Insper

Programa Institucional de Bolsas de Iniciação Científica (PIBIC)

Ciências Econômicas

Extração, Estruturação e Interpretação Automatizada de Notícias: Um Modelo Baseado em Aprendizado Profundo, Reconhecimento Ótico de Caracteres e Alocação Latente de Dirichlet

Klaus Colletti Oehling
Orientador: Miguel Bandeira

São Paulo

2025

Programa Institucional de Bolsas de Iniciação Científica (PIBIC)

Klaus Colletti Oehling
Ciências Econômicas

Extração, Estruturação e Interpretação Automatizada de Notícias: Um Modelo Baseado em Aprendizado Profundo, Reconhecimento Ótico de Caracteres e Alocação Latente de Dirichlet

Relatório final apresentado ao Programa Institucional de Bolsas de Iniciação Científica (PIBIC)

Orientador: Prof. Miguel Bandeira

São Paulo
2025

Resumo

Este relatório de iniciação científica desenvolve um método automatizado para extração, estruturação e interpretação de notícias veiculadas no jornal impresso *O Estado de S. Paulo*, no período de 2001 a 2023. A pesquisa integra o estudo macroeconômico de expectativas fiscais apresentado em *Fiscal Shocks or Fiscal News? Evidence from Daily Fiscal Expectations* (Bandeira; Bonomo; Carvalho; Mendes, 2025). A metodologia adotada consistiu, inicialmente, na construção de um conjunto de dados anotado por meio da ferramenta Label Studio, na qual foram criadas caixas delimitadoras com regiões correspondentes a títulos, textos, tipos, notas, autores, subtítulos e outros nas capas e páginas da seção de economia do periódico. Esse corpus serviu de base para treinar um modelo de aprendizado profundo com a biblioteca LayoutParser, permitindo a detecção automática das estruturas jornalísticas. Mais adiante, foi implementado um algoritmo autoral para agrupar coerentemente caixas delimitadoras da mesma notícia e aplicar o Reconhecimento Ótico de Caracteres (OCR) ao conteúdo extraído. Por fim, utilizou-se a modelagem de tópicos via Alocação Latente de Dirichlet (LDA) para identificar e classificar notícias fiscais. A série temporal resultante demonstra forte alinhamento com eventos fiscais relevantes das últimas décadas.

Palavras-chave: LayoutParser. OCR. Modelagem de Tópicos. Notícia Fiscal. Jornal

Abstract

This scientific initiation report develops an automated method for extracting, structuring, and interpreting news articles published in the print edition of *O Estado de S. Paulo* from 2001 to 2023. The research is part of the macroeconomic study on fiscal expectations presented in *Fiscal Shocks or Fiscal News? Evidence from Daily Fiscal Expectations* (Bandeira; Bonomo; Carvalho; Mendes, 2025). The adopted methodology initially involved building an annotated dataset using the Label Studio tool, in which bounding boxes were created to identify regions corresponding to headlines, main texts, types, notes, authors, subtitles, and others on the front pages and economy section pages of the newspaper. This corpus was used to train a deep learning model via the LayoutParser library, enabling the automatic detection of journalistic structures. Subsequently, a custom algorithm was implemented to coherently group bounding boxes belonging to the same article and to apply Optical Character Recognition (OCR) to the extracted content. Finally, topic modeling using Latent Dirichlet Allocation (LDA) was employed to identify and classify fiscal news articles. The resulting time series shows strong alignment with major fiscal events from recent decades.

Keywords: LayoutParser. OCR. Topic Modeling. Fiscal News. Newspaper

Lista de Ilustrações e Tabelas

Figura 1 – Exemplo de capa anotada manualmente 7

Figura 2 – Exemplo de página anotada.....8

Quadro 1 – Modelos disponíveis 9

Gráfico 1 – Decaimento periódico da taxa de aprendizado para o modelo 1.....10

Gráfico 2 – Decaimento periódico da taxa de aprendizado para o modelo 2.....10

Gráfico 3 – Precisão média por classe de caixa delimitadora do modelo 1 11

Gráfico 4 – Precisão média por classe de caixa delimitadora do modelo 2 11

Gráfico 5 – Perda por atividade do modelo 1.....12

Gráfico 6 – Perda por atividade do modelo 2.....12

Quadro 2 – Parâmetros resultantes do modelo.....13

Figura 3 – Representação das coordenadas e intersecção 15

Figura 4 – Exemplo de trecho da planilha resultante para capas.....19

Figura 5 – Exemplo de trecho da planilha resultante para a seção de economia ..19

Quadro 3 – Palavras representativas de tópicos fiscais em capa 21

Quadro 4 – Palavras representativas de tópicos fiscais na seção de economia 22

Gráfico 7 – Notícias fiscais diárias em capa COMPOSITE 23

Gráfico 8 – Notícias fiscais diárias em capa PURE.....23

Gráfico 9 – Notícias fiscais mensais em capa COMPOSITE 24

Gráfico 10 – Notícias fiscais mensais em capa PURE.....24

Gráfico 11 – Mudança de notícias fiscais mensais em capa COMPOSITE 25

Gráfico 12 – Mudança de notícias fiscais mensais em capa PURE.....25

Gráfico 13 – Precisão média do modelo 1 30

Gráfico 14 – Precisão média para diferentes tamanhos do modelo 1 30

Gráfico 15 – Tempo estimado para conclusão do treinamento do modelo 1 30

Gráfico 16 – Outras modalidades de perda do modelo 1 31

Gráfico 17 – Acurácia e falso negativo do modelo 1 31

Gráfico 18 – Precisão média do modelo 2 32

Gráfico 19 – Precisão média para diferentes tamanhos do modelo 2 32

Gráfico 20 – Tempo estimado para conclusão do treinamento do modelo 2 32

Gráfico 21 – Outras modalidades de perda do modelo 2 33

Gráfico 22 – Acurácia e falso negativo do modelo 2 33

Gráfico 23 – Mudança de notícias fiscais diárias em capa COMPOSITE 34

Gráfico 24 – Mudança de notícias fiscais diárias em capa PURE.....34

Gráfico 25 – Notícias fiscais diárias em páginas de economia COMPOSITE 34

Gráfico 26 – Notícias fiscais diárias em páginas de economia PURE.....35

Gráfico 27 – Mudança de notícias fiscais diárias em páginas de economia COMPOSITE.....35

Gráfico 28 – Mudança de notícias fiscais diárias em páginas de economia PURE35

Gráfico 29 – Notícias fiscais mensais em páginas de economia COMPOSITE 36

Gráfico 30 – Notícias fiscais mensais em páginas de economia PURE.....36

Gráfico 31 – Mudança de notícias fiscais mensais em páginas de economia COMPOSITE.....36

Gráfico 32 – Mudança de notícias fiscais mensais em páginas de economia PURE 37

Sumário

1 Introdução 5

2 Metodologia.....6

2.1 Modelo de aprendizado profundo.....6

2.1.1 Anotação das bases de treinamento 6

2.1.2 Qualidade dos modelos.....8

2.2 Algoritmo de agrupamento 13

2.3 Reconhecimento ótico de caracteres (OCR) 18

3 Modelagem de tópicos 19

3.1 Processamento do texto.....20

3.2 Alocação latente de Dirichlet (LDA).....20

4 Resultados 23

5 Conclusão 27

Referências 29

Apêndice A – Resultados adicionais: Modelo 1.....30

Apêndice B – Resultados adicionais: Modelo 2.....32

Apêndice C – Resultados adicionais: Sinais fiscais 34

1 Introdução

A digitalização de documentos históricos e periódicos impressos tem se consolidado como uma ferramenta essencial para a preservação da informação e para a viabilização de análises quantitativas e qualitativas em larga escala. No contexto econômico, as notícias veiculadas na imprensa desempenham um papel fundamental na divulgação de informações sobre políticas econômicas, influenciando expectativas de agentes e mercados, e possivelmente impactando na disseminação de choques.

O jornal *O Estado de S. Paulo*, um dos mais influentes periódicos brasileiros, possui um vasto acervo impresso que contém informações valiosas sobre acontecimentos econômicos ao longo das últimas décadas. No entanto, a ausência de uma estrutura padronizada para a organização dessas notícias dificulta a análise sistemática dos conteúdos. A alta complexidade e variabilidade dos layouts, especialmente para capas de jornal, torna a organização automatizada um desafio técnico. Nesse sentido, o desenvolvimento de modelos computacionais capazes de lidar com essas estruturas de documentos é de alta relevância científica, podendo simplificar a criação de trabalhos empíricos no ramo econômico.

Adicionalmente, não há estruturação explícita de que temas foram tratados em cada dia, tornando difícil a criação de séries temporais com tópicos econômicos. Portanto, também é de alta relevância implementar métodos de modelagem de tópicos para que se criem dados estruturados de temas discutidos em cada notícia.

Tendo isso em vista, este trabalho tem como objetivo propor um modelo baseado em aprendizado profundo, reconhecimento ótico de caracteres (OCR) e alocação latente de Dirichlet (LDA) para a extração, estruturação e interpretação automatizada de notícias do jornal impresso *O Estado de S. Paulo*. Após a primeira etapa, teremos duas bases de dados em paralelo. Uma com todas as notícias em capa de janeiro de 2001 a dezembro de 2023 e outra com todas as notícias na seção de economia no mesmo período. Finalmente, essas bases serão utilizadas para criar séries temporais de notícias fiscais ao longo do tempo. Espera-se que o produto do presente trabalho seja utilizado em *Fiscal Shocks or Fiscal News? Evidence from Daily Fiscal Expectations* (Bandeira; Bonomo; Carvalho; Mendes, 2025). Além disso, supõe-se que a fase final de modelagem de tópicos via LDA sirva de ponto de partida para desvendar outros tópicos fora o fiscal, podendo ser replicado para diversos trabalhos promovidos pelo Centro de Finanças e Macroeconomia do Insper (CeFin).

2 Metodologia

Esta seção descreve as etapas utilizadas para a implementação do sistema de extração de texto para capas e páginas da seção de economia. O processo foi dividido em três partes principais: o treinamento de um modelo de aprendizado profundo para segmentação de elementos do documento, a aplicação de um algoritmo de agrupamento para associar elementos da mesma notícia, assim como minimizar erros provenientes da primeira etapa, e a utilização de OCR para converter a informação em texto digital. As primeiras duas etapas são diferentes para cada conjunto de dados, enquanto a última é idêntica. Cada uma das etapas é detalhada nas subseções a seguir.

2.1 Modelo de aprendizado profundo

A primeira etapa se trata da criação de um modelo de aprendizado profundo que irá detectar e diferenciar caixas delimitadoras nas imagens. Portanto, através de um treinamento, o modelo será capaz de apontar coordenadas delimitando as regiões onde se encontram conjuntos textuais. Além disso, ele indicará o tipo de texto presente dentro desses objetos, e.g. título ou autor.

2.1.1 Anotação das bases de treinamento

Para o treinamento e validação dos dois modelos, foi necessário criar um conjunto de imagens anotadas através da ferramenta Label Studio. Dada a alta variabilidade do layout do Estadão observada de 2001 a 2023, foram selecionadas imagens aleatórias uniformemente distribuídas no período. O Estadão disponibiliza um link único para acesso à cada página de seu jornal impresso digitalizado. Com o uso destes, coletamos 276 páginas para ambos os modelos – uma por mês de 2001 a 2023. Desta forma, podemos captar os diversos padrões de design no decorrer dos anos.

A anotação das caixas delimitadoras das 276 capas foi feita manualmente com discriminação entre texto, título e outros. A figura 1 demonstra um exemplo de registro.

