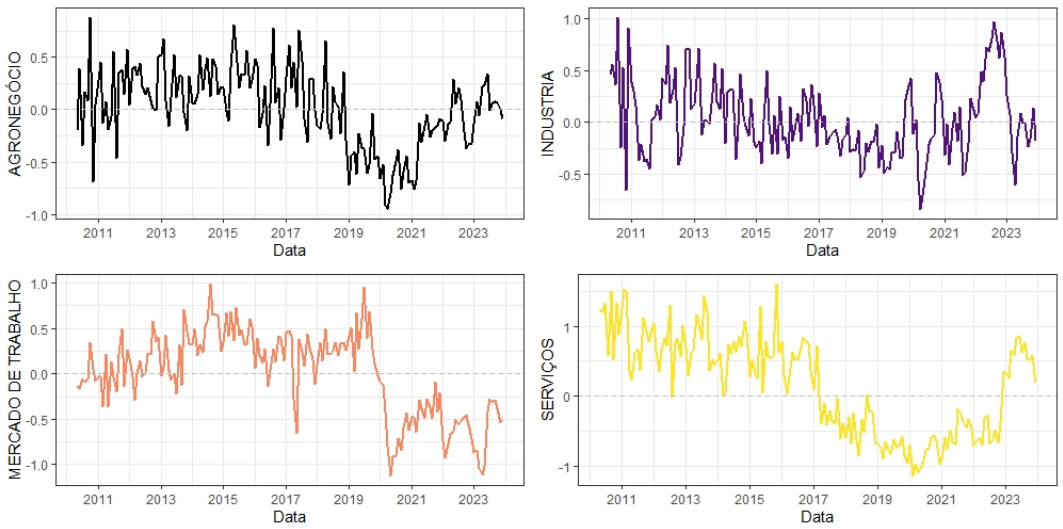
Fonte: Elaboração Própria. A tabela fornece um resumo descritivo do conjunto de dados (corpus) processado para a elaboração dos Índices TESI e TEPU.

Para o Índice de Sentimento Econômico Textual (TESI): em uma análise inicial, os índices apresentados, com exceção do mercado financeiro, mostram-se em harmonia com os períodos de instabilidade econômica observados: (a) durante a recessão de 2015 e 2016, caracterizada por uma contração de 6,8% na atividade econômica e um aumento significativo no nível de desemprego médio, de 8,43% para 12,88%, os índices demonstraram uma tendência de queda no mesmo período; (b) durante a pandemia de COVID-19 em 2020 e 2022, a economia brasileira contraiu-se em 3,3%, seguida por uma rápida recuperação de 5% no período subsequente, enquanto o nível médio de desemprego aumentou em 1,43 pontos percentuais. Nesse contexto, observa-se que o comportamento dos índices (exceto o mercado financeiro) está alinhado com as expectativas previamente estabelecidas.

O Índice de Incerteza da Política Econômica Textual (TEPU) reflete a incerteza em relação à política econômica do governo, diferentemente do TESI. Essa característica é evidente no TEPU da indústria, que, apesar de baixas oscilações, manteve uma tendência estável. No final de 2023, após o anúncio de um novo plano industrial, o índice dobrou. Os índices do mercado de trabalho e de serviços também mostraram elevações significativas durante os períodos pandêmicos de 2020-22, sinalizando incertezas sobre as medidas do governo. No caso dos serviços, a incerteza aumentou antes da pandemia, entre 2016 e 2017, devido a greves em diversas áreas, como saúde, transportes e reajustes de combustíveis. Quanto ao Agronegócio, o índice se manteve consistentemente acima de zero, refletindo o grau de incerteza, em parte devido à crescente atenção a questões sustentáveis nas notícias classificadas como EPU.

