# Novelty Detection: Identificação de Temas Emergentes em Corpora de Notícias

1st Martín Avila Buitron
*Eng da Computação*
*Universidade Tecnológica Federal de Paraná*
Toledo, Brasil
martinavila@alunos.utfpr.edu.br

***Abstract*—A *Novelty Detection*, ou detecção de novidade em documentos, é uma tarefa desafiadora e de grande relevância na atualidade. Este tema é especialmente significativo na área de Linguagem Natural (NLP). O intuito é diferenciar documentos que pertencem a um conjunto já conhecido daqueles que apresentam características inéditas ou emergentes. Este trabalho explora diferentes algoritmos para a detecção de novidade, comparando os resultados obtidos utilizando o *dataset* de *benchmark* TAP-DLND 1.0. Os resultados destacaram as diferenças entre as técnicas avaliadas e suas implicações em diferentes tópicos, destacando a necessidade de explorar métodos mais avançados.**

***Index Terms*—Novelty Detection, TAP-DLND 1.0, Ciência de dados**

## I. Introdução

Quando se fala sobre detecção de novidade, é inevitável mencionar o conceito de KDD. "Knowledge Discovery in Databases (KDD), definido como o processo não trivial de identificar informações válidas, novas, potencialmente úteis e, em última análise, conhecimento compreensível a partir dos dados" [4]. Partindo desse conceito, o objetivo de detecção de novidade está na identificação de padrões inéditos em relação a um conjunto de referências previamente conhecido.

A identificação de dados novos ou anômalos em relação ao conjunto inicial desempenha um papel essencial em diversas áreas, como segurança da informação, monitoramento de fraudes e análise de conteúdo. No contexto de notícias, essa tarefa exige métodos capazes de diferenciar material redundante de documentos genuinamente originais, considerando aspectos como semântica das palavras, sinônimos, análises léxicas, dados temporais e estrutura contextual.

Além de sua aplicabilidade prática, o avanço nessa área contribui para o desenvolvimento de ferramentas que valorizem a produção de conhecimento autêntico, reforçando a necessidade de soluções robustas e precisas. Neste trabalho, avaliamos o desempenho de três algoritmos de detecção de novidade em notícias, com resultados em torno de 0.6 de *F1-score*, evidenciando a necessidade de explorar métodos mais avançados e integrados para alcançar desempenhos mais satisfatórios nessa tarefa.

Este trabalho, a partir de notícias inéditas de um *dataset* de *benchmark*, visa responder às seguintes perguntas de pesquisa:

1) Com base em um conjunto conhecido de notícias, qual é o desempenho dos algoritmos *Local Outlier Factor (LOF)*, *Isolation Forest* e *Elliptic Envelope* na detecção de novidades em conjuntos desconhecidos, utilizando representações baseadas em vetores TF-IDF?
2) Quais são as diferenças temáticas mais significativas entre artigos classificados como novidade e não novidade?
3) Como avaliar os resultados dos algoritmos de detecção de novidades tanto de forma geral (para todas as categorias) quanto para cada categoria individualmente?

Identify applicable funding agency here. If none, delete this.

## II. Processamento de dados

### A. Obtenção de dados

Para este estudo, foi utilizado o corpus **TAP-DLND 1.0** [1], disponibilizado publicamente. O conjunto de dados está organizado em categorias, com subpastas divididas em duas seções principais: `source`, que contém artigos iniciais considerados como a "base" da análise, e `target`, que reúne textos classificados como *novel* (novidade) ou *non-novel* (não novidade). No total, o corpus inclui 6104 notícias e 10 categorias como pode ser visto na Figura 1. Os documentos da pasta `source` servem como referência para identificar se os textos na pasta `target` apresentam informações inéditas. Essa estrutura fornece uma base sólida para aplicar os objetivos planejados da pesquisa.