Figura 1 – Exemplo de capa anotada manualmente

Fonte: O Estado de S. Paulo, adaptado pelo autor

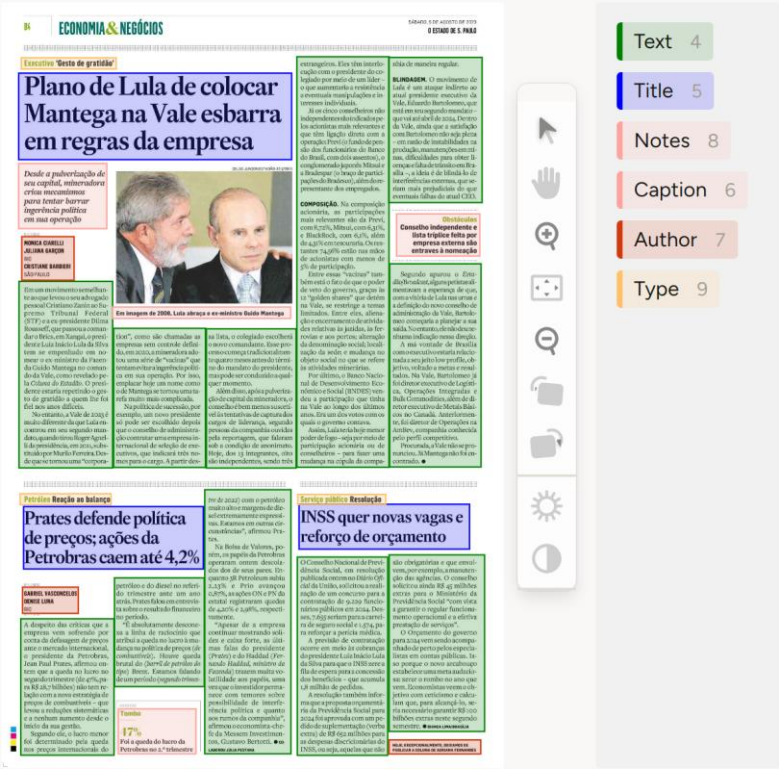
Por sua vez, as 276 páginas da seção de economia possuíam padrões estéticos mais variados, sendo necessário uma discriminação maior entre as caixas. Para essas, a anotação foi dividida entre título, texto, destaques, autores, tipo e subtítulo. Dada essa maior complexidade no registro das caixas delimitadoras seria ineficiente fazê-lo manualmente como feito com as capas e tornou-se necessário adaptar o processo.

Inicialmente, 92 páginas foram selecionadas e anotadas manualmente no Label Studio, respeitando os critérios de segmentação. O modelo resultante desse primeiro ciclo foi treinado e posteriormente integrado ao software por meio de uma URL de backend personalizada. Isto é, o próprio modelo passou a ser utilizado para realizar mais anotações no lugar do pesquisador. Ele previu caixas delimitadoras nas 92 páginas seguintes, criando uma nova base de treinamento com 184 páginas. Todas elas eram manualmente validadas para que os erros do primeiro modelo não fossem carregados para as seguintes versões.

Após esse primeiro ciclo, outro modelo foi treinado, agora com o dobro de dados. Este, novamente previu caixas delimitadoras de mais 92 itens. Após mais uma rodada de validação manual, foi obtido uma base de treinamento de 276 páginas.

Essa abordagem semiautomatizada de aprendizado supervisionado reduziu significativamente o tempo de anotação manual ao mesmo tempo em que manteve a acurácia da base de dados. Trata-se de um processo de auto-refinamento, em que o próprio modelo contribui para a construção incremental de seu conjunto de treinamento com supervisão ativa do pesquisador. A figura 2 demonstra um exemplo de registro das páginas da seção de economia.

Figura 2 – Exemplo de página anotada



Fonte: O Estado de S. Paulo, adaptado pelo autor

2.1.2 Qualidade dos modelos

A partir dos 276 itens e seguindo as recomendações em Shen, Zhang e Dell (2020) foi utilizada a biblioteca LayoutParser – um kit de ferramentas unificado para análise de imagens de documentos com base em aprendizado profundo. Como esclarecido anteriormente, esse processo foi feito repetidamente para o caso das páginas da seção de economia com um treinamento final nos 276 itens.

Essa ferramenta possui modelos de segmentação pré-treinados pelos desenvolvedores, vide Quadro 1, mas como nenhum deles performou de forma satisfatória no conjunto pretendido, o treinamento de um modelo próprio foi indispensável.

Quadro 1 – Modelos disponíveis

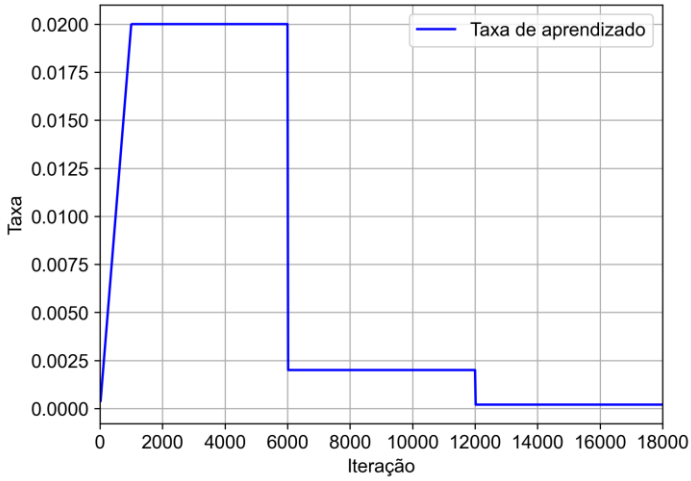
Modelo	Descrição
HJDataset	Para documentos históricos japoneses
PubLayNet	Para artigos científicos contidos no PubMed Central Open Access Subset
PrimaLayout	Para uma alta variedade de documentos
NewspaperNavigator	Para elementos visuais do jornais históricos presentes no Chronicling America
TableBank	Para tabelas
MFD	Para fórmulas matemáticas

Fonte: Elaborado pelo autor

Usando o conjunto produzido no Label Studio, para ambos os modelos foi feita uma divisão de 85% das imagens para treinamento e 15% para teste, i.e., 235 imagens para treinamento e 41 para teste. O exercício percorreu 18000 iterações para o modelo das capas e 30000 para o modelo da seção de economia. Isso foi feito para compensar a maior complexidade do segundo modelo. Além disso, o primeiro modelo treinou com pouco mais de 3% de todas as possíveis capas – 8400 capas de 2001 a 2023 para 276 usadas no treinamento – e o segundo com pouco menos de 0,2% de todas as possíveis páginas da seção de economia – média de 18 páginas dessa seção por dia.

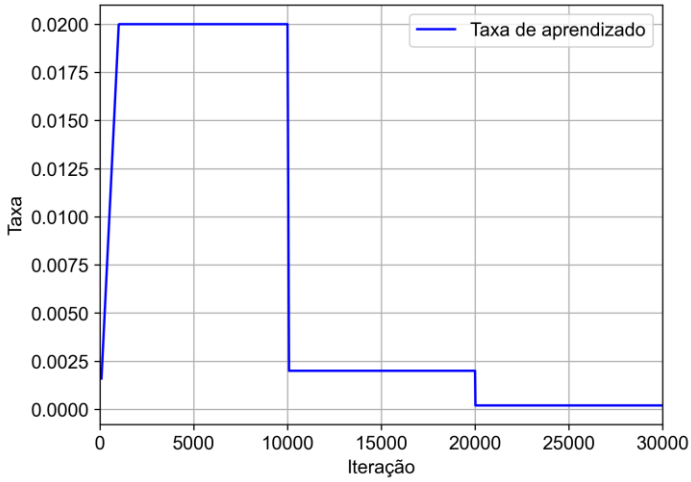
Também foram introduzidos intervalos para decaimento da taxa de aprendizado – em um terço e dois terços do treinamento total – a fim de otimizar o encontro de soluções. Com altas taxas de aprendizado no início, o modelo aumenta sua acurácia rapidamente. Conforme a perda se estabiliza, é ideal que a taxa de aprendizado seja diminuída gradualmente, de forma a possibilitar um aprendizado fino. Para evitar divergência inicial do treinamento, o modelo também inclui uma rodada de 1000 iterações de preparação na qual a taxa de aprendizado é aumentada gradualmente até atingir o primeiro alvo.

Gráfico 1 – Decaimento periódico da taxa de aprendizado para o modelo 1



Fonte: Elaborado pelo autor

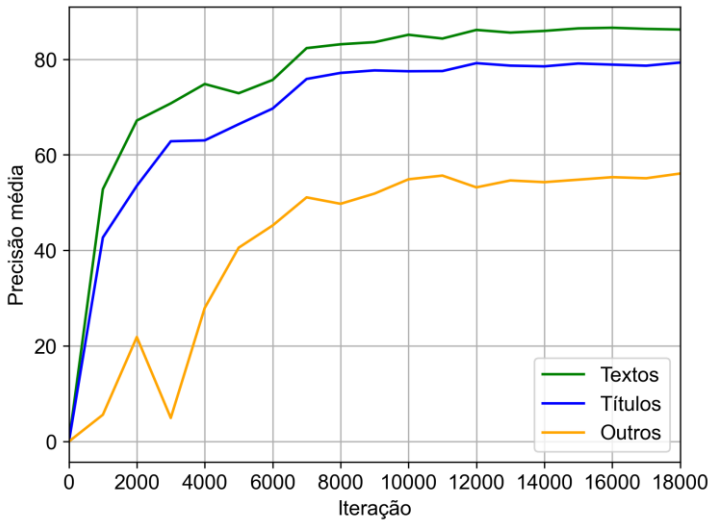
Gráfico 2 – Decaimento periódico da taxa de aprendizado para o modelo 2



Fonte: Elaborado pelo autor

É possível visualizar também a trajetória de aprendizado dos modelos. Como vemos no gráfico 3, há grande diferença na precisão média por classe. Enquanto títulos e textos terminaram com altos valores, a classe *outros* apresentou resultados muito abaixo dos demais. Apesar disso, esta classe é a de menor importância pois aparece com pouca frequência e adiciona informações pontuais à notícia – muitas vezes dados ou citações que não afetam na interpretação.

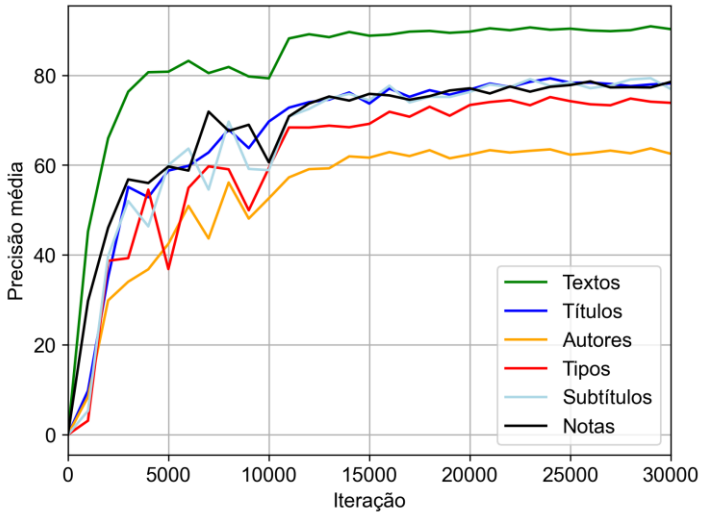
Gráfico 3 – Precisão média por classe de caixa delimitadora do modelo 1



Fonte: Elaborado pelo autor

Similarmente, no segundo modelo (Gráfico 4) também há certa divergência na precisão de identificação por classe. A classe de textos é a mais precisa, enquanto títulos, subtítulos, notas e tipos têm um nível de precisão 15 pontos abaixo. Por último temos as caixas delimitadoras de autores com um nível de precisão de 63 pontos. Assim como no modelo 1, isso não representa um problema significativo já que as informações menos relevantes são as menos precisas.

