Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
https://sol.sbc.org.br/journals/index.php/rbie

Submission: dd/Mmm/yyyy; 1st round notif.: dd/Mmm/yyyy; New version: dd/Mmm/yyyy; 2nd round notif.: dd/Mmm/yyyy;
Camera ready: dd/Mmm/yyyy; Edition review: dd/Mmm/yyyy; Available online: dd/Mmm/yyyy; Published: dd/Mmm/yyyy;

Identificação de tecnologias emergentes a partir de dados de patentes usando ML na área da construção civil

Dimitri Prado
Universidade de São Paulo
dimitriprado@usp.br

Luis Henrique Moraes
Universidade de São Paulo
luishmoraes@usp.br

Pedro Oliveira Fernandes
Universidade de São Paulo
pedro.oliveiraf@usp.br

Rodrigo Lyusei Suguimoto
Universidade de São Paulo
rodrigo.lyusei@usp.br

Resumo

A identificação precoce de tecnologias emergentes constitui um desafio para setores consolidados como o da construção civil, onde métodos tradicionais de análise podem ser insuficientes. Este estudo propõe uma metodologia para a identificação automática de tecnologias emergentes no setor da construção civil brasileira, utilizando a análise de dados de patentes brasileiras por meio de algoritmos de aprendizado de máquina (ML). Para isso, primeiro, é realizado a coleta e o pré-processamento de patentes brasileiras da base WIPO PATENTSCOPE. Em segundo lugar, utilizando os dados no intervalo de tempo mais antigo serão usados através de uma abordagem combinada de aprendizado semissupervisionado e ativo para treinar dois algoritmos de classificação distintos, Random Forest e Support Vector Machine, visando uma análise comparativa e a redução de vieses. Após a avaliação dos modelos por meio de métricas como acurácia e F1-score, eles são aplicados a um conjunto de patentes recentes para classificá-las quanto ao seu potencial de emergência. Finalmente, as patentes classificadas são utilizadas para mapear o ecossistema, gerando uma rede de nós baseada em termos-chave que permite identificar tecnologias emergentes e inovações e tendências futuras para o setor.

Palavras-chave: *Construção Civil; Patentes; Tecnologias Emergentes; Machine Learning*

Cite as: Prado, D., Moraes, L. H., Fernandes, P. O. & Suguimoto, R. L. (2025). Identificação de tecnologias emergentes a partir de dados de patentes usando ML na área da construção civil. Revista Brasileira de Informática na Educação, vol, pp-pp. https://doi.org/10.5753/rbie.yyyy.id.

1 Introdução

A inovação tecnológica é um pilar fundamental para a competitividade e o desenvolvimento sustentável em todos os setores da economia. Para indústrias consolidadas, como a da construção civil, a capacidade de identificar tecnologias emergentes de forma antecipada é um diferencial estratégico crucial, permitindo que as empresas se adaptem, lancem novos produtos e otimizem processos existentes (Choi et al., 2021).Também, através da identificação de tendências promissoras, possibilita o direcionamento de investimentos em pesquisa e desenvolvimento para áreas com maior potencial de retorno, garantindo uma vantagem competitiva (Lee et al., 2018).

Contudo, os métodos tradicionais de pesquisa tecnológica, muitas vezes baseados em análises retrospectivas ou na avaliação de especialistas, podem ser insuficientes no cenário atual. Abordagens que dependem de indicadores como citações de patentes, por exemplo, não são prospectivas, pois só podem ser realizadas em estágios avançados do desenvolvimento tecnológico devido ao tempo necessário para que as patentes sejam citadas (Lee et al., 2018). Sendo assim essa defasagem temporal representa um risco significativo para um setor dinâmico como o da construção, que precisa de ferramentas ágeis para o planejamento estratégico. Nesse contexto, a análise de patentes se destaca como uma das abordagens mais eficazes para o monitoramento da inovação, uma vez que as patentes são amplamente aceitas como indicadores da atividade inventiva e uma fonte rica de informações tecnológicas (Kreuchauff & Korzinov, 2017). A vasta cobertura de áreas tecnológicas, países e a acessibilidade dos dados fazem das patentes uma fonte de dados indispensável para investigar tendências (Choi et al., 2021). No entanto, o volume crescente de registros torna a análise manual impraticável, exigindo a adoção de métodos computacionais.