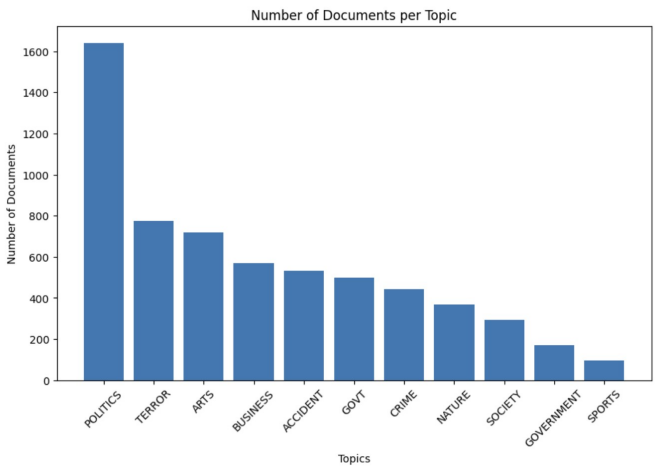


Fig. 1. Quantidade de documentos por Categoria.

O corpus contém diversas características que auxiliam no processo de análise e detecção de novidade. Cada documento no conjunto de dados está associado a múltiplos atributos, incluindo:

- **category**: indica a categoria temática do artigo, `ARTS`, `BUSINESS`, `CRIME`, `GOVT`, `NATURE`, `POLITICS`, `SOCIETY`, `SPORTS` ou `TERROR`.
- **event_id** e **news_id**: identificam de forma única cada evento e cada notícia.
- **content**: contém o texto completo do artigo.
- **is_source**: indica se a notícia pertence à pasta `source` (base) ou `target` (análise).
- **DOP**: data de publicação do artigo.
- **publisher**: fonte responsável pela publicação.
- **title**: título do artigo.
- **topic**: tema geral da notícia.
- **sentence**: número de sentenças no artigo.
- **words**: contagem total de palavras no texto.
- **DLA** (*Decision Label Annotation*): especifica se o artigo na pasta `target` é classificado como novidade ou não novidad.
- **SLNS**: Não foi descrito pelo autor.

Essa estrutura detalhada permite uma realizar análise adequada, embora seja necessário realizar algumas etapas de preparação e limpeza dos dados para garantir a qualidade das informações a serem utilizadas nas análises.

### *B. Processamentos Aplicados*

Na etapa de pré-processamento, inicialmente foi realizada a *tokenização* de palavras, seguida da remoção de pontuação e de *stopwords*, como "say", que apresentava alta frequência, mas não agregava relevância ao contexto textual. Por fim, foi aplicada a lematização para reduzir as palavras à sua forma base. O resultado dessas etapas pode ser observado nas Figuras 2 e 6, que destacam as palavras mais relevantes do tópico `Sports`, após a remoção de termos irrelevantes.


Fig. 2. Wordcloud para Sports.

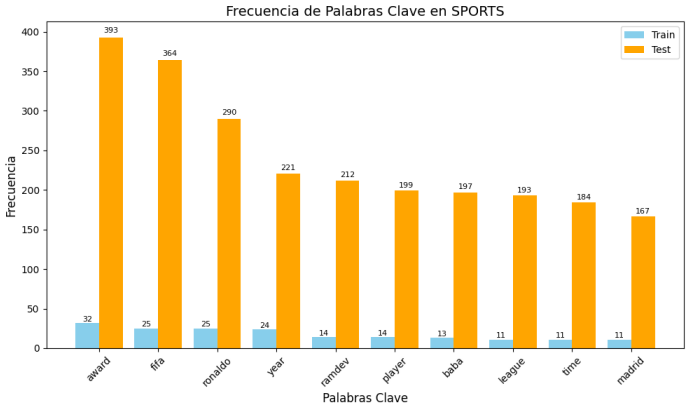


Fig. 3. Comparação da frequência de palavras nos conjuntos de treinamento e teste de SPORTS.

Na etapa seguinte, foram excluídas *features* irrelevantes, redundantes ou de difícil interpretação, como `DOP`, `eventname`, `Publisher` e `topic`. A Figura 4 ilustra a distribuição de publicadores, evidenciando que a identificação da fonte não contribui significativamente para discriminar o conteúdo textual. Variáveis como `ids`, títulos e dados temporais também foram removidas, já que apresentavam valores majoritariamente nulos. A variável `SLNS` foi descartada devido à sua baixa representatividade e falta de clareza, o que poderia comprometer o desempenho do modelo.