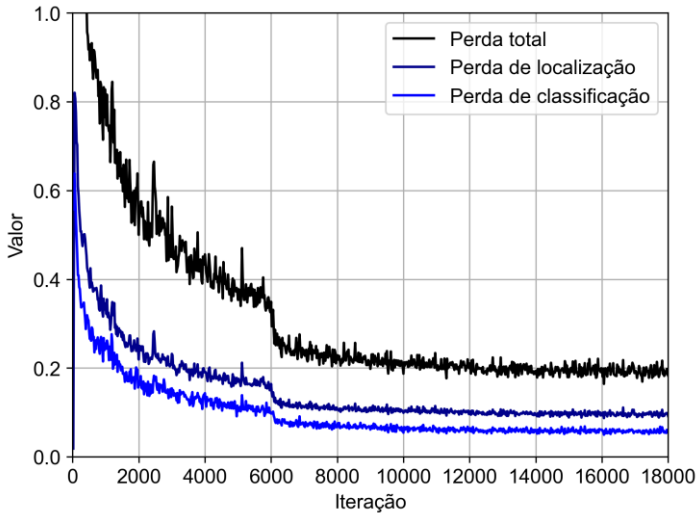
Gráfico 4 – Precisão média por classe de caixa delimitadora do modelo



Fonte: Elaborado pelo autor

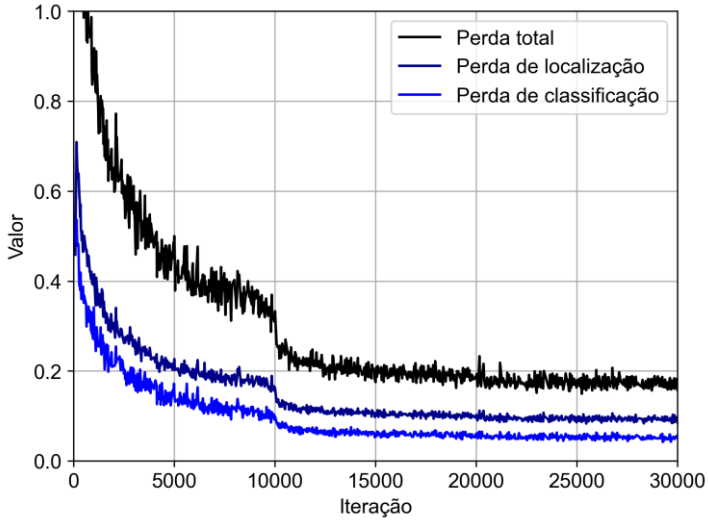
No gráfico 5 é possível visualizar a diminuição constante da perda total no modelo 1 com alto decaimento por volta da iteração 6000. Isso se dá pelo ajuste da taxa de aprendizado nessa época. Além disso, a perda de classificação foi menor do que a de localização, indicando que o modelo encontrou mais dificuldades em localizar corretamente as caixas delimitadoras, do que em classificá-las. O mesmo padrão pode ser visto no segundo modelo no gráfico 6. Mais estatísticas de desempenho estão disponíveis nos apêndices A e B.

Gráfico 5 – Perda por atividade do modelo 1



Fonte: Elaborado pelo autor

Gráfico 6 – Perda por atividade do modelo 2



Fonte: Elaborado pelo autor

O treinamento durou cerca de 25 horas para o primeiro modelo e 50 horas para o segundo. Ambos geraram um arquivo PTH contendo pesos e inferências resultantes. Este foi usado em conjunto com a biblioteca LayoutParser na linguagem de programação Python para prever as caixas delimitadoras de todo o conjunto de dados de 2001 a 2023. O modelo 1 rodou em todas as capas de jornal e o modelo 2 rodou em todas as páginas da seção de economia. Em seguida foi desenvolvido um método de agrupamento para estruturação das notícias (ver subcapítulo 3.2).

2.2 Algoritmo de agrupamento

A partir do devido reconhecimento da classe e localização das caixas delimitadoras, se faz necessário um algoritmo de atribuição cruzada dos elementos a fim de agrupá-los de forma coerente, i.e. em notícias individuais. Esse algoritmo é muito semelhante para as capas e páginas de economia. Por isso, quando uma etapa for diferente para cada modelo, isso será sinalizada claramente na explicação. Além disso, é relevante o entendimento da estrutura das informações geradas pelo modelo para justificar o algoritmo. Ele retorna, por imagem, uma lista de estruturas do tipo Layout, na qual há informações chave de cada caixa delimitadora (Quadro 2) encontrada no documento. As coordenadas $(X1, Y1)$ representam o canto superior esquerdo da caixa e as $(X2, Y2)$ representam o canto inferior direito. As coordenadas são medidas em pixels, sendo o canto superior esquerdo da imagem a origem do plano cartesiano.

Quadro 2 – Parâmetros resultantes do modelo

Parâmetros	Descrição
block	Sistema de coordenadas $(X1, Y1)$, $(X2, Y2)$
text	O texto detectado pelo OCR dentro dos limites da caixa
id	Identificação do bloco
type	A classe do bloco
parent	Identificação do bloco parente
next	Identificação do bloco seguinte
score	A confiança de previsão do bloco

Fonte: Elaborado pelo autor

O primeiro filtro serve para casos em que a caixa delimitadora é desenhada de forma inversa – com $(X1, Y1)$ no canto inferior direito e $(X2, Y2)$ no canto superior esquerdo. Após correção dessas falhas, os objetos foram ordenados de acordo com a importância da classe – título, texto e outros para capas e tipo, título, texto, notas, subtítulos e autores para páginas de economia, respectivamente. Adicionalmente, dentro de cada classe, também foram ordenados de acordo com o score. Isso foi feito para que os filtros de exclusão e ajustes usados em seguida alterem primeiro os elementos menos relevantes.

Para excluir caixas sobrepostas, foi usada uma regra de sobreposição. Sejam A e B caixas delimitadoras. Como a caixa A aparece antes na lista ela é, por definição, de menor relevância do que B e, portanto, deve ser a que sofrerá alterações. Ainda, caso alguma das condições nas equações (1), (1.2) e (1.3) sejam satisfeitas, a caixa A será excluída. Isso será feito para todos os pares de caixas existentes.

$$\frac{\mathrm{A} \cap \mathrm{B}}{\mathrm{A} \cup \mathrm{B}} > 0.6 \tag{1.1}$$

$$\frac{\mathrm{A} \cap \mathrm{B}}{\mathrm{A}} > 0.8 \tag{1.2}$$

$$\frac{\mathrm{A} \cap \mathrm{B}}{\mathrm{B}} > 0.8 \tag{1.3}$$

$$0.003 < \frac{\mathrm{A} \cap \mathrm{B}}{\mathrm{A} \cup \mathrm{B}} < 0.6 \tag{2}$$

Após isso, foi implementada uma lógica similar para o ajuste de caixas levemente sobrepostas. Se a condição da equação (2) for satisfeita, será feito um cálculo de que dimensão – altura ou largura – da caixa deve ser ajustada. Para isso, sejam LI e LA as larguras da intersecção e da caixa A, respectivamente. Ainda, sejam AI e AA as alturas da intersecção e da caixa A, respectivamente. A visualização do problema pode ser vista na Figura 3. Se a condição (3.1) se satisfizer, a largura da caixa deve ser ajustada. Alternativamente, se a condição (3.2) se satisfizer, a altura da caixa deve ser ajustada.

$$\frac{LI}{LA} > \frac{AI}{AA} \tag{3.1}$$

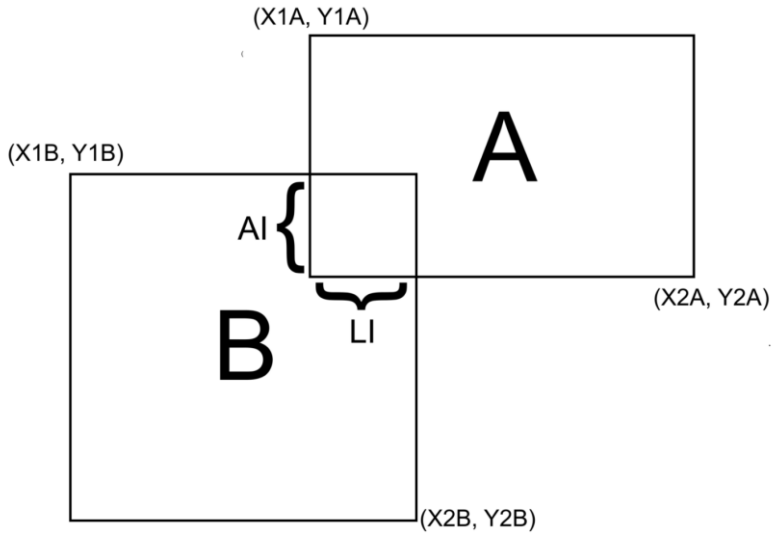
$$\frac{LI}{LA} < \frac{AI}{AA} \tag{3.2}$$

Por fim, caso a condição (3.1) se satisfaça e a largura seja a dimensão que deve ser ajustada, sejam $X1A$, $X1B$, $X2A$ e $X2B$ as coordenadas $X1$ e $X2$ de ambos os objetos. Se a condição (4.1) se satisfizer, i.e. a caixa A está à direita da caixa B, $X1A$ será redefinido como $X2B$. Alternativamente se a condição (4.2) se satisfizer, i.e. a caixa A está à esquerda da caixa B, $X2A$ será redefinido como $X1B$. A mesma lógica de (4.1) e (4.2) vale para ajustes de altura caso a condição (3.2) se satisfaça em detrimento da (3.1). Nesse caso, $(X1A, X2A, X1B, X2B)$ seria trocado por $(Y1A, Y21, Y1B, Y2B)$.

$$X1A > X1B \tag{4.1}$$

$$X2A < X2B \tag{4.2}$$

Figura 3 – Representação das coordenadas e interseção



Fonte: Elaborado pelo autor

Como o algoritmo ainda está sujeito a incongruências, foi aplicado um segundo filtro no qual caixas com coordenadas $(X1, Y1)$ maiores que $(X2, Y2)$ são excluídas. Após essa etapa de tratamento dos dados, se inicia de fato a atribuição cruzada de

objetos. Para isso, os objetos de títulos são reordenados de forma decrescente na coordenada Y para capas. Já para páginas da seção de economia, os objetos de títulos são reordenados de forma decrescente na coordenada Y inclinada em 2 graus para a esquerda. Isto é, títulos são ordenados de baixo para cima e levemente da direita para a esquerda. Isso foi feito, pois na seção de economia havia diversas notícias que se estendiam horizontalmente, fazendo necessária também a existência de uma ordenação horizontal.

Para as páginas da seção de economia exclusivamente, o primeiro passo foi passar por todos os tipos, autores e subtítulos, separadamente, para encontrar o título que está mais próximo de cada uma. Portanto, o objeto primordial que identifica notícias é o título. Tendo feito isso, nos resta somente atribuir textos e notas aos títulos – uma etapa mais sofisticada. Esse passo é idêntico para as capas, mas no lugar das notas, temos outros.