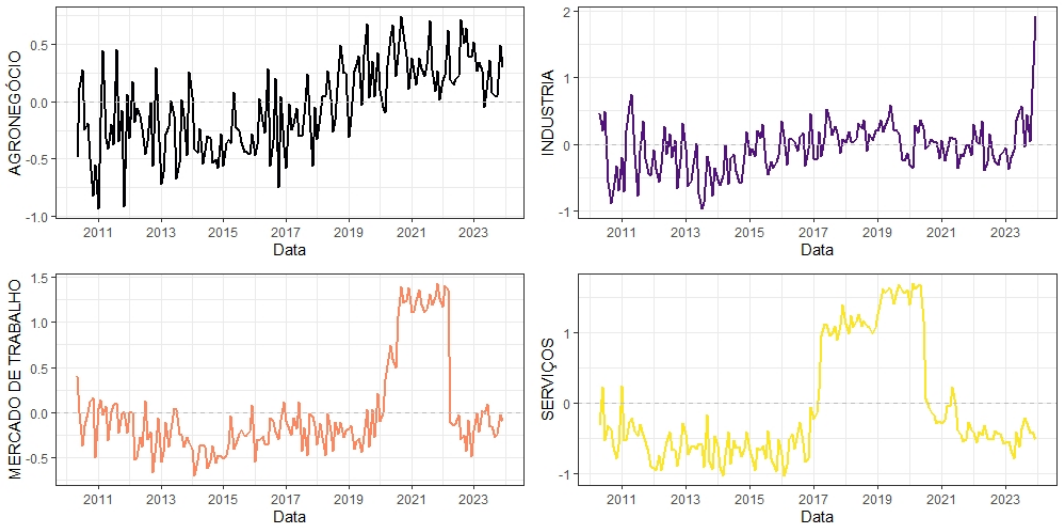
Para o Mercado Financeiro, os índices mostraram relação limitada com indicadores oficiais, como o IBOVESPA (IBVSP). O TESI apresenta baixa relação com os eventos macroeconômicos, enquanto o TEPU teve maior sensibilidade, aumentando em períodos de recessão. Esses resultados são aceitáveis pois os índices visam fornecer dados de alta frequência por meio da análise textual, permitindo atualizações frequentes, enquanto o mercado financeiro já possui seus próprios mecanismos de avaliação.

Figure 2 – Índices de Sentimento Econômico (TESI)



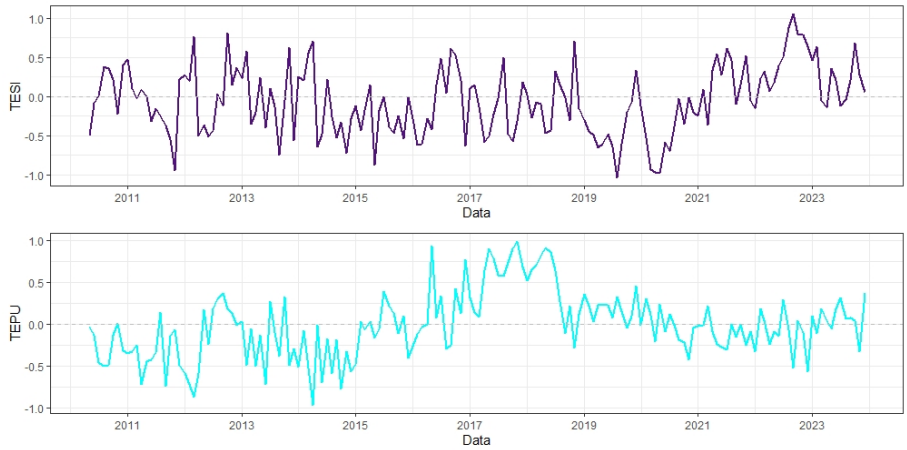
Fonte: Elaboração Própria. Esta figura apresenta os gráficos relacionados ao índice TESI aplicado ao Brasil. Esses índices foram calculados diariamente e agregados mensalmente para refletir a frequência das variáveis econômicas. Posteriormente, padronizados de modo a apresentarem média $\mu = 0$ e $\sigma^2 = 1$.

Figure 3 – Índices de Incerteza da Política Econômica (TEPU)



Fonte: Elaboração Própria. Esta figura apresenta os gráficos relacionados ao índice TEPU aplicado ao Brasil. Esses índices foram calculados diariamente e agregados mensalmente para refletir a frequência das variáveis econômicas. Posteriormente, padronizados de modo a apresentarem média $\mu = 0$ e $\sigma^2 = 1$.

Figure 4 – Mercado Financeiro: TESI e TEPU



Fonte: Elaboração Própria. Esta figura apresenta os gráficos relacionados aos índices TESI e TEPU aplicado ao Mercado Financeiro. Esses índices foram calculados diariamente e agregados mensalmente para refletir a frequência das variáveis econômicas. Posteriormente, padronizados $\mu = 0$ e $\sigma^2 = 1$.