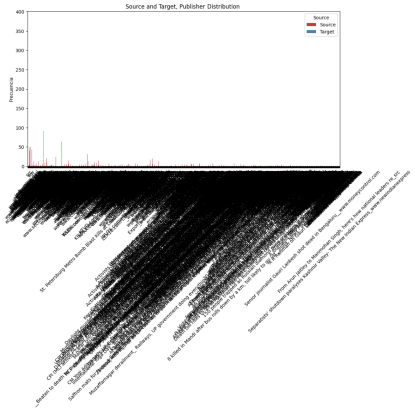


Fig. 4. Quantidade de documentos por publicador.

Durante a etapa de análise das *features*, a coluna `DLA`, que indica se um artigo é novidade ou não novidade, apresentou inconsistências na categorização. A distribuição percentual revelou as seguintes caraterísticas: **Non-Novel** (43.61%), **Novel** (42.39%), **non-novel** (7.95%), **novel** (6.02%), **Non-Novelvel** (0.02%) e **NovNon-Novelel** (0.02%). Os dados `Non-Novelvel` e `NovNon-Novelel` representam erros de formatação humana no conjunto de dados e foram tratadas durante o pré-processamento para minimizar seu impacto. Além disso, variáveis categóricas, como `category`, foram processadas utilizando a função *OneHotEncoder* de scikit-learn [5], que transforma os valores categóricos em atributos booleanos, evitando a introdução de pesos arbitrários.

Como resultado do pré-processamento, com o entendimento das *features* e o conjunto de dados uniformizado e otimizado, foram gerados gráficos que contribuíram para a análise exploratória. Essa etapa foi fundamental para descrever as perguntas de pesquisa iniciais e fornecer informações relevantes para a escolha e implementação de algoritmos de detecção de novidade. Nesse contexto, foi utilizada a biblioteca Scikit-learn [5], que fornece os três algoritmos selecionados: *Local Outlier Factor (LOF)*, *Isolation Forest* e *Elliptic Envelope*. Para viabilizar a aplicação desses algoritmos, empregou-se a ferramenta `TfidfVectorizer`, para representa o corpus no formato de vetores.

Além disso, foram realizadas análises complementares, incluindo a avaliação das distribuições das variáveis, a aplicação do *Principal Component Analysis* (PCA) e a investigação de correlações. Como ilustrado na Figura 5, o PCA permitiu reduzir a dimensionalidade dos dados, condensando *features*, como tamanho e quantidade de palavras, em variáveis representativas, preservando ao máximo a informação relevante em apenas uma característica.

Devido à elevada dimensionalidade dos vetores gerados durante a aplicação do **TF-IDF**, configurado com *ngram_range* igual a 1 — o que significa que cada palavra individual foi tratada como uma característica —, o PCA foi utilizado como método de redução dimensional. Essa abordagem garantiu que as características mais significativas fossem preservadas, permitindo um processamento mais eficiente e representações mais compactas dos dados.

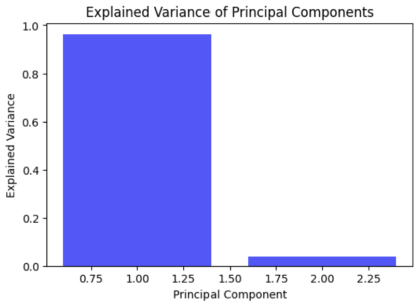


Fig. 5. Quantidade de documentos por publicador.

Por fim, para configurar os algoritmos de forma otimizada, foi utilizado o método de *Grid Search*, que testa combinações pré-definidas de hiperparâmetros. Esse procedimento garantiu que cada algoritmo fosse ajustado para alcançar os melhores resultados no contexto da detecção de novidade.