Para cada título é estipulada uma área na qual se devem buscar blocos da classe texto e outros, ou notas. No eixo X, essa área se inicia na coordenada $X1$ da caixa do título e se encerra na coordenada $X2$ da caixa do título, mais um excedente de 0.2% de pixels da altura da página. Isso evita que textos diretamente à direita do título não sejam pegos. No eixo Y, a área se inicia 50 pixels acima da coordenada $Y1$ da caixa do título – a fim de incluir caixas marginalmente acima da caixa do título – e se encerra Z (5) pixels abaixo de $Y2$. Se algum pixel entre o centro e o canto superior esquerdo de alguma caixa de texto ou outros estiver dentro da área definida, o objeto será atribuído ao título.

$$Z = \begin{cases} \sqrt{\acute{a}rea}, & \acute{a}rea_{título} < \acute{a}rea_{capa} * 0.002 \\ 10 * \sqrt{\acute{a}rea}, & \acute{a}rea_{capa} * 0.002 < \acute{a}rea_{título} < \acute{a}rea_{capa} * 0.005 \\ 3 * \sqrt{\acute{a}rea}, & \acute{a}rea_{título} > \acute{a}rea_{capa} * 0.005 \end{cases} \quad (5)$$

A regra delimitada em (5) foi definida com o objetivo de buscar caixas em uma altura grande abaixo do título caso ele tenha área média, buscar caixas em uma altura menor abaixo do título caso ele tenha área grande e buscar caixas em uma altura menor ainda abaixo do título caso ele tenha área pequena. Isso reflete as características observadas na amostra. Títulos pequenos tendem a possuir poucos textos atribuídos a eles. Textos médios tendem a ser os que tem textos mais distantes

atribuídos a eles. E textos grandes são usualmente um balanço entre os últimos dois. Ainda, os textos e outros atribuídos a cada título foram ordenados com base na distância com o título.

Após todos esses filtros é ideal que todos os textos identificados na capa tenham sido atribuídos a algum título. Caso isso não tenha acontecido com algum, será rodado outro filtro para que estes textos sem título sejam inseridos na estrutura e não sejam excluídos. Para cada um desses textos é procurado algum título que esteja à esquerda do $X2$ do texto e acima do $Y2$ do texto. Além disso, o canto inferior direito do bloco do título deve estar a uma distância máxima do canto superior esquerdo do texto. Distância essa definida como 3% da altura da capa. Portanto, esse filtro está buscando títulos com certa proximidade e que estejam acima e à esquerda do texto. Caso mais de um título satisfaça essas condições, o texto será atribuído ao mais próximo deles. Como este título provavelmente já tem outros textos ordenados atribuídos a ele, terá de ser feita uma nova ordenação a fim de inserir o novo texto na colocação correta. Isso é feito, novamente, de acordo com a distância entre os textos e os títulos.

Ainda, foi desenvolvido um passo para lidar com títulos com poucos objetos atribuídos a eles. Para páginas na seção de economia, títulos sem atribuições foram simplesmente excluídos. Já nas capas, foi implementado um filtro somente para quando a manchete possui um ou menos textos. Para detectar a manchete, foi selecionado o título com maior área. Caso este de fato tenha um ou menos textos, será buscado um título em uma área determinada abaixo da manchete. No eixo X a área se inicia na coordenada $X1$ e se encerra na coordenada $X2$ com um excedente de 1% de pixels da altura da capa. No eixo Y não há limites, podendo o título estar em qualquer altura da capa. Se um título for encontrado na área delimitada, ele será transformado em texto e será atribuído, junto aos textos anteriormente atribuídos a ele, à manchete. Esse mecanismo serve para corrigir um problema recorrente, no qual um texto é erroneamente classificado como título. Se o texto mal classificado, por acaso, estiver diretamente abaixo da manchete, ele irá coletar os textos da manchete, deixando-a com um ou menos textos atribuídos a ela. Como essa é a notícia de maior importância na capa, esse acontecimento é inadmissível.

Por fim, as áreas de todas as classes por notícia são calculadas e adicionadas ao conjunto de dados. Ainda, é criado uma série de dados binária para indicar se a notícia é manchete ou não no caso das capas.

2.3 Reconhecimento ótico de caracteres (OCR)

Tendo resolvido toda a questão de tratamento das caixas delimitadoras, chegamos ao efetivo reconhecimento do texto. Para isso se faz o uso do reconhecimento ótico de caracteres, rodado para cada objeto das páginas. Foi carregado um modelo adaptado para a língua portuguesa que foi utilizado da mesma forma tanto nas capas de jornal quanto na seção de economia. Os hifens no final de linha foram apagados, as quebras de linha substituídas por espaços e os erros como espaços duplos substituídos por espaços únicos. Também, diferentes textos atribuídos ao mesmo título foram unidos, mantendo uma separação de quebra de linha para que os vários blocos possam ser visualmente identificados.

3 Modelagem de tópicos

Ao estabelecer todo o processo de análise citado, o produto foi um arquivo em formato CSV como mostra o exemplo da Figura 4 para capas e Figura 5 para páginas da seção de economia.

Figura 4 – Exemplo de trecho da planilha resultante para capas

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Título	Texto	Outros	Manchete	Área título	Área textos	Área outros	Área total	Área capa	Ranking por área do título	Ranking por área total	Data	Links
38318	Presid	O presidente		0	45776.93	99269.82	0.0	145046.75	7774144.0	6	4	20/06/2008	https://acervo.estadao.com.br/publicados/2008/06/20/g/20080620-41884-nac-1-pri-a1-not-aahwwah.jpg
38319	11 mil	Quat		1	367088.45	310709.21	0.0	677797.67	7774144.0	1	1	20/06/2008	https://acervo.estadao.com.br/publicados/2008/06/20/g/20080620-41884-nac-1-pri-a1-not-aahwwah.jpg
38320	Obam	O can US$ 27		0	57732.46	98887.54	27292.03	183912.03	7774144.0	2	2	20/06/2008	https://acervo.estadao.com.br/publicados/2008/06/20/g/20080620-41884-nac-1-pri-a1-not-aahwwah.jpg
38321	Exérci	.++ Para o mir		0	29371.1	59402.08	0.0	88773.18	7774144.0	7	7	20/06/2008	https://acervo.estadao.com.br/publicados/2008/06/20/g/20080620-41884-nac-1-pri-a1-not-aahwwah.jpg
38322	ee sa	\o Seleção foi		0	14829.03	24065.57	0.0	38894.6	7774144.0	10	9	20/06/2008	https://acervo.estadao.com.br/publicados/2008/06/20/g/20080620-41884-nac-1-pri-a1-not-aahwwah.jpg
38323	Tráfic	Traficantesda		0	46623.94	85505.8	0.0	132129.74	7774144.0	5	6	20/06/2008	https://acervo.estadao.com.br/publicados/2008/06/20/g/20080620-41884-nac-1-pri-a1-not-aahwwah.jpg
38324	Safra t	O governo vai		0	53279.56	89010.48	0.0	142290.04	7774144.0	4	5	20/06/2008	https://acervo.estadao.com.br/publicados/2008/06/20/g/20080620-41884-nac-1-pri-a1-not-aahwwah.jpg
38325	Grupo	O grupo norte		0	54254.14	96547.56	0.0	150801.7	7774144.0	3	3	20/06/2008	https://acervo.estadao.com.br/publicados/2008/06/20/g/20080620-41884-nac-1-pri-a1-not-aahwwah.jpg
38326	TPSA	L o Longa de Da		0	15690.29	17965.79	0.0	33656.08	7774144.0	9	10	20/06/2008	https://acervo.estadao.com.br/publicados/2008/06/20/g/20080620-41884-nac-1-pri-a1-not-aahwwah.jpg
38327	Home	o Versão mod		0	14319.15	18264.57	0.0	32583.72	7774144.0	12	11	20/06/2008	https://acervo.estadao.com.br/publicados/2008/06/20/g/20080620-41884-nac-1-pri-a1-not-aahwwah.jpg
38328	Brasil	o Meta é destı		0	14782.43	17495.6	0.0	32278.04	7774144.0	11	12	20/06/2008	https://acervo.estadao.com.br/publicados/2008/06/20/g/20080620-41884-nac-1-pri-a1-not-aahwwah.jpg
38329	Golpe	Ruy Pereira C		0	6999.35	20009.73	0.0	27009.07	7774144.0	13	13	20/06/2008	https://acervo.estadao.com.br/publicados/2008/06/20/g/20080620-41884-nac-1-pri-a1-not-aahwwah.jpg
38330	Um pc	Alibe		0	18966.89	43129.36	0.0	62096.25	7774144.0	8	8	20/06/2008	https://acervo.estadao.com.br/publicados/2008/06/20/g/20080620-41884-nac-1-pri-a1-not-aahwwah.jpg