Para superar as limitações das abordagens tradicionais, este estudo propõe e detalha uma metodologia que emprega o aprendizado de máquina (Machine Learning - ML) para a identificação automática de tecnologias emergentes a partir de dados de patentes do setor da construção civil brasileira. Especificamente, o trabalho utiliza uma abordagem híbrida de aprendizado semi supervisionado e ativo, na qual um pequeno conjunto de dados avaliados treina um modelo para classificar um universo muito maior de patentes não rotuladas, otimizando o processo e reduzindo a necessidade de intervenção humana.

O processo metodológico estrutura-se em três etapas principais: a primeira consiste na coleta e pré-processamento dos dados de patentes da base WIPO PATENTSCOPE; a segunda envolve o treinamento e a comparação de dois algoritmos de classificação distintos — Random Forest e Support Vector Machine — para prever o potencial de emergência das tecnologias; e, finalmente, as patentes recentes classificadas como promissoras são utilizadas para mapear o ecossistema tecnológico, identificando inovações e tendências futuras. Com isso, a pesquisa visa oferecer um método sistemático para auxiliar empresas e formuladores de políticas na tomada de decisões estratégicas, fomentando a inovação na construção civil brasileira.

2 Fundamentos Teóricos

2.1 Aprendizado de máquina semi-supervisionado

Os dados utilizados neste estudo originalmente não possuem rótulos, e rotulá-los pode aumentar muito a complexidade do processo. Uma técnica de aprendizado semi-supervisionado pode simplificar esse processo, para que os dados possam ser processados pelos algoritmos de aprendizado de máquina. A utilizada nesse estudo será a de *self-training*, que se baseia em, com base em um pequeno conjunto de dados rotulados, treinar um modelo classificador que seja capaz de gerar pseudo-rótulos para o resto do *dataset*.

O *self-training* pode ser implementado com base em diferentes algoritmos para gerar os classificadores. Um deles é o *random forest*, que utiliza de múltiplas árvores de decisão para garantir uma boa precisão, de forma que o output agregado das árvores é utilizado para definir o rótulo final. Outro é o *support vector machine*, que separa os dados em classes por meio de um hiperplano definido por vetores de suporte. Utilizando o hiperplano encontrado através do treinamento, os resultados da classificação são determinados com base em coordenadas referentes aos valores do vetor de entrada.

2.2 Aprendizado de máquina ativo

O aprendizado de máquina ativo pode ser combinado com o semi-supervisionado para fortalecer a precisão da predição de rótulos. Nessa técnica, podem rotular uma parte dos dados, tal que esta é escolhida por algoritmos. Assim, é possível construir um conjunto de dados com rótulos precisos que vai melhor a performance e o tempo de convergência do modelo.

3 Trabalhos Relacionados

A análise de patentes tornou-se uma fonte importante para rastrear tendências tecnológicas e identificar inovações emergentes. Métodos tradicionais são úteis, mas apresentam limitações, como subjetividade, dificuldade em processar grandes volumes de dados e ambiguidade de palavras-chave (Beatto & Back, 2022; Choi et al., 2021; Ranaei & Suominen, 2017). Nesse cenário, a análise de dados e o aprendizado de máquina (ML) vêm transformando a prospecção tecnológica, permitindo uma transição de análises retrospectivas para abordagens preditivas (Erdogan et al., 2022; Kreuchauff & Korzinov, 2017; Lee et al., 2018; Zhou et al., 2021).

Um estudo fortemente relacionado à nossa pesquisa é o de Choi, Park e Lee (Choi et al., 2021). Os autores propõem uma abordagem híbrida que combina percepções de especialistas com dados de patentes para identificar tecnologias promissoras. Eles utilizam aprendizado semi-supervisionado e ativo, aplicando classificadores como Random Forest (RF), Support Vector Machine (SVM) e Extreme Gradient Boosting (XGBoost) no setor de baterias para veículos elétricos. Os resultados mostram que a incorporação do conhecimento de especialistas melhora significativamente a acurácia das previsões. Nosso estudo converge com esse trabalho, pois também empregamos RF e SVM para a classificação de patentes no setor da construção civil.

