Conjunto de notícias para avaliação de investimentos em regiões do estado de São Paulo

Gabriel L. Melo[1], João V. C. Neres Sousa[1], Willian D. Oliveira[1], Lucas Mingardo[2], Carlos Freire[2], Agma J. M. Traina[1], Caetano Traina Jr.[1]

[1]Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo (USP)

[2]Fundação Sistema Estadual de Análise de Dados (Seade)
São Paulo – SP – Brasil

{melogabriel,joaovneres}@usp.br,{willian,agma,caetano}@icmc.usp.br

{lucasmingardo,carlosfreire}@seade.gov.br

Resumo. *Este artigo apresenta um conjunto de dados com notícias jornalísticas sobre investimentos produtivos no Estado de São Paulo (2016–2024), coletadas e classificadas pela Fundação Seade e que antecede a curadoria efetuada pela Pesquisa de Investimentos Anunciados no Estado de São Paulo (PIESP). A base inclui dados como título, fonte, texto integral e classificação manual de relevância. Descrevemos o processo de coleta e organização dos dados. Em seguida, discutimos potenciais aplicações como agrupamento semântico, extração de entidades nomeadas e sumarização automática. Também abordamos o desbalanceamento de classes nas notícias recentes e estratégias de amostragem para mitigação. O conjunto de dados visa apoiar pesquisas em economia regional, mineração de textos e aprendizado de máquina.*

Abstract. *This paper presents a structured dataset of news reports about productive investments in the São Paulo State of Brazil (2016–2024), collected and classified by Fundação Seade and preceding the curation performed by the Survey of Announced Investments in the State of São Paulo (PIESP). The dataset includes data such as title, source, full text, and manual relevance labels. We describe the process of data collection and organization. We proceed to discuss potential applications, such as semantic clustering, named entity recognition, and automatic summarization. We also tackle class imbalances in recent data and possible mitigations through sampling. The dataset is intended to support research in regional economics, text mining, and machine learning.*

1. Introdução

O crescimento acelerado do volume de informações disponíveis na internet tem transformado a maneira como instituições públicas e privadas monitoram eventos socioeconômicos. No contexto econômico, notícias sobre investimentos produtivos representam uma fonte estratégica para análises de mercado, estudos regionais e formulação de políticas públicas. Contudo, a coleta sistemática dessas informações enfrenta desafios relacionados à dispersão, heterogeneidade e falta de estrutura das publicações jornalísticas, exigindo processos robustos de captura, curadoria e análise [Reips et al. 2023].

Nesse cenário, a Fundação Sistema Estadual de Análise de Dados (Seade)[1], vinculada ao Governo do Estado de São Paulo, conduz desde 1998 a Pesquisa de Investimentos Anunciados no Estado de São Paulo (PIESP)[2], que atualmente monitora cerca de quarenta veículos jornalísticos. A coleta é realizada por meio de uma combinação de *web scraping*, clipagem digital e validação manual por especialistas, resultando em uma base de dados contínua e confiável. Para fins da PIESP, são considerados como 'de investimento' os anúncios relacionados à aplicação de capital voltada à ampliação da capacidade produtiva de empresas privadas ou entidades não estatais, excluindo-se investimentos exclusivamente estatais ou oriundos de organizações sem fins lucrativos (exceto nos setores de saúde e educação) [Fundação Sistema Estadual de Análise de Dados (SEADE) 2025].

Este artigo descreve e disponibiliza, em parceria entre os pesquisadores do Centro de Ciência de Dados para Estatísticas Públicas (CCDEP-FAPESP)[3] e a Fundação SEADE, o conjunto de dados estruturado de notícias que antecede a curadoria da PIESP referente ao período de 1º jan. 2016 a 31 dez. 2024, composto por mais de quatrocentos mil registros jornalísticos. O *dataset* inclui dados como título, texto completo da notícia, data de publicação, fonte jornalística e sua classificação quanto à relevância para o tema de investimentos.

Esta é a primeira publicação estruturada e pública deste *dataset*, e tem por objetivo disponibilizar sua reutilização em pesquisas acadêmicas e aplicações práticas. Acreditamos que a disponibilização pode facilitar estudos sobre economia regional, fomentar pesquisas em aprendizado de máquina voltadas à classificação automática de textos jornalísticos, além de subsidiar avaliações de políticas públicas voltadas à atração de investimentos produtivos. Embora outros conjuntos de dados com finalidades similares tenham sido disponibilizados para a exploração de dados para políticas públicas ([Freitas et al. 2023] e [Davis 2022]), este é, no melhor do nosso conhecimento, o primeiro conjunto de dados a estruturar notícias coletadas e classificadas como investimentos produtivos relevantes por especialistas de uma instituição oficial de estatística.