Fonte: Elaborado pelo autor

Figura 5 – Exemplo de trecho da planilha resultante para a seção de economia

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Título	Texto	Destaques	Autores	Tipo	Subtítulo	Área título	Área textos	Área destaques	Área subtítulo	Área total	Área capa	Ranking por área do título	Ranking por área total	ID	Link
113559	IPCA fecha	A inflação r	IBGE ainda	Jacqueline	CONJUNTU	mentos cor	507820,6	979143,38	43018,80873	102487,6007	1632470	7774144	1	1	11-07-2008-20-b1	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-20-eco-b1-not-kqgsqxa.jpg
113560	"Preços de	* O presidente	Luiz Iná >	Mais informações na:	Segundo pr		145289,5	637967,14	0	88371,07415	871627,7	7774144	2	2	11-07-2008-20-b1	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-20-eco-b1-not-kqgsqxa.jpg
113561	Contínua fa	expectativa de que a in	CELSO MING	releo mino	Wwosunr		52800,84	731263,5	0	0	784064,3	7774144	3	1	11-07-2008-21-b2	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-21-eco-b2-not-pagsqxa.jpg
113562		Mais consumo -	O relatório	mensal da Agência	Internacior		5362,23	161784,19	0	0	167146,4	7774144	8	5	11-07-2008-21-b2	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-21-eco-b2-not-pagsqxa.jpg
113563	Por vias tor	xão. Setiver fundament	Rogério L. F	Opinião			36694,32	550185,17	0	0	586879,5	7774144	4	2	11-07-2008-21-b2	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-21-eco-b2-not-pagsqxa.jpg
113564	Brasil lidera	0 O desenvolvimento	de projetos e	— RELATÓRIO	DAAIE		16656,82	97424,312	0	0	114081,1	7774144	7	7	11-07-2008-21-b2	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-21-eco-b2-not-pagsqxa.jpg
113565	eoo	"Esper	Abdalla Salem	el-Badri,	secretário-geral	da Organização d	68705,86	20404,706	0	0	89110,56	7774144	2	8	11-07-2008-21-b2	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-21-eco-b2-not-pagsqxa.jpg
113566	Previsões d	ma dessa previsão.	Essainflação	deverá responder,	neste		109234	472542,58	0	0	581776,5	7774144	1	3	11-07-2008-21-b2	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-21-eco-b2-not-pagsqxa.jpg
113567	EF prenhe	7 0 A Polícia Federal (PF)	prendeu ur	OPERAÇÃO BICHO	MIR		17075,62	135783,17	0	0	152858,8	7774144	5	6	11-07-2008-21-b2	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-21-eco-b2-not-pagsqxa.jpg
113568	Ruralistas	0 O setor agropecuário	da Argentina	pretende voltar a exib			16827,56	287428,78	0	0	304256,3	7774144	6	4	11-07-2008-21-b2	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-21-eco-b2-not-pagsqxa.jpg
113569	Inflação vo	O presidente do Banco	Célia Frouf	CONJUNTU	Para Meirel		224147	796445,64	0	121383,6294	1141976	7774144	1	1	11-07-2008-22-b3	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-22-eco-b3-not-wggsqxa.jpg
113570	Em 15 dias	Após meses de ação di	Fernando	Nakagawa Fi	C intensific		84923,84	504751,51	0	52358,68948	642034	7774144	2	2	11-07-2008-22-b3	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-22-eco-b3-not-wggsqxa.jpg
113571	Lficcindo t	Namorado	querendo d	Mariti Ribei	PESOQUIS	Há gastado	335974,5	467504,13	0	59615,88231	863094,5	7774144	1	2	11-07-2008-23-b4	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-23-eco-b4-not-xegsqxa.jpg
113572	Eban an cn	Pelo segundo mês con	Márcia De	FINANCAS			68876,88	847602,36	0	0	916479,2	7774144	3	1	11-07-2008-23-b4	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-23-eco-b4-not-xegsqxa.jpg
113573	Banco dá	Os grandes Acúmulo d	Rente Pere	COMPETIÇ	Maiores ins		83583,11	499567,35	42534,25669	62928,44504	688613,2	7774144	2	3	11-07-2008-23-b4	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-23-eco-b4-not-xegsqxa.jpg
113574	L dd d pc	M RA Além de Dotações d	Ribamar Oi	FINANCAS	Ações e del		317470	907083,76	42663,7516	90139,71347	1357297	7774144	1	1	11-07-2008-24-b5	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-24-eco-b5-not-qhgsqxa.jpg
113575	BNDES terr	O Banco Nacional de D	Alberto Komatsu	RIO	Banco dob		74344,83	612802,63	0	47448,29279	734595,7	7774144	2	2	11-07-2008-24-b5	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-24-eco-b5-not-qhgsqxa.jpg
113576	Governo qu	O governo quer criar	parcerias com	ainiciativa p	Participaçã		220476,1	415854,52	0	107173,6267	743504,3	7774144	1	1	11-07-2008-25-b6	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-25-eco-b6-not-asgsqxa.jpg
113577	Lobbão prc	O ministro de Minas e E	RIO				62865,22	281886,85	0	0	344752,1	7774144	3	3	11-07-2008-25-b6	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-25-eco-b6-not-asgsqxa.jpg
113578	Constituiçã	Deacordoc	Constituiçã	Daniele Ca	ENERGIA E COMBUSTÍ		38487,82	188066,24	43566,81634	0	270120,9	7774144	4	4	11-07-2008-25-b6	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-25-eco-b6-not-asgsqxa.jpg
113579	Nova propr	O ministro do Meio Am	Leonardo Goy	BRASÍLIA			73638,08	349892,79	0	0	423530,9	7774144	2	2	11-07-2008-25-b6	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-25-eco-b6-not-asgsqxa.jpg
113580	OMC já terr	Os mediadores da Orgı	GENEBRA	COMÉRCIC	Textos revis		271036,3	469968,21	0	63434,2958	804438,8	7774144	1	1	11-07-2008-26-b7	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-26-eco-b7-not-gkgsqxa.jpg
113581	Brasil inves	O Ministério do Desenv	Renata Veríssimo	BRASÍLIA			61584,08	358109,55	0	0	419693,6	7774144	2	2	11-07-2008-26-b7	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-26-eco-b7-not-gkgsqxa.jpg
113582	Sarkozy:	"A s.s O presidente da	França, Nicolas	Sarkozy,	também atua		34549,92	345865,43	0	0	380415,3	7774144	3	3	11-07-2008-26-b7	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-26-eco-b7-not-gkgsqxa.jpg
113583	Juiz libera	O juiz Luiz Roberto	Ayoub,	coorder	NEGÓCIOS	NO AR	309555,6	382765,86	0	0	692321,5	7774144	1	2	11-07-2008-27-b8	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-27-eco-b8-not-epgsqxa.jpg
113584	Odebrecht	e Ticdoft een f	Doços p	Cláudia Tre	INTERNACIONALIZAÇÃ		60249,94	379018,93	0	0	439268,9	7774144	3	3	11-07-2008-27-b8	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-27-eco-b8-not-epgsqxa.jpg
113585	Aeroviários	s.s Os aeroviários da Vi	Alberto Komatsu	RIO	Valor se lim		33603,54	177237,54	0	60158,16625	270999,2	7774144	5	4	11-07-2008-27-b8	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-27-eco-b8-not-epgsqxa.jpg
113586	Lula diz que pode have	Lula diz que pode haver	cooperação com o Vie				42588,93	0	41273,83769	0	83862,77	7774144	4	6	11-07-2008-27-b8	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-27-eco-b8-not-epgsqxa.jpg
113587	Petroleiros	* . . * * O Sindicato dos	Nicola Parr	TRABALHO	Categoria a		207657	838288,89	0	52548,20285	1098494	7774144	2	1	11-07-2008-27-b8	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-27-eco-b8-not-epgsqxa.jpg
113588	Temporada se.	Metalúrgicos ligados à CUT	fizeram ontem	manifestaçã			20340,22	115614,57	0	0	135954,8	7774144	6	5	11-07-2008-27-b8	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-27-eco-b8-not-epgsqxa.jpg
113589	Bernanke v	O presidente do Federal	Reserve (F	INTERNACIE	m depoim		211248,2	280579,6	0	131083,4594	622911,2	7774144	1	1	11-07-2008-28-b9	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-28-eco-b9-not-hwgsqxa.jpg
113590	Inflação alt	O presidente do Banco	MUNIQUE				73429,13	414025,56	0	0	487454,7	7774144	2	2	11-07-2008-28-b9	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-28-eco-b9-not-hwgsqxa.jpg
113591	Segundo el	Na avaliaçã	Segundo el	WASHINGTON			40343,21	285233,16	39345,22076	0	364921,6	7774144	3	3	11-07-2008-28-b9	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-28-eco-b9-not-hwgsqxa.jpg
113592	Petróleo di	As bolsas seguiram	voláteis, ontem	CENÁRIO F	Índices de a		231646	212068,81	0	112470,6908	556185,5	7774144	1	1	11-07-2008-29-b10	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-29-eco-b10-not-sxgsqxa.jpg
113593	Preço do oi	Acombinação	petróleo em alta	dólarem	queda no	exterio	19121,35	84110,316	0	0	103231,7	7774144	3	2	11-07-2008-29-b10	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-29-eco-b10-not-sxgsqxa.jpg
113594	BrT sobe m	BrT PN subiu 8,33%,	repercutindo	declarações do	ministr		19023,04	51926,881	0	0	70949,92	7774144	4	3	11-07-2008-29-b10	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-29-eco-b10-not-sxgsqxa.jpg
113595	FRASE	Ferr	"IPCA	tirou o gás do	receio de	piora das	24876,65	32423,221	0	0	57299,87	7774144	2	4	11-07-2008-29-b10	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-29-eco-b10-not-sxgsqxa.jpg
113596	Soja antecı	Os preços da soja	registraram	forte alta,	pelo segundo dia		19007,45	119871,4	0	0	138878,8	7774144	1	1	11-07-2008-30-b11	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-30-eco-b11-not-pqesqxa.jpg
113597	Governo Ki	Emcrise,	companhia vira	alvo de política	de nacionalizaçã		17713,82	21620,089	0	0	39333,91	7774144	6	9	11-07-2008-31-b12	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-31-eco-b12-not-waesqxa.jpg
113598	Depois do i	Empresa começa a vender	hoje o novo	iPhone,	que adota		17478,53	20687,435	0	0	38165,96	7774144	8	10	11-07-2008-31-b12	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-31-eco-b12-not-waesqxa.jpg
113599	nnnnn Agre	Nocomeço da manhã c	Patrícia Cai	OLSA	Bolsa paral		402381,2	1015707,8	0	95653,7132	1513743	7774144	1	1	11-07-2008-31-b12	https://acervo.estadao.com.br/publicados/2008/07/11/g/20080711-41905-nac-31-eco-b12-not-waesqxa.jpg

Fonte: Elaborado pelo autor

A partir desse texto, serão feitas 3 etapas para modelagem de tópicos de cada notícia: processamento do texto, vetorização e alocação latente de Dirichlet. Essa cadeia de métodos para extração do conteúdo de textos segue a sugestão de Kashyap (2024) – uma abordagem amplamente usada. Através dessas etapas, serão identificados automaticamente os temas predominantes em cada corpo textual.

3.1 Processamento do texto

Após importar o conjunto de dados das figuras 4 e 5, foi necessário agregar todos os textos em uma coluna. Isso foi feito para ambas as bases de dados. Entretanto, na segunda, a coluna de autores não foi incluída na nova coluna com o conteúdo textual bruto. Isso foi feito pois essa informação não ajudaria na classificação de tópicos.

A partir disso, realizou-se o pré-processamento textual, fundamental para transformar o texto em uma forma que possa ser analisada por algoritmos de aprendizado de máquina. Inicialmente foi realizada uma tokenização que consistiu na divisão dos textos em palavras individuais. Após isso foi feita a remoção de *stop words*. Essas são palavras comuns da língua, como “o”, “é”, “de”, que não carregam valor semântico relevante. Ainda, foi feita uma filtragem de termos na qual foram mantidas apenas as palavras que aparecem em pelo menos 5 notícias e em no máximo 75% das notícias. Isso tem o objetivo de eliminar palavras muito raras ou excessivamente frequentes como as próprias *stop words*.

Por fim, as notícias pré-processados foram vetorizadas usando a técnica TF-IDF (Term Frequency-Inverse Document Frequency). Essa abordagem transforma os textos em uma matriz, onde cada linha representa uma notícia e cada coluna representa um termo. Os valores correspondem à importância relativa de cada termo em cada notícia. Para as notícias presentes na seção de economia, essa matriz possui 280915 linhas – notícias – e 28556 colunas, ou seja, 28556 termos. Já para as notícias presentes em capa, essa matriz possui 129737 linhas e 5004 colunas, ou seja, somente 5004 termos. Aqui fica claro que o processamento do texto reduz muito a quantidade de termos que o computador analisará a fim de discriminar os dados.

3.2 Alocação latente de Dirichlet (LDA)

A partir das matrizes de termos foi possível aplicar o LDA para a modelagem de tópicos. Para isso, se faz necessário definir um número de tópicos – uma decisão mais argumentativa. Para as notícias em capas, foi definido um número de 80 tópicos já que nela pode haver qualquer tipo de notícia. Já para as páginas da seção de economia com temas menos abrangentes, foram definidos somente 40 tópicos. Vale

lembrar que um número grande de tópicos piora a performance do modelo. Portanto é sempre preferível número menores.

A aplicação do LDA foi feita utilizando a implementação disponível na biblioteca sklearn. Esse modelo probabilístico assume que cada notícia é composta por uma combinação de tópicos e que cada tópico é composto por um conjunto de palavras mais prováveis. Portanto o resultado foi outra matriz em que cada linha representa uma notícia, cada coluna um tópico e cada valor indica a proporção de presença do tópico na respectiva notícia.

Podemos também interpretar semanticamente cada tópico identificado, extraindo as palavras mais representativas. Essa extração fornece os termos com maior peso em cada tópico, permitindo uma interpretação. As tabelas com as 10 palavras mais frequentes para os 80 tópicos de capas e 40 tópicos de páginas da seção de economia estão disponíveis no apêndice C. Os tópicos que mais se assemelharam a temas fiscais para capas foram os números 4, 8, 25 e 68 como podemos visualizar no quadro 3. Já para páginas da seção de economia, os tópicos que mais se assemelharam a temas fiscais foram os números 10, 12, 34 e 37 como visto no quadro 4.