Outros trabalhos relevantes reforçam essa direção. Lee et al. (Lee et al., 2018) aplicam redes neurais e RF para detectar tecnologias emergentes por meio de indicadores de patentes. Kwon e Geum (Kwon & Geum, 2020) utilizam SVM e RF para identificar invenções promissoras. Já Mohammadi et al. (Mohammadi et al., 2024) exploram redes de colaboração em patentes para identificar tendências tecnológicas. Em conjunto, esses estudos demonstram como técnicas de ML podem mapear ecossistemas de patentes, identificar tecnologias emergentes e antecipar inovações.

4 Metodologia

A metodologia proposta está estruturada para identificar sistematicamente e analisar tecnologias emergentes no setor da construção civil, utilizando dados de patentes brasileiras e técnicas de aprendizado de máquina. Inspirado na abordagem de Choi (Choi et al., 2021) , o processo é dividido em três fases macro: (1) Coleta e Pré-Processamento dos Dados, (2) Modelagem e Classificação de Patentes, e (3) Mapeamento do Ecossistema de Inovação.

4.1 Coleta e Pré-Processamento de Dados

Os dados brutos de patentes serão coletados da base de dados PATENTSCOPE, mantida pela Organização Mundial da Propriedade Intelectual (WIPO). A busca será realizada por meio de uma query booleana complexa, que combinará palavras-chave e termos técnicos relevantes para o domínio da construção civil. Exemplos de termos incluem: "construção civil", "engenharia civil", "edificação inteligente", "materiais compósitos para construção", "obras de infraestrutura", entre outros.

A busca será direcionada aos campos de título e resumo para garantir a relevância dos documentos. Será aplicado um filtro geográfico para restringir os resultados a patentes depositadas junto a órgãos brasileiros, como o Instituto Nacional da Propriedade Industrial (INPI). Os dados extraídos (autor, ano de depósito, título, resumo, classificações, etc.) serão exportados em formato de planilha para processamento subsequente.

Os dados não são diretamente utilizáveis pelos algoritmos. Portanto, uma fase de engenharia de atributos será conduzida. A partir de colunas como "nº de inventores", "nº de reivindicações"e do conteúdo textual de "título"e "resumo", serão extraídas e calculadas variáveis preditoras. Inspirado no artigo de referência, estas variáveis podem ser agrupadas em categorias como: características da invenção; características do depositante; características de palavras-chave.

A variável-alvo, que define uma tecnologia como "emergente"ou "promissora", será operacionalizada. Uma abordagem comum, como vista no artigo, é usar um proxy quantitativo, como a frequência de citações futuras (forward citations) para patentes mais antigas, definindo o top 10% como "promissoras". Esta definição será a base para o treinamento inicial do modelo.

4.2 Processamento Analítico e Algoritmos

Nesta fase, algoritmos de aprendizado de máquina serão aplicados para classificar as patentes.

4.2.1 Abordagem de Aprendizado Híbrida

Para superar a limitação de ter um pequeno volume de dados previamente classificados (rotulados), será adotada uma abordagem híbrida, combinando aprendizado semi-supervisionado com aprendizado ativo.

- Aprendizado Semi-Supervisionado: Um modelo inicial será treinado com um pequeno subconjunto de patentes mais antigas, cujos rótulos ("promissora"vs "não promissora") são conhecidos. Este modelo será então usado para classificar iterativamente o grande volume de dados não rotulados.
- Aprendizado Ativo: Para refinar o modelo, o processo permitirá a intervenção de um especialista (ou analista). As patentes que o modelo classificar com menor grau de confiança (ex: probabilidade próxima de 50%) serão selecionadas para rotulagem manual, tornando o conjunto de treinamento mais robusto e preciso a cada iteração.

4.2.2 Algoritmos de Classificação e Validação

Para garantir a robustez e mitigar vieses de um único algoritmo, dois modelos de classificação serão desenvolvidos e comparados:

- Random Forest (RF): Um algoritmo de ensemble que combina múltiplas árvores de decisão para melhorar a performance e evitar sobreajuste (overfitting).
- Support Vector Machine (SVM): Um classificador que busca encontrar o hiperplano ótimo que melhor separa as classes de dados no espaço de atributos.

A performance dos modelos será avaliada utilizando métricas padrão. Dada a natureza desbalanceada do dataset (onde patentes "promissoras"são minoria), além da Acurácia, será dada atenção especial ao F1-Score, que representa a média harmônica entre precisão e revocação (precision e recall), sendo mais adequado para tais cenários. O modelo com melhor performance será então aplicado ao conjunto de patentes mais recentes para identificar as tecnologias emergentes.