5.1 FORECASTING

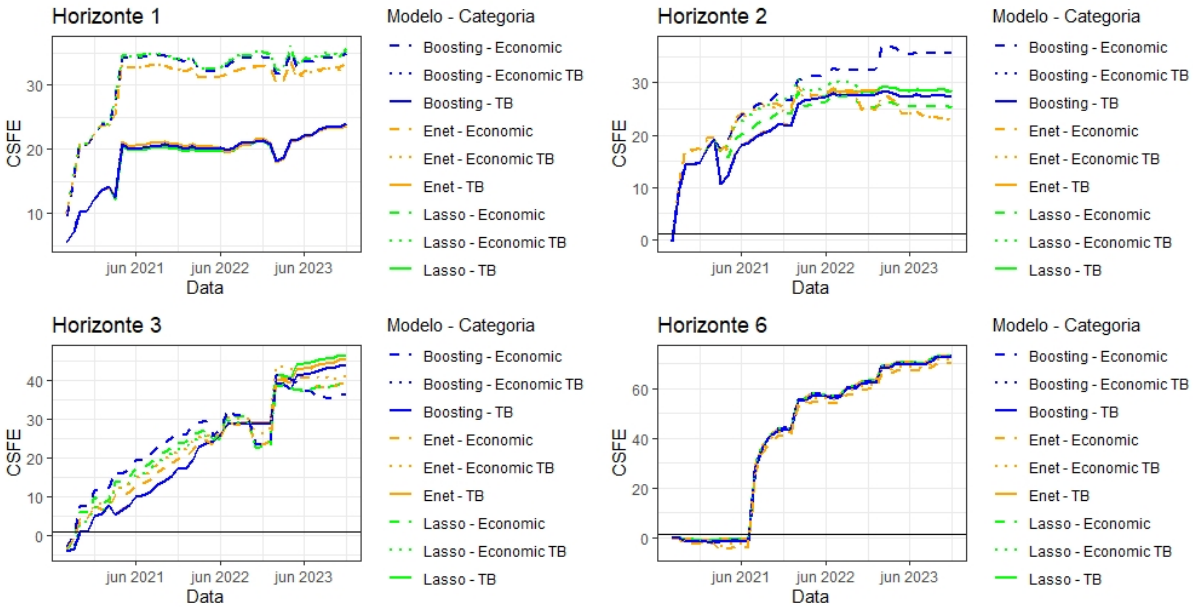
Os resultados do exercício de previsão mostram que a inclusão de notícias aos dados econômicos tem um efeito pouco perceptível, com diferenças mínimas, independentemente do modelo ou horizonte. Isso se alinha ao encontrado por Aprigliano (2023), onde as variáveis de texto e econômicas mostraram melhorias pontuais, mas não significativas. Além disso, observa-se que, no curto prazo, há uma disparidade no poder preditivo entre variáveis textuais e econômicas, mas essa diferença diminui com a expansão do horizonte de previsão, tornando as variáveis textuais um complemento útil na ausência de dados econômicos.

Table 3 – Forecasting Atividade Econômica: RMSE Relativo

Variáveis	Modelo	Horizontes							
		1	**2**	**3**	**4**	**5**	**6**	**9**	**12**
Textual	LASSO	0.84*	0.81	0.74**	0.57	**0.66**	0.65	0.99	**1.00**
	ENET	0.84*	0.82	**0.74*** **	0.57	0.67	0.65	0.97	1.01
	BOOSTING	0.83*	0.82	0.75***	**0.57**	0.67	0.66	**0.97**	1.01
Econômicas	LASSO	0.74*	0.83	0.78	1.05	0.71	**0.65**	1.00	1.03
	ENET	0.76*	0.85	0.78*	0.63	0.68	0.67	1.00	1.09
	BOOSTING	0.75*	**0.81***	0.77	0.60	0.69	0.66	1.01	1.07
Econômicas & Texto	LASSO	0.74*	0.81	0.78	0.99	0.68	0.65	1.00	1.02
	ENET	0.76*	0.85	0.77*	0.63	0.67	0.66	1.00	1.03
	BOOSTING	**0.74***	0.81	0.77*	0.60	0.69	0.66	1.01	1.07

Fonte: Elaboração Própria. A tabela apresenta os resultados das previsões da atividade econômica para o Brasil: valores menores a um indicam que o modelo previu melhor que o benchmark. *, **, e ***, denotam níveis de significância a 10%, 5% e 1% dos valores de p para o teste Diebold-Mariano.