O restante deste artigo está estruturado da seguinte forma: a Seção 2 detalha a metodologia empregada para criação e curadoria dos dados, incluindo análises iniciais que mostram suas características fundamentais; a Seção 3 explora possíveis aplicações do conjunto, destacando desafios e limitações para seu uso efetivo; na Seção 3.3, são apresentadas informações sobre o licenciamento e a localização pública para *download* dos dados; por fim, as considerações gerais estão na Seção 4.

2. NIP-SP: O Conjunto de Notícias sobre Investimentos Produtivos Avaliadas no Estado de São Paulo

Esta seção detalha o conjunto de dados estruturado coletado pela PIESP da Fundação Seade. São documentadas as fontes onde são feitas as coletas, os métodos empregados, os critérios utilizados na seleção e classificação das notícias, e a estrutura final do conjunto disponibilizado. Essa documentação visa o reuso por pesquisadores interessados em economia regional, mineração de textos, análise de mídia e avaliação de políticas públicas.

[1]Fundação Seade. Disponível em: `https://www.seade.gov.br/`. Acesso em: 16-jun-2025.
[2]PIESP. Disponível em: `https://investimentos.seade.gov.br`. Acesso em: 16-jun-2025.
[3]CCDEP. Disponível em: `https://bv.fapesp.br/pt/auxilios/116227` Acesso em: 04-ago-2025.

2.1. Fontes e Metodologia de Coleta

O conjunto de dados foi obtido como resultado do monitoramento diário conduzido por empresa terceirizada, contratada pela Fundação Seade. As informações abrangem o período de 1º de janeiro de 2016 a 31 de dezembro de 2024, incorporando ao longo do período 62 veículos jornalísticos distintos, sendo 50 portais digitais e 12 periódicos impressos. As fontes *online* foram coletadas por meio de técnicas automatizadas de *web scraping*, enquanto os conteúdos dos jornais impressos passaram por digitalização e tratamento via clipagem digital. Ao todo, foram capturados 411.718 registros jornalísticos, dos quais 349.942 ($85\,\%$) são provenientes de fontes online e 61.776 ($15\,\%$) de veículos impressos. Ressalta-se que a seleção inicial das notícias foi baseada em palavras-chave previamente definidas pela equipe da Fundação Seade, como, por exemplo: ampliação, abertura, aquisição, entre outras. Essas palavras-chave são disponibilizadas no conjunto de dados.

Entre as principais fontes monitoradas estão veículos de abrangência nacional e regional com impacto significativo na cobertura econômica paulista. A tabela 1 apresenta os cinco veículos jornalísticos que mais contribuíram com notícias para a base, ordenados pela quantidade de notícias coletadas.

Tabela 1. Cinco fontes jornalísticas com maior número de registros coletados no período de 2016 a 2024

Fonte	Quantidade de notícias	Tipo de veículo
Valor Econômico	127.378	Online
Estadão	66.621	Online
Folha de S. Paulo	65.137	Online
Diário da Região (São José do Rio Preto)	23.476	Online
Valor Econômico (impresso)	20.503	Impresso

2.2. Classificação dos Dados

Após a coleta inicial, todas as notícias passaram por um processo de curadoria e classificação manual. Inicialmente, cada registro recebe a classificação padrão `N` (não classificado)[4]. Essas notícias não são encaminhadas para avaliação dos especialistas, mas são mantidas na base de dados, pois fazem parte do fluxo de coleta. Em seguida, os especialistas reavaliam os registros, atribuindo uma das seguintes categorias, com base nos critérios explicitados a seguir, estabelecidos pela PIESP:

- **I (Investimento)**: notícia relacionada diretamente à aplicação de capital privado para expansão produtiva, abrangendo implantação, ampliação ou modernização de empreendimentos. Esses casos são considerados potenciais investimentos produtivos, que posteriormente é avaliada para inclusão na PIESP;
- **L (Irrelevante)**: notícia explicitamente não relacionada à temática de investimentos produtivos;
- **N (Não classificada)**: notícia ainda não avaliada de forma definitiva.

[4]Notícias que não contêm nenhuma das palavras-chave definidas nos critérios de classificação da PIESP.

