Classificação de gêneros textuais utilizando *Machine Learning*

Enzo Guarnieri, Gustavo V. Mitraud, Júlia C. de Oste

Faculdade de Computação e Informática – Universidade Presbiteriana Mackenzie (UPM) – São Paulo – SP – Brasil

enzo.guarnieri@mackenzista.com.br, gustavo.mitraud@mackenzista.com.br, julia.oste@mackenzista.com.br

Abstract. *This work aims to study the classification of textual genres using Machine Learning techniques. For this purpose, a set of texts representing different genres will be built and preprocessing steps, such as data cleaning and transformation, will be applied. Then, different machine learning algorithms will be analyzed regarding their potential for automatic classification. The dataset used was taken from the American book platform "Goodreads."*

Resumo. *Este trabalho tem como objetivo estudar a classificação de gêneros textuais utilizando técnicas de aprendizado de máquina. Para isso, será construído um conjunto de textos representando diferentes gêneros e aplicadas etapas de pré-processamento, como limpeza e transformação dos dados. Em seguida, diferentes algoritmos de aprendizado de máquina serão analisados quanto ao seu potencial para realizar a classificação automática. O conjunto de dados utilizado foi retirado da loja de livros americana "*Goodreads*".*

1. Introdução

A literatura desempenha um papel central na formação cultural e histórica das sociedades, refletindo valores, estilos e práticas discursivas que organizam em torno de diferentes gêneros textuais, como romance, poesia, conto e drama. Tradicionalmente, a tarefa de identificar o gênero de uma obra literária depende de uma análise manual realizada por especialistas, que levam em consideração critérios linguísticos, formais e semânticos. No entanto, com o avanço da digitalização e a ampliação do acesso a vastos acervos textuais, essa tarefa tornou-se cada vez mais complexa, exigindo apoio de ferramentas tecnológicas para lidar com o volume e a diversidade de dados disponíveis (Stamatatos, Fakotakis & Kokkinakis, 2000).

Nesse cenário, a Inteligência Artificial (IA), em especial as técnicas de Processamento de Linguagem Natural (PLN) e de aprendizado de máquina (do inglês, *Machine Learning — ML)*, oferece ferramentas que permitem a automatização de análise de gêneros literários. Pesquisas demonstram que as abordagens baseadas em frequência lexicais e características estilísticas podem diferenciar gêneros de maneira satisfatória (Stamatatos et al., 2000), enquanto métodos modernos de classificação textual, que vão desde modelos estatísticos até arquiteturas profundas de aprendizado, ampliam as possibilidades de desempenho e generalização (Li et al., 2020).

Diante desse contexto, este projeto propõe o desenvolvimento de um sistema de IA, fundamentado em técnicas de ML, capaz de classificar automaticamente textos de

acordo com seu gênero a partir da sua descrição. A construção desse sistema representa um desafio relevante, dado o caráter interpretativo da linguagem literária, marcada pela ausência de padronização, exploração de ambiguidades, metáforas e jogos de linguagem. Isso exige tanto do leitor humano quanto do sistema de classificação um repertório cultural e histórico mínimo para uma análise mais precisa.

Neste trabalho, optou-se pela abordagem ML/DL/PLN, que consiste em empregar frameworks e ferramentas de ML, *Deep Learning* (DL) e PLN, como *scikit-learn*, para solucionar um problema de classificação. Especificamente, a tarefa consiste em classificar gêneros literários a partir de obras digitalizadas, utilizando bases de dados públicas ou coletadas pelo grupo.

2. Fundamentação teórica

A Classificação de obras literárias é um tema bem embasado na comunidade científica. Ljubešić e Kuzman (2023) fazem uma revisão bibliográfica sobre o tema, apresentando um resumo sobre diferentes abordagens possíveis em cada passo do processo de classificação, como a definição dos *Schemas* de gêneros e coleta do conjunto de dados.

Ademais, Sharof e Lepekhin (2022), discutem a acurácia de diferentes modelos no processo de classificação de gêneros textuais, apresentando, por fim, uma abordagem multi modelo para classificação.

Por outro lado, em trabalhos mais antigos, modelos estatísticos como Análise Discriminante também foram apresentados como uma alternativa à modelos de ML no processo de classificação (KARLGREEN, CUTTING; 1994). Isso mostra a evolução da área, bem como sua consolidação.

Dessa forma, podemos observar que o tema de Classificação de gêneros textuais é um antigo, mas em constante evolução.

3. Descrição do problema

A tarefa de análise e classificação de textos literários é tradicionalmente realizada por profissionais da linguística e literatura, caracterizando-se como um processo subjetivo, manual e de baixa escalabilidade. Esse cenário se torna um problema diante do crescente volume de obras literárias digitalizadas e disponíveis em plataformas online, exigindo métodos mais eficientes para organização e estudo desses materiais.

Nesse contexto, este projeto propõe o desenvolvimento de um sistema de inteligência artificial, baseado em técnicas de ML, com o objetivo de classificar automaticamente textos literários de acordo com o gênero com base na sua descrição.