Figure 5 – CSFE: Horizontes de previsão & Modelos



Fonte: Elaboração Própria. Esta figura apresenta o gráfico da métrica CSFE para todos os modelos e categorias. As cores representam os modelos: azul representa os modelos BOOSTING, laranja representa os modelos ENET e verde representa os modelos LASSO. O estilo da linha representa a categoria dos dados: tracejado representa a categoria de Econômicos, pontilhado representa a categoria dos dados Econômicos & Textuais e Constante representa a categoria dos dados Textuais.

6 CONCLUSÃO

O uso de texto e notícias para previsões econômicas é uma prática relativamente nova que tem ganhado popularidade, sendo adotada por diversas instituições, incluindo Bancos Centrais. A alta produção de notícias possibilita a construção de índices com frequências tão elevadas quanto a capacidade de processamento dos computadores, alcançando um pseudo tempo real.

Foram empregadas técnicas de mineração de dados para construção dos bancos de dados, que, devido ao seu caráter menos estruturado, exigiu um tratamento mais detalhado. Além disso, as técnicas de mineração trouxeram notícias pouco relacionadas ao tópico de busca, tornando necessário um filtro para garantir a qualidade dos dados.

Os índices baseados em texto mostram relação com os eventos macroeconômicos observados, exceto no que diz respeito ao mercado financeiro. O exercício de previsão revelou que as previsões da atividade econômica geradas a partir dos dados textuais superam o benchmark e são estatisticamente significativas nos horizontes de 1 e 3 meses. Embora as variáveis textuais apresentem um nível preditivo inicialmente inferior ao das variáveis econômicas no curto prazo, essa diferença se reduz ao longo do tempo. A combinação das variáveis textuais com as econômicas não demonstrou um incremento significativo na performance preditiva, com alguns modelos atribuindo valor nulo às variáveis textuais em certos casos. Esses resultados não divergem significativamente dos apresentados por (APRIGLIANO et al., 2023).

Os resultados sugerem que tanto o Índice de Sentimento Econômico Textual (TESI) quanto o Índice de Incerteza de Política Econômica Textual (TEPU) podem ser utilizados para fornecer informações complementares na análise econômica. Esses índices podem ser calculados em pseudo tempo real e não dependem de indicadores econômicos oficiais. Na ausência de variáveis econômicas, as variáveis textuais podem se revelar valiosas para a prática de previsões, servindo como uma alternativa ou proxy.

Para trabalhos futuros, o primeiro passo é diversificar a origem das notícias, uma vez que tanto (APRIGLIANO et al., 2023) quanto (SHAPIRO; SUDHOF; WILSON, 2022) utilizam dados de mais de uma fonte. Em um segundo momento, as notícias presentes no banco de dados podem ser tokenizadas para a elaboração de um dicionário próprio, semelhante ao que foi feito por (STONE; DUNPHY; SMITH, 1966), (HENRY, 2008) e (LOUGHRAN; MCDONALD, 2011), seguindo os passos descritos em (SHAPIRO; SUDHOF; WILSON, 2022).

REFERÊNCIAS

APRIGLIANO, Valentina et al. The power of text-based indicators in forecasting Italian economic activity. **International Journal of Forecasting**, Elsevier, v. 39, n. 2, p. 791–808, 2023.

ARDIA, David et al. The R Package sentometrics to Compute, Aggregate, and Predict with Textual Sentiment. **Journal of Statistical Software**, v. 99, n. 2, p. 1–40, 2021.

BAKER, Scott R; BLOOM, Nicholas; DAVIS, Steven J. Measuring economic policy uncertainty. **The quarterly journal of economics**, Oxford University Press, v. 131, n. 4, p. 1593–1636, 2016.

BANTIS, Evripidis; CLEMENTS, Michael P.; URQUHART, Andrew. Forecasting GDP growth rates in the United States and Brazil using Google Trends. **International Journal of Forecasting**, v. 39, n. 4, p. 1909–1924, 2023.

BOX, George EP et al. **Time series analysis: forecasting and control**. [S.l.]: John Wiley & Sons, 2015.

CALDEIRA, Joao F; MOURA, Guilherme V; SANTOS, André AP. Predicting the yield curve using forecast combinations. **Computational Statistics & Data Analysis**, Elsevier, v. 100, p. 79–98, 2016.