A classificação manual visa garantir precisão na identificação e seleção de conteúdos efetivamente relacionados a investimentos produtivos. Esse esforço traz confiança para a classificação, tornando-a apropriado para análises quantitativas e qualitativas, estudos comparativos de classificação automatizada e formulação embasada de políticas públicas. O conjunto disponibilizado combina as notícias obtidas pela coleta inicial com a classificação de investimentos produtivos que antecede a análise e disponibilização pública feita normalmente pela PIESP.

2.3. Pré-processamento

O conjunto de dados foi estruturado com foco na integridade temporal e na padronização das fontes jornalísticas, garantindo coerência e facilidade de utilização. Cada registro possui os atributos descritos na tabela 2. Os textos originais foram preservados integralmente, mantendo-se as quebras de linha, caracteres especiais e eventuais trechos incompletos, características comuns em conteúdos obtidos diretamente de portais de notícias.

Tabela 2. Descrição dos atributos presentes no conjunto de dados

Atributo	Descrição	Tipo do dado
`data_publicacao`	Data de publicação da notícia	Data (yyyy-mm-dd hh:mm:ss)
`titulo`	Título da notícia	Texto
`fonte`	Nome do veículo jornalístico	Texto
`texto`	Corpo completo da notícia	Texto
`flagnoticia`	Classificação manual: I, L, N	Texto (rótulo)

A menos da classificação manual, optou-se por não aplicar técnicas adicionais de normalização textual, remoção de duplicatas ou correções linguísticas, com o intuito de preservar a fidelidade ao conteúdo original, e permitir que outras alternativas de preparação possam ser avaliadas. Ressalta-se que algumas notícias apresentam trechos truncados devido à presença de mecanismos de restrição de acesso (*paywall*)[5].

Os dados originais foram obtidos a partir da base interna mantida pela Fundação Seade. O pré-processamento seguiu as seguintes etapas:

- Leitura integral dos registros originais;
- Equalização da nomenclatura das fontes jornalísticas;
- Padronização e ordenação cronológica das datas de publicação;
- Validação e filtragem dos registros com classificação inválida ou ausente no atributo `flagnoticia`;
- Exportação final dos atributos relevantes dos registros corretos.

A partir dos dados originais, foram criados três subconjuntos, que estão sendo disponibilizados conforme indicado na seção 3.3. Os seguintes subconjuntos estão disponíveis, junto com os respectivos dicionários de dados:

- **Completo (2023–2024)**: notícias com todas as categorias (`I`, `L`, `N`);
- **Investimentos (2016–2024)**: exclusivamente classificadas como `I`;
- **Palavras-chave**: lista completa usada na filtragem inicial.

Esses conjuntos são disponibilizados no formato CSV, trazendo contextos analíticos distintos.

[5]Mecanismo utilizado por veículos jornalísticos para restringir o acesso à notícia para não assinantes.

Tabela 3. Distribuição das classes no subconjunto completo (2023–2024), com detalhamento anual

Classe	2023	2024	Quantidade total	Porcentagem (%)
`I` (Investimento)	956	967	1.923	0,8
`L` (Irrelevante)	75.683	124.994	200.677	83,15
`N` (Não classificada)	35.024	3.717	38.741	16,05
Total	**111.663**	**129.678**	**241.341**	**100,00**

Tabela 4. Percentual de valores ausentes por atributo no subconjunto completo (2023–2024)

Atributo	Porcentagem de nulos (%)
`data_publicacao`	0,00
`titulo`	1,28
`fonte`	0,00
`texto`	0,09
`flagnoticia`	0,00

2.4. Descrição dos Dados

O conjunto de dados produzido pela PIESP está estruturado conforme tabela 2 e contendo as informações essenciais para a análise dos registros jornalísticos. Nessa seção, dois subconjuntos são descritos: o *Subconjunto Completo* e o *Subconjunto Investimentos*.

2.4.1. Subconjunto Completo (2023–2024)

Este subconjunto contém as notícias capturadas entre 1º de janeiro de 2023 e 31 de dezembro de 2024, totalizando 241.341 registros. Cada notícia foi classificada em três categorias: `I` (investimento), `L` (irrelevante) e `N` (não classificada). A tabela 3 apresenta a distribuição dessas classes, com detalhamento por ano, quantidade total e porcentagem em relação ao subconjunto completo.