O principal desafio na construção do sistema de classificação reside no caráter interpretativo da linguagem literária, uma vez que não são textos padronizados ou previsíveis e frequentemente exploram ambiguidades, metáforas e jogos de linguagem. Tais características exigem do leitor e, por consequência, do sistema, um repertório histórico e cultural mínimo para uma interpretação adequada.

Assim, o desenvolvimento da solução demanda o emprego de técnicas sofisticadas de PLN e ML, capazes de lidar com dados textuais complexos e heterogêneos.

4. Ética e responsabilidade no desenvolvimento

O desenvolvimento de sistemas de IA voltados à classificação de gêneros literários demandam atenção a algumas questões éticas, envolvendo tanto questões legais quanto responsabilidades sociais e acadêmicas. Um ponto essencial é o cuidado sobre a curadoria e ao uso dos textos usados, é essencial respeitar os direitos autorais garantindo que as obras utilizadas estejam sob domínio público ou sejam acessadas de forma legal e ética.

Além disso é importante levar em consideração a inclusão de obras e textos de diferentes autores, incluindo minorías étnicas e sociais, de modo a evitar que os modelos treinados possam privilegiar determinadas tradições literárias ou grupos sociais, prevenindo que os modelos reproduzam vieses raciais, de gênero ou sociais.

Por fim é importante manter uma transparência quanto às escolhas realizadas na construção da base de treinamento, permitindo que outros pesquisadores possam compreender possíveis limitações e vieses do sistema.

5. Datasets

Para a exploração inicial do tema, foi utilizado um conjunto de dados disponível na plataforma de código aberto Hugging Face. Essa plataforma é amplamente reconhecida pela vasta coleção de *datasets* prontos para o uso em projetos de ML, além de oferecer uma fácil integração com bibliotecas Python.

Embora o objetivo inicial do projeto fosse trabalhar com dados textuais em português, optou-se por utilizar um conjunto de dados em inglês, devido à maior disponibilidade, variedade e qualidade dos recursos nesse idioma.

O *dataset* escolhido foi o “goodreads-bookgenres” (SZEMRAJ, 2023), que contém cerca de 7 mil dados. Cada entrada inclui informações como o título do livro, uma breve descrição e os gêneros aos quais a obra está associada. Os gêneros disponíveis são:

- História e política;
- Saúde e medicina;
- Mistério e thriller;
- Artes e design;
- Auto-ajuda e bem-estar;
- Não-ficção;
- Ficção científica e fantasia;
- Esportes e recreação;
- Natureza e meio ambiente;
- Negócios e finanças;
- Romance;
- Filosofia e religião;
- Literatura e ficção;
- Ciência e tecnologia;
- Infantojuvenil;
- Culinária;
- Outros.

Este conjunto oferece um cenário rico e diversificado para o treinamento do modelo que será desenvolvido para classificar as obras literárias com base na sua descrição.

6. Metodologia e Resultados Esperados

Serão criados três diferentes modelos baseados nos seguintes algoritmos: *K — Nearest Neighbors* (KNN), Regressão Logística e *Support Vector Machine* (SVM). Dessa forma, o *dataset* será divido em partes de treinamento e de teste, cada modelo será treinado sobre a parte de treinamento e será avaliado baseado nas métricas de avaliação Acurácia e Matriz de confusão.

Espera-se, como resultado, desenvolver algoritmos capazes de reconhecer ao menos um dos gêneros literários presentes em uma obra. Idealmente, os algoritmos também deverão identificar todos os gêneros, caso a obra se enquadre em mais de um.

Referências

JUSSI KARLGREN; CUTTING, D. R. *Recognizing text genres with simple metrics using discriminant analysis*. 5 ago. 1994. Disponível em: https://arxiv.org/abs/cmp-lg/9410008. Acesso em: 20 set. 2025.

KUZMAN, Taja; LJUBEŠIĆ, Nikola. Automatic genre identification: a survey. **Springer Nature Link**, [*s. l.*], 16 set. 2023. DOI /10.1007. Disponível em: https://link.springer.com/article/10.1007/s10579-023-09695-8. Acesso em: 20 set. 2025.

LEPEKHIN, M.; SHAROFF, S. Estimating Confidence of Predictions of Individual Classifiers and Their Ensembles for the Genre Classification Task. **arXiv (Cornell University)**, 1 jan. 2022.

LI, Q.; PENG, H.; LI, J.; XIA, C.; YANG, R.; SUN, L.; YU, P. S.; HE, L. A survey on text classification: from shallow to deep learning. *arXiv preprint arXiv:2008.00364*, 2020. Disponível em: https://arxiv.org/abs/2008.00364. Acesso em: 23 set. 2025

STAMATATOS, E.; FAKOTAKIS, N.; KOKKINAKIS, G. Text genre detection using common word frequencies. In: *Proceedings of the 18th Conference on Computational Linguistics (COLING 2000)*. Saarbrücken: ACL Anthology, 2000. p. 808–814. Disponível em: https://aclanthology.org/C00-2117/. Acesso em 21 set. 2025

SZEMRAJ, Peter. *goodreads-bookgenres*. Hugging Face, 2023. Dataset. Disponível em: https://huggingface.co/datasets/pszemraj/goodreads-bookgenres. Acesso em: 28 set. 2025.