## III. Algoritmos

Na escolha dos algoritmos, foi relevante observar que os autores do próprio *dataset* publicaram um artigo no qual avaliaram diversos algoritmos e métodos de processamento de texto, abrangendo análises semânticas, léxicas, léxico-semânticas e modelos de linguagem, como o KLD.

Nesta pesquisa, optou-se por utilizar exclusivamente uma abordagem léxica, empregando o TF-IDF para representar os textos. Essa técnica analisa a frequência dos termos em documentos e ajusta sua relevância com base na ocorrência no conjunto total de documentos. Tal abordagem mostrou-se suficiente para construir representações vetoriais robustas, viabilizando a aplicação dos algoritmos selecionados. Com base em métodos tradicionais e nas recomendações apresentadas na documentação da biblioteca Scikit-learn [5], os algoritmos avaliados foram:

### A. Local Outlier Factor (LOF)

Para compreender o algoritmo *Local Outlier Factor (LOF)*, é importante destacar que ele é um método projetado para identificar outliers em espaços multidimensionais. Essa característica permite sua aplicação tanto em *Outlier Detection* quanto em *Novelty Detection*. Como descrito pelo autor do algoritmo, Markus M. Breunig, ”o LOF introduz um fator de outlier local para cada objeto no conjunto de dados, indicando seu grau de discrepância. Este é, até onde sabemos, o primeiro conceito de outlier que também quantifica o quão longe um objeto está. O fator de outlier é local no sentido de que apenas uma vizinhança restrita de cada objeto é levada em conta.” [4]. Em resumo, o LOF identifica anomalias ao comparar a densidade de um ponto com a densidade dos seus vizinhos mais próximos. Pontos com densidades significativamente menores que a de seus vizinhos são considerados outliers.

### B. Isolation Forest

Partindo da ideia de que anomalias podem ser interpretadas como instâncias suscetíveis à isolação, o método *Isolation Forest* se apresenta como uma ferramenta interessante para ser testada. Segundo Fei Tony Liu, ”essa técnica é diferente das demais porque não utiliza o conceito de distâncias, detectando anomalias puramente com base no conceito de isolamento, sem empregar qualquer medida de distância ou densidade – o que a torna fundamentalmente diferente de todos os métodos existentes.” [2]. Basicamente esta técnica visa construir de árvores de isolamento, que segmentam iterativamente os dados até que as amostras sejam completamente isoladas. Essa abordagem é usada, já que as instâncias consideradas como novidade são isoladas mais rapidamente em comparação às amostras regulares.

### C. Elliptic Envelope

A partir da definição de Mohammad Ashrafuzzaman, ”O algoritmo *Elliptic Envelope* é um método de aprendizado de máquina não supervisionado que utiliza estimativas de covariância em dados com distribuição gaussiana. O envelope elíptico tenta formar um agrupamento elíptico e se ajusta às principais instâncias dessa classe. Instâncias distantes do agrupamento são então consideradas como anomalias.” [3]. No contexto desta pesquisa, os pontos que estão fora da elipse definida pelo modelo são classificados como outliers ou novidades. Esse método é particularmente útil em cenários onde os dados apresentam distribuições aproximadamente normais.

Finalmente, foram implementadas três funções, uma para cada técnica selecionada. Cada função foi projetada para receber as principais *features* identificadas como relevantes para a tarefa, incluindo `PC1_normalized`, `category` e `texto_limpo`. Os dados foram separados em conjuntos de treinamento, correspondendo à pasta `source`, e de teste, correspondendo à pasta `target`. Os resultados foram avaliados com base em três métricas principais: *precision*, *recall* e *F1-score*.

## IV. Resultados

A partir dos vetores gerados, foi realizada uma análise considerando tanto o conjunto completo de categorias quanto cada uma de forma individual. O principal enfoque foi dado ao *F1-score*, bem como à proporção de documentos classificados como *Novelty(N)* e *Non-Novelty(NN)* . Essa abordagem permitiu avaliar o desempenho dos algoritmos em termos de equilíbrio entre as classes e sua capacidade de discriminar padrões novos e já conhecidos.