Adicionalmente, a tabela 4 exibe o percentual de valores ausentes por atributo no subconjunto completo de notícias entre 2023 e 2024. Observa-se que os campos `titulo` e `texto` apresentam proporções de nulos de $1{,}28\,\%$ e $0{,}09\,\%$, respectivamente. Análises adicionais revelaram que os registros com título ausente são exclusivamente provenientes de fontes do tipo *impresso*, o que sugere limitações na etapa de digitalização e clipagem desses documentos. Por outro lado, os casos com texto ausente ocorrem exclusivamente em fontes *online*, o que pode ser atribuído à presença de mecanismos de restrição de acesso ao conteúdo completo, como *paywalls*, ou falhas pontuais no carregamento da página durante o processo de raspagem automatizada.

Não são armazenados registros onde os campos `fonte` e `texto` estejam simultaneamente ausentes, o que demonstra consistência na estrutura mínima dos dados coletados. Entre os 212 registros que apresentam texto ausente, mas possuem título, $67{,}92\,\%$ não contém nenhuma das palavras-chave utilizadas na filtragem inicial. Ainda assim, todos esses registros foram mantidos com a classificação `N` (não classificada). Esse resultado sugere que, mesmo sem o corpo da notícia, o título foi considerado suficientemente

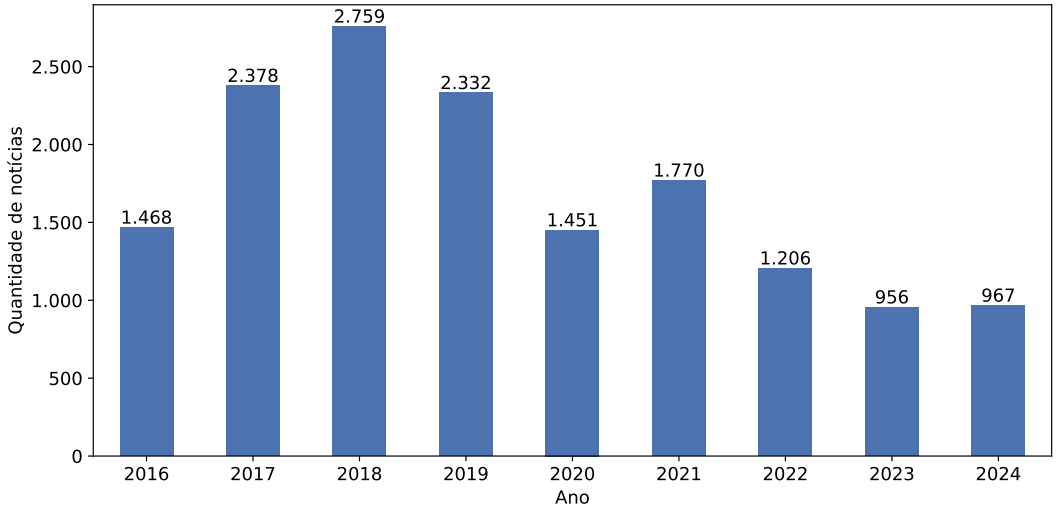

Figura 1. Número de notícias classificadas como investimento por ano (2016–2024)

informativo para justificar sua retenção no sistema, possibilitando avaliação posterior pela equipe de curadoria.

2.4.2. Subconjunto Investimentos (2016–2024)

Este subconjunto inclui exclusivamente notícias classificadas manualmente como investimento (`I`), relacionadas à aplicação produtiva de capital no Estado de São Paulo. Ele contém 15.287 registros coletados entre janeiro de 2016 e dezembro de 2024. A figura 1 apresenta a distribuição anual desses registros, com aumento no número de notícias até 2018 e variações nos anos subsequentes. Essas flutuações podem refletir mudanças na cobertura jornalística, nos critérios de classificação adotados ou na própria dinâmica dos investimentos produtivos ao longo do período analisado.

A figura 2 apresenta as dez fontes jornalísticas com maior número de notícias classificadas como investimentos, destacando veículos de abrangência nacional e regional estratégicos para a análise econômica. Cabe destacar que não existem valores ausentes nos campos `titulo` e `texto` neste subconjunto, o que reforça sua completude e adequação para análises baseadas em conteúdo textual.

2.4.3. Comparativo entre Subconjuntos: Fonte e Densidade Textual

Dois aspectos descritivos relevantes são comparados entre os subconjuntos *Completo (2023–2024)* e *Investimentos (2016–2024)*: a origem das fontes jornalísticas e as características dos campos textuais.