Quadro 3 – Palavras representativas de tópicos fiscais em capa

	Topic 04	**Topic 08**	**Topic 25**	**Topic 68**
1	governo (251.349)	b12 (109.266)	fmi (163.309)	fiscal (139.149)
2	bilhões (183.960)	renda (95.096)	bancos (153.762)	contas (122.367)
3	reforma (154.194)	economia (82.222)	economia (99.417)	meta (107.340)
4	economia (153.281)	pág (76.875)	sal (84.931)	governo (93.886)
5	servidores (149.597)	crédito (70.515)	brasil (83.928)	superávit (93.796)
6	pág (148.753)	tributária (62.879)	banco (83.674)	bilhões (90.048)
7	salário (143.793)	imposto (61.485)	bilhões (79.729)	gastos (85.926)
8	previdência (142.936)	pobreza (57.045)	pág (74.569)	públicas (79.157)
9	orçamento (132.543)	cartão (56.706)	argentina (73.585)	pág (76.223)
10	congresso (128.465)	receita (55.537)	pré (72.450)	déficit (75.370)

Fonte: Elaborado pelo autor

Quadro 4 – Palavras representativas de tópicos fiscais na seção de economia

	Topic 10	Topic 12	Topic 34	Topic 37
1	justiça (527.586)	governo (1061.059)	crédito (954.647)	declaração (352.424)
2	lei (480.590)	bilhões (906.730)	bancos (730.507)	imposto (297.303)
3	federal (463.009)	pib (628.746)	banco (659.799)	renda (188.799)
4	governo (430.480)	fiscal (583.933)	fundos (507.368)	receita (186.511)
5	câmara (402.766)	ano (569.299)	bndes (476.316)	contribuinte (173.585)
6	proposta (395.687)	arrecadação (549.274)	caixa (440.809)	ir (146.228)
7	presidente (361.161)	despesas (546.240)	juros (422.007)	rendimentos (138.326)
8	tribunal (359.218)	gastos (522.813)	recursos (408.178)	valor (136.999)
9	projeto (356.874)	superávit (456.831)	financiamento (348.098)	contribuintes (114.718)
10	decisão (347.788)	contas (415.802)	poupança (336.919)	restituição (93.584)

Fonte: Elaborado pelo autor

Apesar do tópico 10 no quadro 4 poder ser relacionado ao âmbito fiscal, ele não é tão objetivo quanto os outros. Portanto não será incluído como tema fiscal. Além disso, os tópicos mais claramente fiscais são os 68 nas capas e 12 na seção de economia. Por conta disso, serão feitos dois sinais fiscais para cada caso. Um sinal puro que é simplesmente o tópico 68 ou 12, e um sinal composto combinando todos. No sinal composto em capas, temos 50% de peso no tópico 68, 25% no 4, 15% no 8 e 10% no 25. Já no de páginas de economia, temos 65% de peso no tópico 12, 20% no 37 e 15% no 34. Esses valores foram definidos com raciocínio lógico argumentativo.

Por fim, nos resta agregar as probabilidades em diferentes notícias dentro do mesmo dia. Para isso vamos fazer uma média ponderada baseada na área que a notícia ocupa dentro da página. A equação 6 demonstra o cálculo, onde n é o número de notícias dentro do mesmo dia.

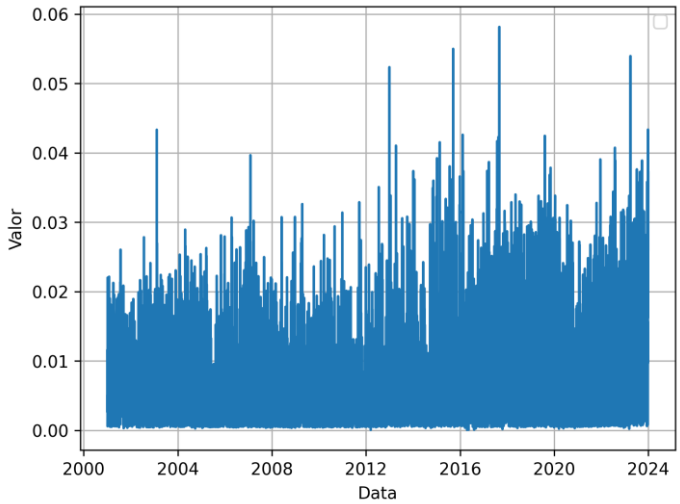
$$probabilidade_{dia} = \sum_{i=0}^{n} \frac{área_{notícia\ i}}{área_{página}} * probabilidade_{notícia\ i} \quad (6)$$

Também podemos agregar para dados mensais somando as probabilidades diárias. Por fim, seria valioso também calcular essas séries temporais na diferença a fim de observar picos de variação em notícias fiscais; ou seja, choques fiscais. Os gráficos finais e interpretações seguem na próxima seção.

4 Resultados

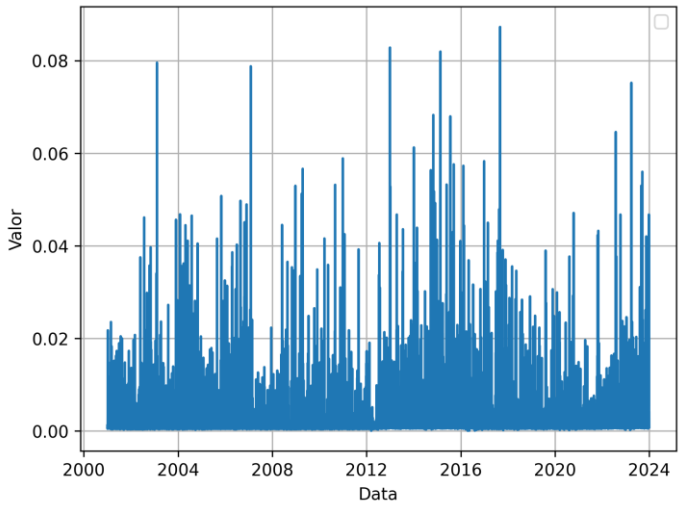
Tendo determinado o método de agregação, podemos agora simplesmente plotar as séries temporais de notícias fiscais. Inicialmente temos nos gráficos 7 e 8 os sinais fiscais compostos e puros para capas. Como se trata de uma frequência diária, é difícil visualizar padrões.

Gráfico 7 – Notícias fiscais diárias em capa COMPOSITE



Fonte: Elaborado pelo autor

Gráfico 8 – Notícias fiscais diárias em capa PURE

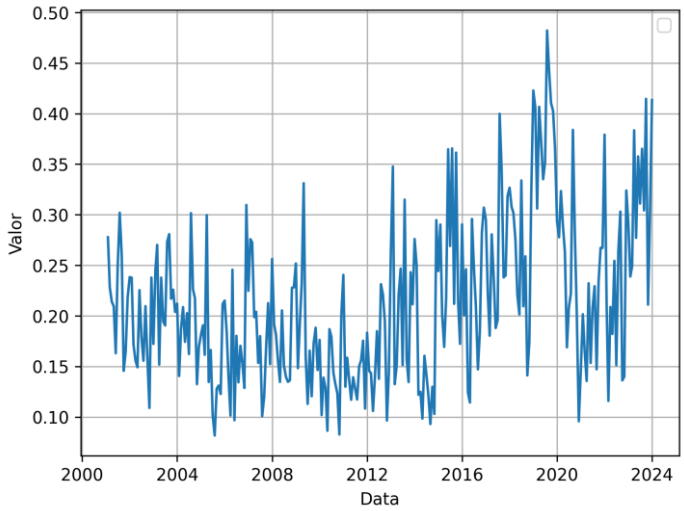


Fonte: Elaborado pelo autor

Com a agregação mensal dos gráficos 9 e 10, fica mais fácil a identificação de padrões. No sinal composto, vemos claramente um pico de notícias em 2015, 2019 e

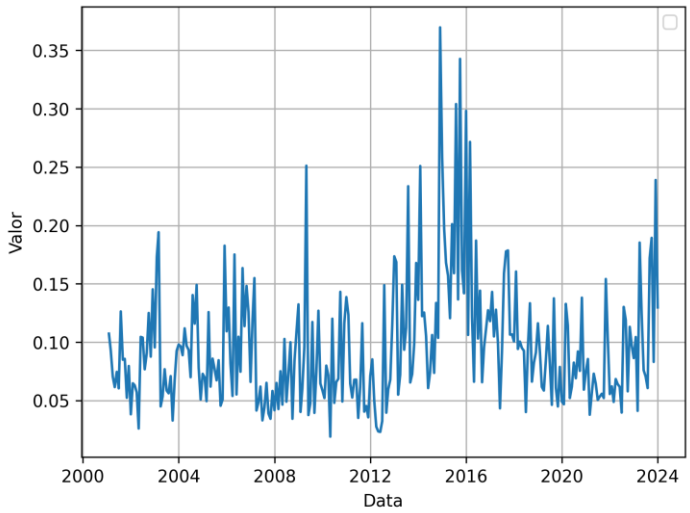
2023. No sinal puro, esse pico está também evidente, mas em torno dos anos de 2008, 2015 e 2023. A análise combinada desses dois sinais demonstra uma boa semelhança com a realidade. Os anos de 2008 e 2009 foram marcados por estímulos pós crise global. Já 2015 e 2016 foi o auge, marcado por situação fiscal terrível, ajuste fiscal e impeachment. 2019 também foi marcante com a reforma da previdência. Por fim, em 2023 muito se falou do arcabouço fiscal e dos riscos quanto a nova gestão populista no Brasil.

Gráfico 9 – Notícias fiscais mensais em capa COMPOSITE



Fonte: Elaborado pelo autor

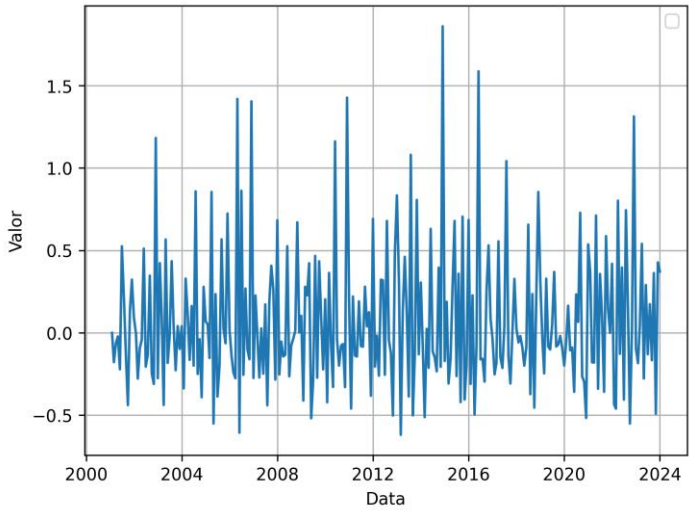
Gráfico 10 – Notícias fiscais mensais em capa PURE



Fonte: Elaborado pelo autor

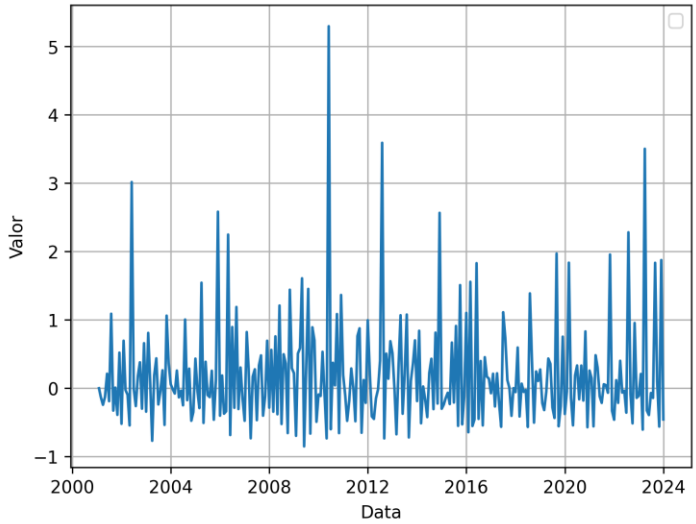
As variações mensais desse sinal também podem trazer interpretações valiosas. No sinal composto temos picos claros no fim de 2002, começo e fim de 2006, meio e fim de 2010, meio de 2013, fim de 2014, meio de 2016, fim de 2017 e fim de 2022. No sinal puro, os picos claros são no meio de 2002, final de 2005, começo de 2006, meio de 2010, meio de 2012, fim de 2014, meio de 2022 e começo de 2023.

Gráfico 11 – Mudança de notícias fiscais mensais em capa COMPOSITE



Fonte: Elaborado pelo autor

Gráfico 12 – Mudança de notícias fiscais mensais em capa PURE



Fonte: Elaborado pelo autor

Cruzando essas duas informações, temos alguns picos moderados. No meio e fim de 2002: Crise de confiança pré-eleição de Lula; contração do crédito internacional

e das economias latinas; necessidade de ajuste fiscal. Final de 2005: Dilma Rousseff entra como ministra da Casa Civil; mensalão contribui para crise política. Começo e fim de 2006: Palocci sai e Guido Mantega assume o cargo de ministro da Fazenda; Programa de Aceleração do Crescimento é lançado com vários subsídios e benefícios. Meio de 2013: Aumento de subsídios e estímulos públicos frente às manifestações de junho; contexto da Nova Matriz Econômica de Dilma. Fim de 2017: Debate sobre reforma da previdência. Meio e fim de 2022: PEC Kamikaze fura Teto de Gastos e amplia benefícios.