TABLE I
Resultados - Isolation Forest

| Category | Precision | Recall | Best F1 | N F1-S | NN F1-S |
|---|---|---|---|---|---|
| All CAT | 0.485 | 0.493 | **0.435** | 0.179 | 0.691 |
| ARTS | 0.568 | 0.515 | 0.399 | - | - |
| BUSINESS | 0.631 | 0.594 | 0.588 | - | - |
| CRIME | 0.425 | 0.464 | 0.356 | - | - |
| GOVT | 0.490 | 0.491 | 0.490 | - | - |
| NATURE | 0.430 | 0.483 | 0.433 | - | - |
| POLITICS | 0.445 | 0.473 | 0.398 | - | - |
| SOCIETY | 0.515 | 0.508 | 0.282 | - | - |
| SPORTS | 0.694 | 0.541 | 0.459 | - | - |
| TERROR | 0.437 | 0.458 | 0.437 | - | - |

TABLE II
Resultados - Elliptical Envelope

| Category | Precision | Recall | Best F1-S | N F1-S | NN F1-S |
|---|---|---|---|---|---|
| All CAT | 0.589 | 0.588 | **0.589** | 0.521 | 0.650 |
| ARTS | 0.517 | 0.515 | 0.498 | - | - |
| BUSINESS | 0.583 | 0.587 | 0.577 | - | - |
| CRIME | 0.488 | 0.492 | 0.414 | - | - |
| GOVT | 0.454 | 0.417 | 0.324 | - | - |
| NATURE | 0.456 | 0.460 | 0.457 | - | - |
| POLITICS | 0.513 | 0.513 | 0.513 | - | - |
| SOCIETY | 0.386 | 0.500 | 0.436 | - | - |
| SPORTS | 0.217 | 0.500 | 0.302 | - | - |
| TERROR | 0.597 | 0.592 | 0.594 | - | - |

TABLE III
Resultados - LOF

| Category | Precision | Recall | Best F1-S | N F1-S | NN F1-S |
|---|---|---|---|---|---|
| All CAT | 0.607 | 0.606 | **0.606** | 0.539 | 0.674 |
| ARTS | 0.316 | 0.484 | 0.334 | - | - |
| BUSINESS | 0.350 | 0.369 | 0.356 | - | - |
| CRIME | 0.284 | 0.337 | 0.278 | - | - |
| GOVT | 0.496 | 0.495 | 0.493 | - | - |
| NATURE | 0.409 | 0.484 | 0.426 | - | - |
| POLITICS | 0.343 | 0.434 | 0.340 | - | - |
| SOCIETY | 0.449 | 0.440 | 0.340 | - | - |
| SPORTS | 0.808 | 0.812 | 0.809 | - | - |
| TERROR | 0.346 | 0.465 | 0.391 | - | - |

O *Local Outlier Factor (LOF)* foi configurado com os seguintes parâmetros, a contaminação foi ajustada para 0.4, devido à distribuição equilibrada entre novidade e não novidade. A métrica de distância euclidiana apresentou os melhores resultados, em conjunto com a configuração de uma vizinhança composta por apenas 3 elementos. O algoritmo demonstrou um desempenho geral promissor, com *F1-score* de 0.606. O *Novelty F1-Score* foi de 0.539, enquanto o *No Novelty F1-Score* alcançou 0.674, indicando que o algoritmo teve melhor desempenho na identificação de padrões já conhecidos não novidade. Entre as categorias, o melhor desempenho foi observado em `SPORTS`, com *F1-score* de 0.809, sugerindo que o algoritmo é mais eficaz com menos quantidade de documentos e com temas bem definidos. Em contrapartida, categorias como `CRIME` (*F1-score* de 0.278) e `SOCIETY` (*F1-score* de 0.340) apresentaram resultados abaixo do esperado, possivelmente devido à maior variabilidade ou à ausência de padrões claros.