A predominância das fontes digitais é evidente em ambos os subconjuntos. No período recente (2023–2024), fontes online representam 226.160 registros (93,71 %), enquanto os periódicos impressos contribuíram com 15.181 registros (6,29 %). Já no subconjunto de investimentos (2016–2024), as fontes digitais correspondem a 12.964 regis-

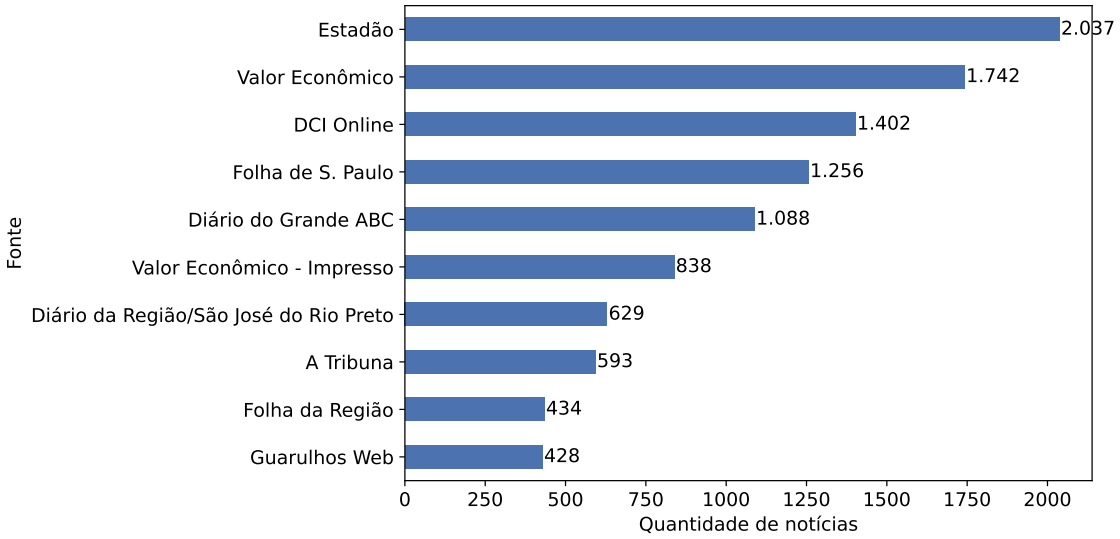

Figura 2. As 10 fontes com mais notícias classificadas como investimento (2016–2024)

Tabela 5. Estatísticas descritivas dos campos textuais em ambos os subconjuntos

Campo	Subconjunto	Média	Mediana	Desvio padrão
`titulo`	Completo (2023-2024)	68,8	74,0	25,3
	Investimentos (2016-2024)	58,1	59,0	20,0
`texto`	Completo (2023-2024)	3.667,6	2.791,0	4.061,5
	Investimentos (2016-2024)	3.060,2	2.665,0	2.453,6

tros (84,80 %) e as fontes impressas a 2.323 registros (15,20 %). A figura 3 contém essa distribuição com predominância dos veículos digitais durante o período analisado.

Além da origem das fontes, os subconjuntos diferem também quanto à densidade informacional dos campos textuais. A tabela 5 apresenta estatísticas descritivas dos atributos `titulo` e `texto` em ambos os conjuntos, incluindo média, mediana e desvio padrão do número de caracteres. Observa-se que os títulos das notícias no subconjunto de investimentos (2016–2024) são, em média, mais curtos (58,1 caracteres) do que os do subconjunto completo (2023–2024), cuja média é de 68,8 caracteres. Já os corpos textuais são, em média, mais extensos no subconjunto completo (3.667,6 caracteres) do que no de investimentos (3.060,2 caracteres), embora este último apresente maior homogeneidade, com menor desvio padrão.

Essas diferenças podem refletir características editoriais distintas entre os subconjuntos: enquanto o conjunto completo agrega uma variedade maior de conteúdos e estilos jornalísticos, o conjunto de investimentos tende a concentrar textos mais objetivos e técnicos, focados em anúncios específicos de aplicação produtiva de capital.