Seguindo esses choques, temos também os picos fortes. No meio de 2012: Políticas de desoneração; contexto da Nova Matriz Econômica; sinais de desaceleração da economia. Começo de 2023: Proposta do arcabouço fiscal como substituto do Teto de Gastos; debate sobre sustentabilidade fiscal.

Por fim, temos 3 períodos que representaram as maiores variabilidades das notícias. No meio e fim de 2010: Gastos crescentes com funcionalismo; contabilidade criativa nos dados fiscais; expansionismo fiscal desde a crise de 2008; Dilma é eleita com promessas de impulsão dos investimentos. Fim de 2014: Déficit crescente; reeleição de Dilma; promessa de ajuste fiscal e troca de ministro da Fazenda para Joaquim Levy. Meio de 2016: Temer assume após impeachment; discussões sobre ajuste fiscal; PEC do Teto de Gastos.

5 Conclusão

Diante do que foi exposto, pode-se concluir que o objetivo primordial deste trabalho foi concluído. Apesar disso, houve muitos empecilhos no caminho que impossibilitaram o desenvolvimento de um modelo e bases de dados perfeitos.

Inicialmente, o algoritmo de agrupamento de notícias se mostrou problemático em alguns casos. A alta complexidade dos layouts torna quase impossível a existência de um programa que os agrupe de forma correta sempre. Com certa frequência caixas de texto estavam demasiadamente longe de seus títulos, fazendo com que elas fossem atribuídas a outro título, ou fossem apagadas. Felizmente, para grande parte dos estudos econômicos isso não é um empecilho, pois o interesse está na identificação de um tema para cada dia. A troca de textos entre notícias da mesma página não afeta isto. Ainda, a forma como o algoritmo foi construído impossibilita que uma caixa identificada como título seja excluída, preservando o máximo de informações possível, mesmo que não estejam devidamente organizadas.

Sendo assim, o problema central dessa metodologia para o propósito acadêmico são os erros que o próprio modelo de aprendizado profundo comete. Eles podem ser de localização ou classificação das caixas delimitadoras, sendo ambos altamente problemáticos pois resultam na exclusão de textos relevantes. Infelizmente, o ganho de treino com mais iterações é marginal como se comprova através da estabilização da precisão média e perda total nos Gráficos 3, 4, 5 e 6. A alternativa restante é o aumento dos exemplos para treinamento. Como isso representa um ganho pequeno de performance e as ocorrências de erro ainda não são conhecidas, este tópico não foi considerado.

Adicionalmente, houve um erro fatal no modelo para a seção de economia. Ao rodar o programa, o computador ficou em torno de 10 dias lendo e digitalizando todo o conteúdo. Por conta desse período extensivo, em diversos momentos, a máquina encontrou problemas ao acessar o link das páginas – e.g. problemas do servidor e de conexão com internet. O código, por sua vez, ignorou as páginas com esses problemas e continuou rodando. Isso fez com que somente em torno de 70% das páginas de economia tenham sido de fato lidas. As outras foram ignoradas e não estão presentes na base de dados final. Por conta disso, a única série temporal analisada na seção de resultados foi a advinda de notícias em capa. Apesar disso, o caminho está pavimentado e o trabalho feito. Consertar esse empecilho é um mero problema

operacional e poderia ser facilmente corrigido com técnicos de informática – alguém familiarizado com os métodos envolvendo um código tão demorado.

Por fim, mesmo restrita às notícias de capa, a série temporal de notícias fiscais mostrou-se extremamente satisfatória, refletindo acontecimentos reais e possuindo interpretação direta. Seu uso em conjunto com séries de PIB ou risco país revela grande potencial para a compreensão da propagação de choques fiscais no Brasil, de modo que se espera que seja utilizada com tal propósito em *Fiscal Shocks or Fiscal News? Evidence from Daily Fiscal Expectations* (Bandeira; Bonomo; Carvalho; Mendes, 2025). Também se espera que sirva de inspiração para a criação de outras séries temporais de notícias comerciais ou políticas de interesse do Centro de Finanças e Macroeconomia do Insper. Em suma, este relatório final apresenta a conclusão de um método de extração, estruturação e interpretação de notícias veiculadas no jornal impresso *O Estado de S. Paulo*, cujos resultados demonstram o sucesso da linha de análise desenvolvida, ainda que desafios como a identificação correta de caixas delimitadoras revelem as limitações do procedimento. Nesse sentido, permanece em aberto a correção da base de dados das páginas de economia, de modo que a série temporal de notícias fiscais se torne ainda mais fidedigna. Em última análise, este trabalho contribui para a estruturação de dados favoráveis ao CeFin e oferece uma base sólida para a criação de novas bases de dados provenientes de fontes desestruturadas.

Referências

BANDEIRA, Miguel; BONOMO, Marco; CARVALHO, Carlos; MENDES, Marcos. Fiscal shocks or fiscal news? Evidence from daily fiscal expectations. São Paulo: Insper – Centro de Finanças e Macroeconomia, 2025. (Working Paper).

KASHYAP, Piyush. *Topic modeling with Latent Dirichlet Allocation (LDA)*. Medium, [s. l.], [2024]. Disponível em: https://medium.com/@piyushkashyap045/topic-modeling-with-latent-dirichlet-allocation-lda-d2ab3fcfba68. Acesso em: 31 jul. 2025.

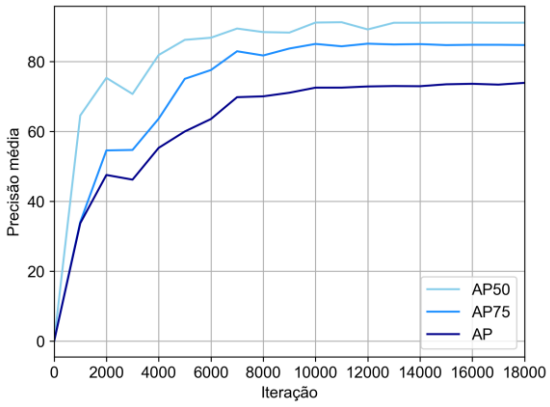
MALYUK, Michael; LIUBIMOV, Nikolai. Label Studio. California, 2025. Disponível em: https://labelstud.io/. Acesso em: 18 out. 2024.

SHEN, Zejiang; DELL, Melissa; LEE, Benjamin. Layout Parser. Washington, 2021. Disponível em: https://layout-parser.github.io/. Acesso em: 18 out. 2024.

SHEN, Zejiang; ZHANG, Kaixuan; DELL, Melissa. A Large Dataset of Historical Japanese Documents with Complex Layouts. Harvard University, abr. 2020. Disponível em: https://arxiv.org/pdf/2004.08686. Acesso em: 18 out. 2024.

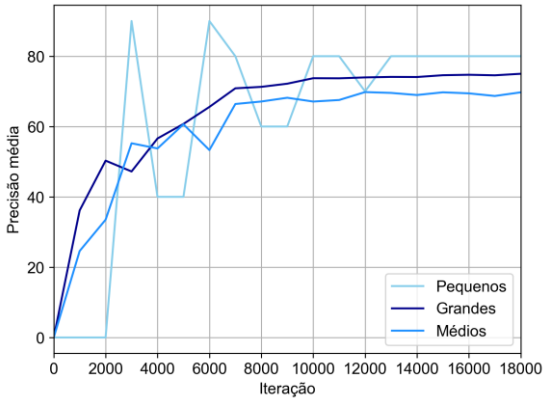
Apêndice A – Resultados adicionais: Modelo 1

Gráfico 13 – Precisão média do modelo 1



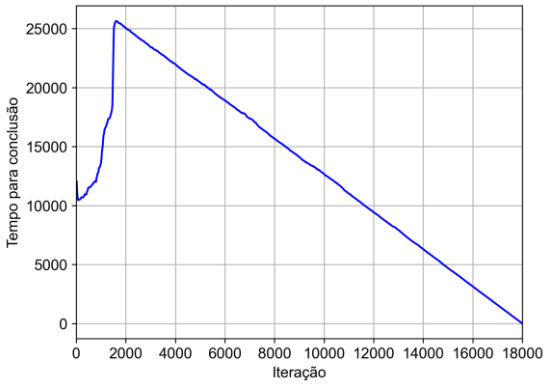
Fonte: Elaborado pelo autor

Gráfico 14 – Precisão média para diferentes tamanhos de objeto do modelo 1



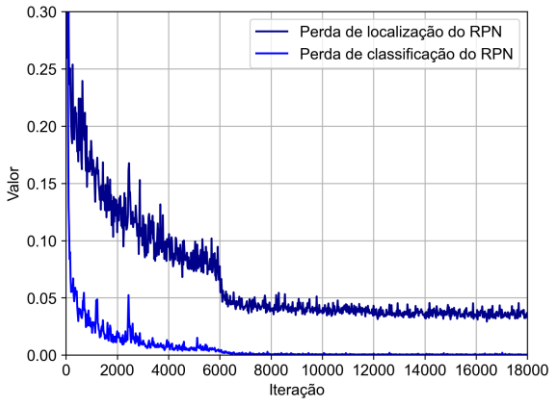
Fonte: Elaborado pelo autor

Gráfico 15 – Tempo estimado para conclusão do treinamento do modelo 1



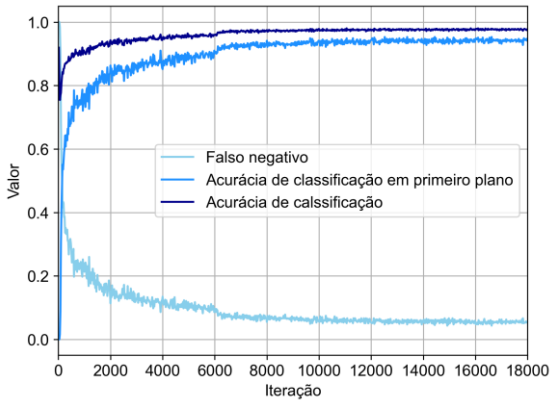
Fonte: Elaborado pelo autor

Gráfico 16 – Outras modalidades de perda do modelo 1



Fonte: Elaborado pelo autor

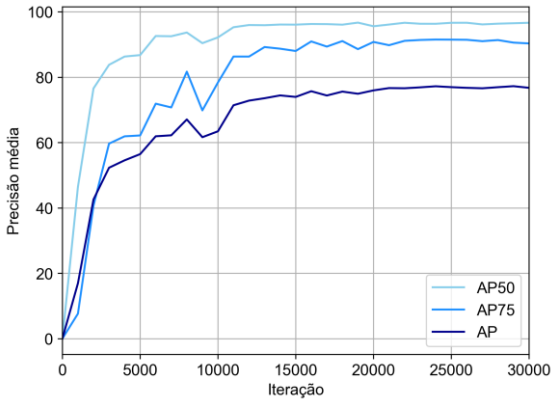
Gráfico 17 – Acurácia e falso negativo do modelo 1



Fonte: Elaborado pelo autor

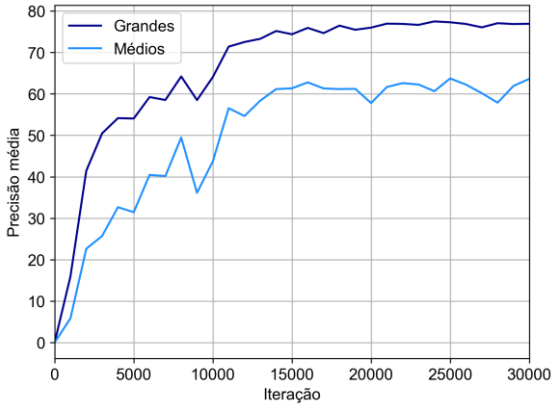
Apêndice B – Resultados adicionais: Modelo 2