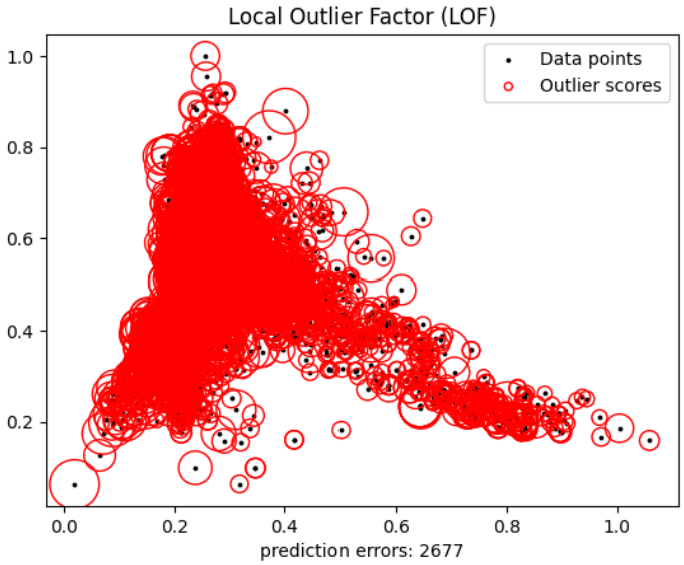


Fig. 6. Gráfico gerado pelo *Local Outlier Factor (LOF)*

O gráfico gerado para todas as categorias pelo *Local Outlier Factor (LOF)* complementa os resultados obtidos, destacando a eficácia do algoritmo em identificar anomalias em categorias estruturadas, como `SPORTS`, que apresentou um *F1-score* de 0.809. No entanto, a alta densidade de pontos no núcleo reflete as dificuldades do LOF em discriminar padrões em categorias mais dispersas, como `CRIME` (*F1-score* de 0.278). Os pontos mais afastados no gráfico representam, com alta probabilidade, os outliers detectados pelo algoritmo.

Para o *Isolation Forest*, foram utilizados os parâmetros `n_estimators` de 500 e contaminação de 0.4. O algoritmo apresentou um desempenho geral inferior, com *F1-score* de 0.435. O *Novelty F1-Score* foi de apenas 0.179, enquanto o *No Novelty F1-Score* atingiu 0.691, evidenciando a dificuldade do modelo em identificar documentos classificados como

novidade. Entre as categorias, o melhor desempenho foi observado em BUSINESS, com *F1-score* de 0.588, enquanto categorias como SOCIETY (*F1-score* de 0.282) e CRIME (*F1-score* de 0.356) apresentaram resultados muito aquém do esperado.

Já o *Elliptic Envelope* configurado com uma contaminação maior, de 0.5, e support_fraction definido como None. O algoritmo alcançou um *F1-score* geral de 0.589, com *Novelty F1-Score* de 0.521 e *No Novelty F1-Score* de 0.650, mostrando um desempenho aleatório tanto na detecção de novidades quanto na identificação de padrões familiares não novidade. Entre as categorias, o melhor desempenho foi registrado em TERROR, com *F1-score* de 0.594. No entanto, categorias como CRIME (*F1-score* de 0.414) e SOCIETY (*F1-score* de 0.435) apresentaram resultados medianos, evidenciando dificuldades do modelo em lidar com dados mais dispersos.