2.5. Considerações finais

Por fim, o terceiro conjunto disponibilizado corresponde à lista completa de palavras-chave utilizadas na etapa de filtragem inicial das notícias. Ele é composto por um único

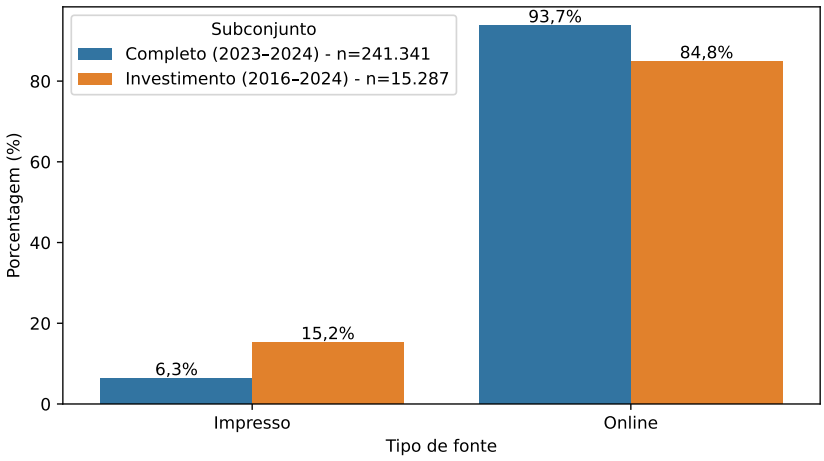

Figura 3. Distribuição por tipo de fonte jornalística nos subconjuntos Completo (2023–2024) e Investimentos (2016–2024)

atributo, contendo 118 palavras distintas, que abrangem termos relacionados à instalação, inauguração, ampliação, entre outros eventos vinculados a investimentos produtivos. Essas palavras foram definidas pela equipe da Fundação Seade com base na experiência acumulada ao longo da execução da PIESP, servindo como critério preliminar de relevância durante a coleta automatizada.

É importante ressaltar que o conjunto aqui descrito não está publicamente disponível nos resultados oficiais da PIESP. Este conjunto constitui uma base independente, especialmente estruturada para estudos acadêmicos e análises metodológicas sobre notícias relacionadas a investimentos produtivos. A Fundação Seade divulga oficialmente na PIESP apenas os resultados consolidados após a validação técnica e institucional.

Os conjuntos de dados descritos nesta seção oferecem oportunidades analíticas diversas, bem como desafios metodológicos específicos. Algumas dessas possibilidades e limitações, que são alvo das atividades atuais da PIESP, são discutidas na próxima seção.

3. Aplicações e Desafios

O conjunto de dados disponibilizado oferece amplas possibilidades de reuso em tarefas de mineração de textos, aprendizado de máquina e apoio à formulação de políticas públicas. Por conter milhares de notícias jornalísticas classificadas quanto à presença de investimentos produtivos, ele permite análises semânticas, extração de padrões e automação de processos de triagem.

3.1. Agrupamento Temático com BERTopic

Como estudo exploratório, aplicou-se uma técnica de modelagem de tópicos baseada no método BERTopic [Grootendorst 2022], com foco exclusivo nas 15.287 notícias classificadas como investimento (`I`). Utilizou-se o modelo BERTimbau [6] [Souza et al. 2020]

[6]O BERTimbau, algoritmo desenvolvido pela NeuralMind que consiste no BERT (Bidirectional Encoder Representations from Transformers), desenvolvido pelo Google para melhorar o motor de buscas da plataforma, treinado para língua portuguesa.

para gerar *embeddings* semânticos dos textos, seguido por redução de dimensionalidade com UMAP [McInnes et al. 2018] e agrupamento por densidade via HDBSCAN [Campello, R J G B.. et al. 2013]. O resultado foi a identificação de 51 grupos temáticos coerentes, com $14{,}17\,\%$ de pontos considerados *outliers*.

A representação dos tópicos foi gerada por meio do método c-TF-IDF, que considera a frequência de termos por *cluster*, produzindo palavras-chave para cada grupo de notícias. Essa abordagem revelou padrões recorrentes no conteúdo jornalístico, como menções a inaugurações de fábricas, investimentos em energia renovável e expansão de setores logísticos. A Figura 4 ilustra a distribuição dos tópicos em espaço bidimensional, onde se pode constatar que o *dataset* de notícias inclui dados com boa separabilidade e potencial para exploração empírica, bem como servir de base para o desenvolvimento de novas técnicas de análise.