Gráfico 18 – Precisão média do modelo 2



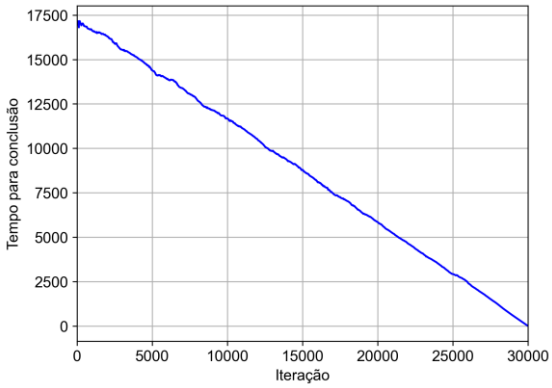
Fonte: Elaborado pelo autor

Gráfico 19 – Precisão média para diferentes tamanhos de objeto do modelo 2



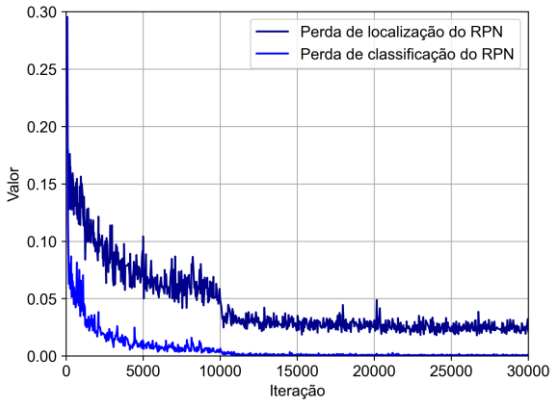
Fonte: Elaborado pelo autor

Gráfico 20 – Tempo estimado para conclusão do treinamento do modelo 2



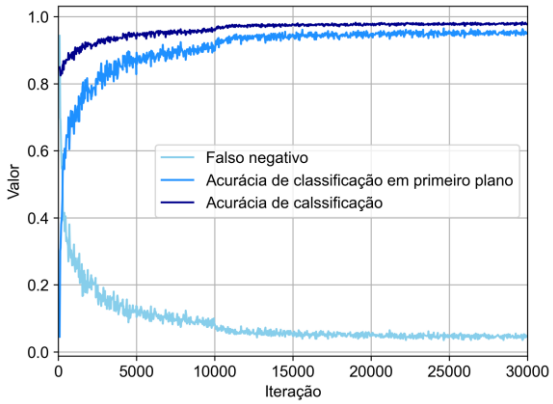
Fonte: Elaborado pelo autor

Gráfico 21 – Outras modalidades de perda do modelo 2



Fonte: Elaborado pelo autor

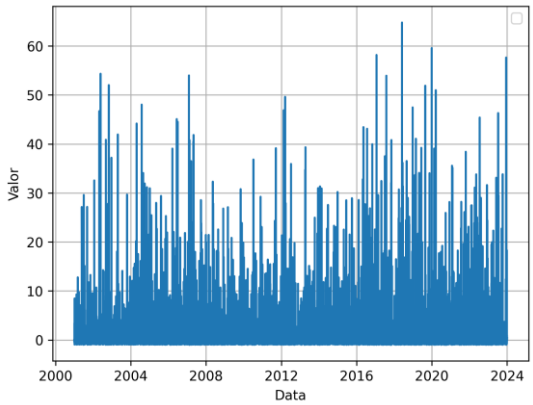
Gráfico 22 – Acurácia e falso negativo do modelo 2



Fonte: Elaborado pelo autor

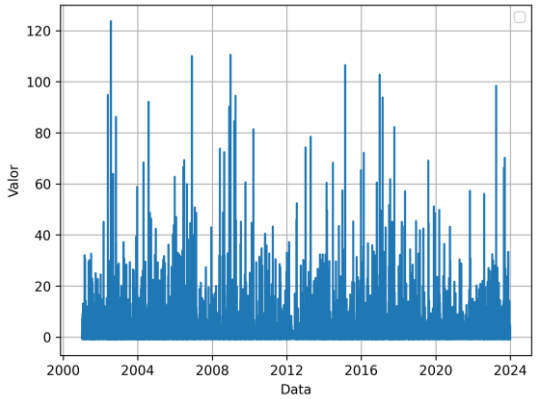
Apêndice C – Resultados adicionais: Sinais fiscais

Gráfico 23 – Mudança de notícias fiscais diárias em capa COMPOSITE



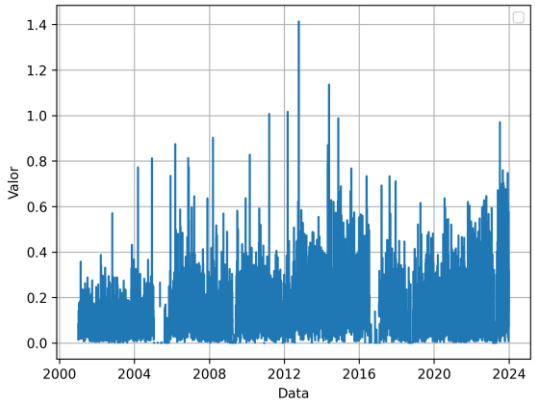
Fonte: Elaborado pelo autor

Gráfico 24 – Mudança de notícias fiscais diárias em capa PURE



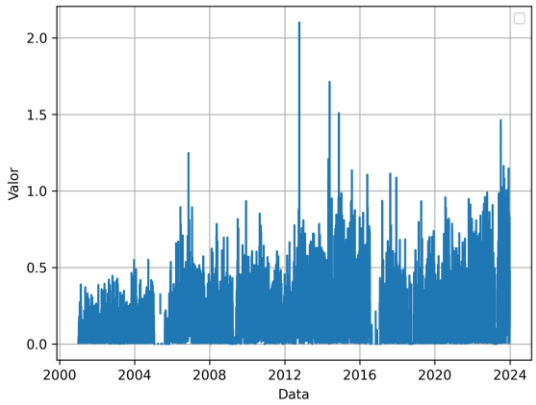
Fonte: Elaborado pelo autor

Gráfico 25 – Notícias fiscais diárias em páginas de economia COMPOSITE



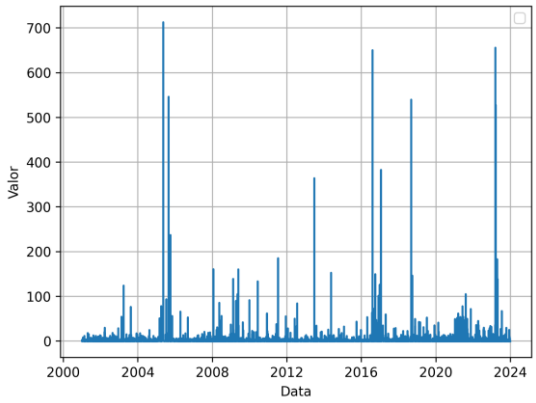
Fonte: Elaborado pelo autor

Gráfico 26 – Notícias fiscais diárias em páginas de economia PURE



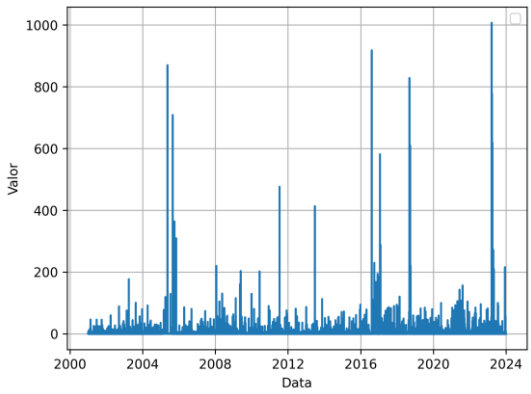
Fonte: Elaborado pelo autor

Gráfico 27 – Mudança de notícias fiscais diárias em páginas de economia COMPOSITE



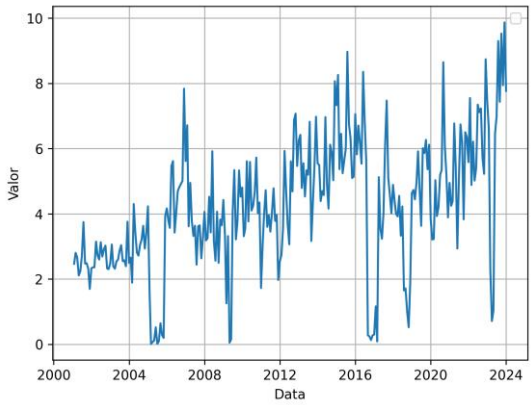
Fonte: Elaborado pelo autor

Gráfico 28 – Mudança de notícias fiscais diárias em páginas de economia PURE



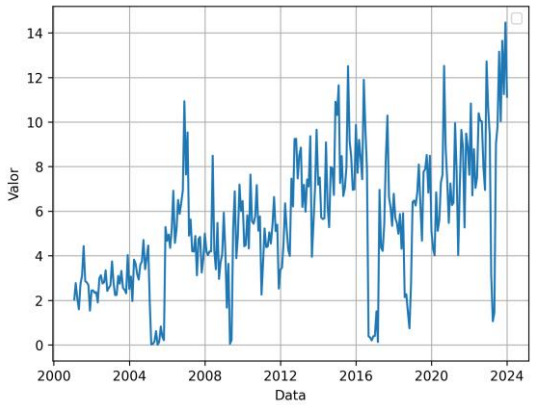
Fonte: Elaborado pelo autor

Gráfico 29 – Notícias fiscais mensais em páginas de economia COMPOSITE



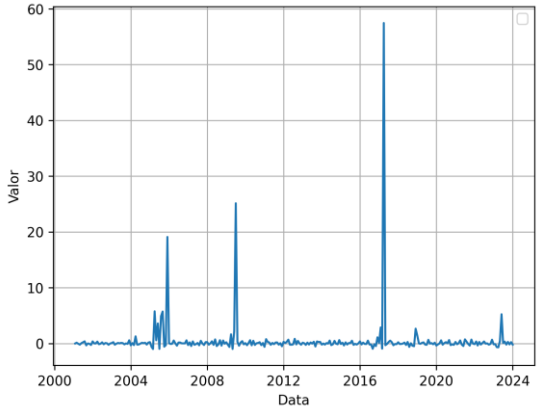
Fonte: Elaborado pelo autor

Gráfico 30 – Notícias fiscais mensais em páginas de economia PURE



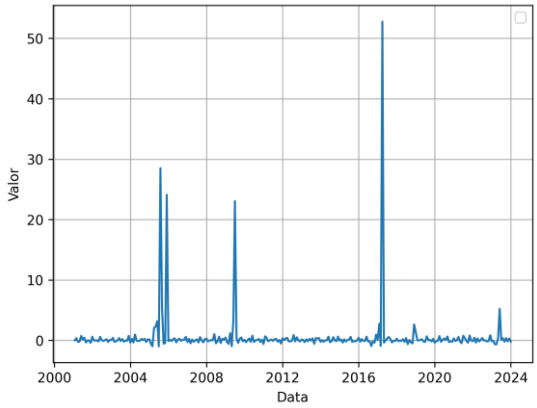
Fonte: Elaborado pelo autor

Gráfico 31 – Mudança de notícias fiscais mensais em páginas de economia COMPOSITE



Fonte: Elaborado pelo autor

Gráfico 32 – Mudança de notícias fiscais mensais em páginas de economia PURE



Fonte: Elaborado pelo autor