## V. Considerações finais e próximos passos

- A falta de inconformações sobre algumas características desempenhou um papel importante na pesquisa, resultando na exclusão de certas variáveis do conjunto de dados que, em outras circunstâncias, poderiam ser relevantes. Um exemplo disso é a ausência de dados temporais, cuja disponibilidade permitiria que os algoritmos interpretassem novidades não apenas com base no contexto, mas também considerando a dimensão temporal.
- Para tarefas focadas na detecção de novidades, o *Elliptic Envelope* apresentou um desempenho equilibrado, mas ainda insuficiente. A maioria das predições foram aleatórias, evidenciando a dificuldade do algoritmo em identificar padrões consistentes para discriminar documentos novos.
- Para a identificação de não novidade, o *Isolation Forest* demonstrou ser o menos eficaz, apresentando dificuldades em identificar diferenças notáveis entre documentos novos e já conhecidos.
- Finalmente, o *LOF* ao comparar os resultados de nossa proposta com os apresentados no artigo do TAPN, observamos similaridades consideráveis. Ambas as abordagens empregam análise léxica e a métrica de distância euclidiana para a tarefa de detecção de novidade. No entanto, enquanto nosso modelo utiliza um tamanho de n-grama igual a 1, o TAPN adota um tamanho de n-grama igual a 3. Apesar dessa diferença, os resultados obtidos pelo **TAPN** em termos de *F1-score* são valores próximos a 0,66 para a classe *Non-Novelty* e 0,73 para a classe *Novelty*, respectivamente. Embora esses resultados sejam promissores, é importante destacar que a pesquisa do TAPN alcançou um *F1-score Macro avg* menor que **0,7**, enquanto nossa pesquisa chegou a **0,6**. Por outro lado, ao empregar técnicas mais sofisticadas, como modelos de linguagem, o TAPN alcançou valores superiores de *F1-score* de 0,7. Isso indica que ainda há um amplo espaço para aprimorar as técnicas empregadas no nosso algoritmo *LOF*, explorando abordagens mais avançadas como modelos de linguagem.

## VI. Conclusão

Os resultados obtidos neste trabalho demonstraram que a tarefa de detecção de novidade em notícias apresenta desafios significativos, especialmente em contextos onde não há dados temporais e o processamento é baseado exclusivamente em abordagens léxicas como o TF-IDF. Entre os algoritmos avaliados, o *Local Outlier Factor (LOF)* mostrou-se o mais promissor, apresentando o melhor *F1-score* geral e desempenho consistente em categorias estruturadas, como SPORTS. Em contrapartida, o *Isolation Forest* revelou-se limitado, com a maioria das categorias classificadas como não novidade, evidenciando dificuldades na identificação de diferenças relevantes. Já o *Elliptic Envelope*, embora equilibrado, apresentou dificuldades em capturar padrões temáticos claros, resultando em predições próximas ao acaso em muitos casos.

Além disso, a análise temática entre artigos novidade e não novidade revelou diferenças significativas, mas também desafios relacionados à variabilidade e dispersão dos dados em algumas categorias, como CRIME e SOCIETY. Essa análise reforça a importância de explorar abordagens mais robustas, que integrem métodos semânticos e dados temporais, para melhorar a detecção de novidades. Como etapa final, o estudo também destacou a necessidade de avaliar os algoritmos de maneira diferenciada, tanto em todas as categorias quanto em cada uma individualmente, fornecendo insights mais detalhados sobre o desempenho e os limites de cada método. Os resultados obtidos abrem caminho para o refinamento dos modelos e o desenvolvimento de soluções mais precisas e robustas no campo da detecção de novidades.

## Disponibilidade

O código-fonte deste trabalho está disponível publicamente no seguinte repositório: https://gitlab.com/mab0205/introcd2-novelty-detectio. Quaisquer dúvidas ou inconvenientes podem ser comunicados diretamente ao autor.

## References


[1] T. Ghosal, A. Salam, S. Tiwari, A. Ekbal, and P. Bhattacharyya, "TAP-DLND 1.0: A Corpus for Document Level Novelty Detection," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. [Online]. Available: https://aclanthology.org/L18-1559

[2] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation-Based Anomaly Detection," *ACM Trans. Knowl. Discov. Data*, vol. 6, no. 1, pp. 3:1–3:39, Mar. 2012. [Online]. Available: https://doi.org/10.1145/2133360.2133363

[3] M. Ashrafuzzaman, S. Das, A. A. Jillepalli, Y. Chakhchoukh, and F. T. Sheldon, "Elliptic Envelope Based Detection of Stealthy False Data Injection Attacks in Smart Grid Control Systems," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1131–1137, Dec. 2020. doi: 10.1109/SSCI47803.2020.9308523

[4] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, Dallas, Texas, USA, pp. 93–104, May 2000. doi: 10.1145/342009.335388

[5] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, Oct. 2011. [Online]. Available: https://scikit-learn.org/stable/index.html