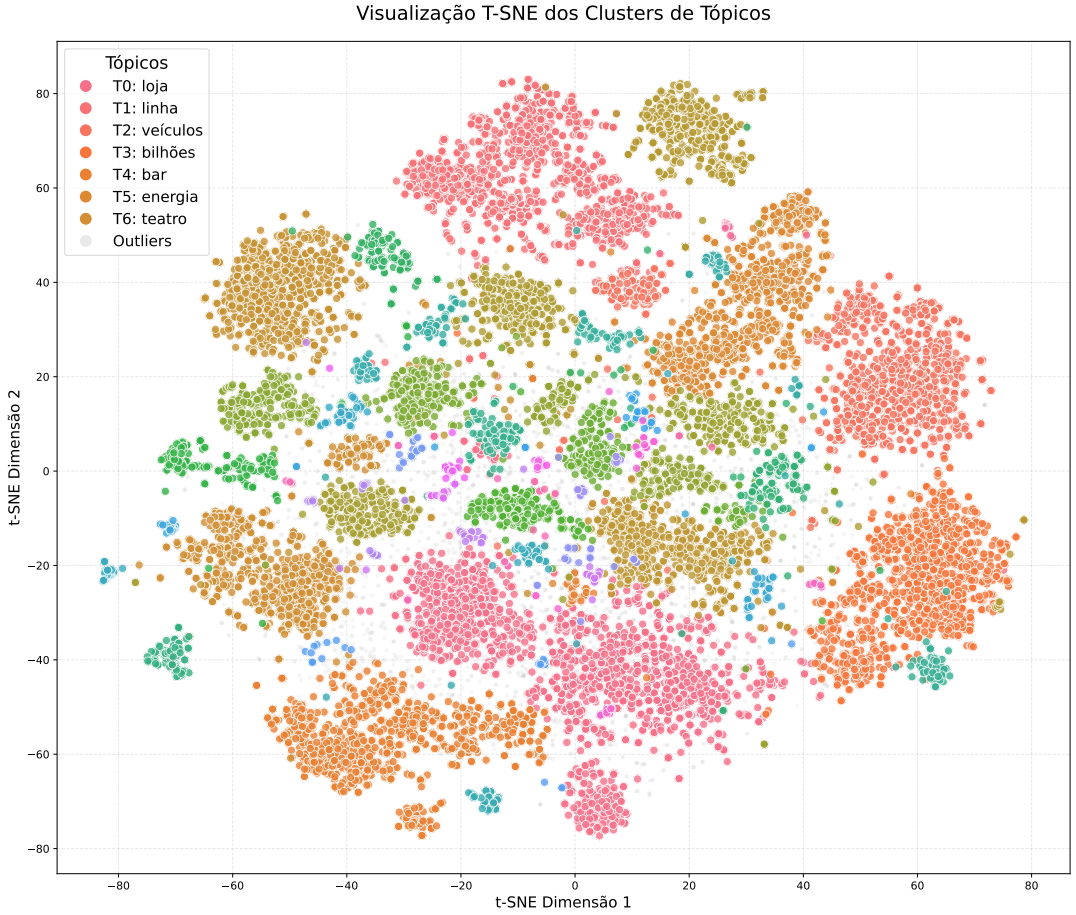

Figura 4. Representação bidimensional dos tópicos identificados via BERTopic nas notícias classificadas como investimento

3.2. Perspectivas de Uso

Além da modelagem de tópicos, o *dataset* possibilita a aplicação de técnicas como *Named Entity Recognition* (NER) para extrair automaticamente entidades como empresas e municípios envolvidos nos investimentos, o que amplia a aplicabilidade da base para análises geográficas e setoriais [Albuquerque, H. O. et al. 2023].

Outra perspectiva promissora é o uso de modelos de sumarização automática para apoiar especialistas na triagem de notícias. Essa funcionalidade pode acelerar a validação manual das publicações ao oferecer resumos sintéticos, especialmente em contextos com grandes volumes de dados e necessidade de tratamento rápido, como as atualizações mensais da PIESP [Barros, T. et al. 2021].

Por fim, o conjunto completo referente aos anos de 2023 e 2024, que inclui as três classes (`I`, `L`, `N`), apresenta um desbalanceamento significativo entre os rótulos, com a classe `I` representando menos de $1\,\%$ dos registros. Esse desbalanceamento pode comprometer o desempenho de classificadores supervisionados, especialmente em tarefas de detecção de investimentos. Para mitigar esse desafio, técnicas de amostragem e geração sintética de amostras, como o uso de redes adversárias generativas (GANs) [Goodfellow I. et al. 2020] e algoritmos como *SMOTE* [Chawla, N. V. et al. 2002], podem ser investigadas. Tais abordagens têm-se mostrado eficazes em cenários com dados tabulares e rótulos escassos [Cavalcanti, A. et al. 2024].