4.3 Análise dos Resultados

Com as patentes mais recentes devidamente classificadas como "emergentes", será realizada uma análise de co-ocorrência para mapear o ecossistema de inovação. Utilizando os termos-chave extraídos dos títulos e resumos ou os códigos de classificação das patentes (CPC/IPC), será construída uma rede. Nesta rede, os nós podem representar tecnologias (termos-chave), depositantes (empresas/inventores) ou as próprias patentes. E as arestas representarão a força da relação entre os nós (ex: a frequência com que dois termos aparecem juntos nas mesmas patentes promissoras).

A análise da rede permitirá a identificação de clusters (grupos de patentes fortemente conectados), que correspondem aos principais temas tecnológicos emergentes. Além disso, métricas de centralidade na rede ajudarão a identificar os atores-chave (empresas e inventores) que estão liderando a inovação nesses campos. O resultado será um mapa visual do ecossistema futuro, destacando não apenas o quê está emergindo, mas também quem está impulsionando essas inovações, oferecendo uma visão prospectiva e estratégica do setor da construção civil no Brasil.

5 Cronograma

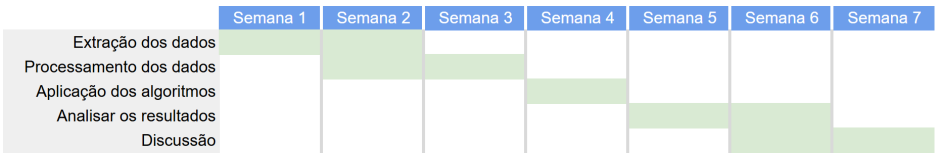

Figura 1: Linha do tempo do projeto.

Referências

Beatto, V. M., & Back, R. B. (2022). Levantamento de patentes tecnológicas que contribuem para a acessibilidade na construção civil. *Revista de Arquitetura IMED*, *11*(1), 151–170. https://doi.org/10.18256/2318-1109.2022.v11i1.4719 [GS Search].

Choi, Y., Park, S., & Lee, S. (2021). Identifying emerging technologies to envision a future innovation ecosystem: A machine learning approach to patent data. *Scientometrics*, *126*(7), 5431–5476. https://doi.org/10.1007/s11192-021-04001-1 [GS Search].

Erdogan, Z., Altuntas, S., & Dereli, T. (2022). Predicting patent quality based on machine learning approach. *IEEE Transactions on Engineering Management*, *71*, 3144–3157. https://doi.org/10.1109/TEM.2022.9904033 [GS Search].

Kreuchauff, F., & Korzinov, V. (2017). A patent search strategy based on machine learning for the emerging field of service robotics. *Scientometrics*, *111*(2), 743–772. https://doi.org/10.1007/s11192-017-2268-3 [GS Search].

Kwon, U., & Geum, Y. (2020). Identification of promising inventions considering the quality of knowledge accumulation: A machine learning approach. *Scientometrics*, *125*(3), 1877–1897. https://doi.org/10.1007/s11192-020-03710-3 [GS Search].

Lee, C., Kwon, O., Kim, M., & Kwon, D. (2018). Early identification of emerging technologies: A machine learning approach using multiple patent indicators. *Technological Forecasting and Social Change*, *127*, 291–303. https://doi.org/10.1016/j.techfore.2017.11.002 [GS Search].

Mohammadi, N., Maghsoudi, M., & Soghi, M. (2024). Innovation ecosystems in retail: Uncovering technological trends and collaboration networks through patent mining. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.10792909 [GS Search].

Ranaei, S., & Suominen, A. (2017). Using machine learning approaches to identify emergence: Case of vehicle related patent data. *2017 Portland International Conference on Management of Engineering and Technology (PICMET)*, 1–8. https://doi.org/10.23919/PICMET.2017.8125290 [GS Search].

Zhou, Y., Dong, F., Liu, Y., & Ran, L. (2021). A deep learning framework to early identify emerging technologies in large-scale outlier patents: An empirical study of CNC machine tool. *Scientometrics*, *126*(2), 969–994. https://doi.org/10.1007/s11192-020-03797-8 [GS Search].