DIEBOLD, Francis X. Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold–Mariano tests. **Journal of Business & Economic Statistics**, Taylor & Francis, v. 33, n. 1, p. 1–1, 2015.

FRIEDMAN, Jerome H. Greedy function approximation: a gradient boosting machine. **Annals of statistics**, JSTOR, p. 1189–1232, 2001.

GENTZKOW, Matthew; KELLY, Bryan; TADDY, Matt. Text as Data. **Journal of Economic Literature**, v. 57, n. 3, p. 535–74, Sept. 2019.

HENRY, Elaine. Are Investors Influenced By How Earnings Press Releases Are Written? **The Journal of Business Communication (1973)**, v. 45, n. 4, p. 363–407, 2008.

HU, Minqing; LIU, Bing. Mining and summarizing customer reviews. In: PROCEEDINGS of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, WA, USA: Association for Computing Machinery, 2004. (KDD '04), p. 168–177.

KIM, Soo-Min; HOVY, Eduard. Determining the sentiment of opinions. In: COLING 2004: Proceedings of the 20th international conference on computational linguistics. [S.l.: s.n.], 2004. P. 1367–1373.

LINDENMEYER, Guilherme; SKORIN, Pedro Pablo; TORRENT, Hudson da Silva. Using boosting for forecasting electric energy consumption during a recession: a case study for the Brazilian State Rio Grande do Sul. **Letters in Spatial and Resource Sciences**, v. 14, n. 2, p. 111–128, Aug. 2021.

LOUGHRAN, Tim; MCDONALD, Bill. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. **The Journal of finance**, Wiley Online Library, v. 66, n. 1, p. 35–65, 2011.

NADKARNI, Prakash M; OHNO-MACHADO, Lucila; CHAPMAN, Wendy W. Natural language processing: an introduction. **Journal of the American Medical Informatics Association**, BMJ Group BMA House, Tavistock Square, London, WC1H 9JR, v. 18, n. 5, p. 544–551, 2011.

NANLI, Zhu et al. Sentiment analysis: A literature review. In: 2012 International Symposium on Management of Technology (ISMOT). [S.l.: s.n.], 2012. P. 572–576.

PARK, Byeong U.; LEE, Young K.; HA, Seung. L2 boosting in kernel regression. **Bernoulli**, Bernoulli Society for Mathematical Statistics and Probability, International Statistical Institute (ISI), v. 15, n. 3, p. 599–613, 2009.

SHAPIRO, Adam Hale; SUDHOF, Moritz; WILSON, Daniel J. Measuring news sentiment. **Journal of Econometrics**, v. 228, n. 2, p. 221–243, 2022.

STONE, Philip J; DUNPHY, Dexter C; SMITH, Marshall S. The general inquirer: A computer approach to content analysis. MIT press, 1966.

TIBSHIRANI, Robert. Regression Shrinkage and Selection via the Lasso. **Journal of the Royal Statistical Society. Series B (Methodological)**, [Royal Statistical Society, Wiley], v. 58, n. 1, p. 267–288, 1996.

WELCH, Ivo; GOYAL, Amit. A comprehensive look at the empirical performance of equity premium prediction. **The Review of Financial Studies**, Society for Financial Studies, v. 21, n. 4, p. 1455–1508, 2008.

ZOU, Hui; HASTIE, Trevor. Regularization and Variable Selection Via the Elastic Net. **Journal of the Royal Statistical Society Series B: Statistical Methodology**, v. 67, n. 2, p. 301–320, Mar. 2005.

APÊNDICE A - FOLHA DE APROVAÇÃO

NATHAN RAMOS DE ALMEIDA

O PODER DAS PALAVRAS: USANDO TEXTO PARA PREVISÕES ECONÔMICAS

Trabalho de conclusão submetido ao Curso de Graduação em Ciências Econômicas da Faculdade de Ciências Econômicas da UFRGS, como requisito parcial para obtenção do título Bacharel em Economia.

Aprovado em: Porto Alegre, 15 de Agosto de 2024.

BANCA EXAMINADORA:

Prof. Dr. Hudson da Silva Torrent
Orientador - UFRGS

Prof. Dr. Fernando Augusto Boeira Sabino da Silva
UFRGS

Prof. Dr. Nelson Seixas dos Santos
UFRGS