3.3. *Download* e requisição de citação

O NIP-SP é disponibilizado publicamente sob a licença CC BY-NC 2.0[7] e disponível em `https://repositorio.seade.gov.br/dataset/noticias-de-investimento-produtivos`. No caso deste conjunto ser utilizado para qualquer propósito, é solicitado que este artigo seja referenciado.

4. Conclusão

Este artigo apresenta e documenta um conjunto de dados composto por notícias jornalísticas sobre investimentos produtivos no Estado de São Paulo, coletadas e classificadas pela Fundação Seade. Trata-se de uma contribuição inédita à comunidade científica, uma vez que os dados aqui disponibilizados correspondem a uma etapa anterior àquela agregada e divulgada oficialmente pela Seade, antecedendo os filtros institucionais e validações técnicas realizadas para a PIESP.

A disponibilização desse conjunto de dados visa fomentar pesquisas em múltiplas áreas, incluindo economia regional, mineração de textos, visualização de dados, aprendizado de máquina e políticas públicas. Por incluir rótulos manuais de classificação, textos integrais e metadados relevantes, o *dataset* se mostra adequado tanto para estudos supervisionados quanto para abordagens exploratórias e não supervisionadas.

Foram destacados exemplos de uso envolvendo agrupamento temático com BERTopic, análise de desbalanceamento de classes, e potenciais aplicações com técnicas de extração de entidades e sumarização automática. Tais possibilidades ilustram a riqueza do material e seu potencial para apoiar análises de tendências econômicas, automação de processos institucionais e desenvolvimento de novas metodologias em ciência de dados.

Ao tornar pública esta base intermediária, espera-se contribuir com maior transparência, reprodutibilidade e inovação nas formas de monitoramento de investimentos e sua comunicação para a sociedade. Estudos futuros podem expandir sua utilização, aplicando novos modelos linguísticos, análises regionais comparativas e técnicas de enriquecimento semântico para aumentar ainda mais seu valor analítico.

[7]Licença CC BY-NC 2.0. Disponível em: https://creativecommons.org/licenses/by-nc/2.0/ . Acesso em: 16 jun. 2025.

Agradecimentos

Este trabalho foi apoiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) (Processos 23/18026-8 e 24/13328-9), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

Referências

Albuquerque, H. O. et al. (2023). Named entity recognition: a survey for the Portuguese language. *Procesamiento del Lenguaje Natural*.

Barros, T. et al. (2021). Sumarizacão automática de notícias crime no contexto da polícia federal. In *Anais Estendidos do XXXVI Simpósio Brasileiro de Bancos de Dados*, pages 127–133, Porto Alegre, RS, Brasil. SBC.

Campello, R J G B.. et al. (2013). Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg.

Cavalcanti, A. et al. (2024). Avaliação de técnicas de balanceamento de dados na detecção de fraude em transações online de cartão de crédito. In *Anais do XXXIX SBBD*, pages 694–700, Porto Alegre, RS, Brasil. SBC.

Chawla, N. V. et al. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Davis, P. (2022). Indicadores e dados municipais: Um banco de dados para avaliar a eficiência das despesas públicas. In *Anais do IV Dataset Showcase Workshop*, pages 79–90, Porto Alegre, RS, Brasil. SBC.

Freitas, J. B., Clarindo, J. P., and Aguiar, C. (2023). Spsafe: um dataset sobre dados de criminalidade no estado de são paulo. In *Anais do V Dataset Showcase Workshop*, pages 48–57, Porto Alegre, RS, Brasil. SBC.

Fundação Sistema Estadual de Análise de Dados (SEADE) (2025). Anexo metodológico — seade investimentos.

Goodfellow I. et al. (2020). Generative adversarial networks. *CACM*, 63(11):139–144.

Grootendorst, M. (2022). BERTopic: neural topic modeling with a class-based TF-IDF procedure. *ArXiv Ref. 2203.05794*, page 10.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Reips, L., Musicante, M., Vargas-Solar, G., Pozo, A., and Hara, C. (2023). Enow - extrator de dados de notícias da web. In *Anais Estendidos do XXXVIII Simpósio Brasileiro de Bancos de Dados*, pages 78–83, Porto Alegre, RS, Brasil. SBC.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In Cerri, R. and Prati, R. C., editors, *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.