Faculdade de
Computação e Informática

Classificador Inteligente de Notícias de Futebol: Uma Abordagem com BERT para Categorização Automática

Daniel Reis Raske - 10223349

Eduardo Marui de Camargo - 10400734

Victor Vergara Marques de Oliveira - 10403378

Vitor dos Santos Souza - 10204809

João Vitor Tortorello - 10402674

1. Resumo

Este trabalho apresenta o desenvolvimento de um sistema automatizado para categorização de notícias de futebol utilizando técnicas avançadas de Processamento de Linguagem Natural (NLP) e Deep Learning. O sistema, baseado no modelo BERT, foi treinado para classificar notícias em cinco categorias principais: resultado, transferência, lesão, tática e outras. A implementação inclui uma interface web intuitiva e um pipeline completo de processamento de texto em português. Os resultados demonstram o sistema de categorização automática de notícias esportivas, com aplicações práticas para jornalistas, torcedores e plataformas de conteúdo esportivo.

2. Introdução

A crescente quantidade de notícias esportivas disponíveis online torna cada vez mais desafiador para leitores e profissionais do setor acompanharem e organizarem o conteúdo relevante. A necessidade de categorização manual de notícias consome tempo e recursos significativos, especialmente em grandes portais de notícias esportivas. Este projeto propõe uma solução automatizada para este problema, utilizando técnicas avançadas de Processamento de Linguagem Natural (NLP) e Deep Learning.

O objetivo principal deste trabalho é desenvolver um classificador automático de notícias de futebol que possa auxiliar na organização e categorização de conteúdo esportivo. O sistema foi projetado para ser utilizado por jornalistas esportivos, torcedores, analistas de futebol e plataformas de agregação de conteúdo, oferecendo uma ferramenta eficiente para a gestão de informações esportivas.

3. Fundamentação

• Processamento de Linguagem Natural e Deep Learning

O projeto utiliza o modelo BERT (Bidirectional Encoder Representations from Transformers), um dos mais avançados modelos de linguagem disponíveis atualmente. O BERT foi escolhido por sua capacidade de compreender o contexto bidirecional do texto e por sua eficácia comprovada em tarefas de classificação de texto. A versão em português do modelo foi especialmente adaptada para o contexto das notícias esportivas brasileiras.

• Coleta e Preparação dos Dados

O conjunto de dados foi compilado a partir de diversas fontes de notícias esportivas em português, abrangendo um período significativo de publicações. O processo de coleta considerou a diversidade de fontes e a representatividade das diferentes categorias de notícias. Após a coleta, os dados passaram por um rigoroso processo de pré-processamento, incluindo:

- Limpeza e normalização do texto
- Remoção de elementos irrelevantes
- Processamento linguístico avançado
- Balanceamento das categorias

• Arquitetura do Sistema

O sistema implementa uma arquitetura neural profunda baseada no BERT, com adaptações específicas para a classificação de notícias esportivas. A arquitetura inclui:

- Camadas de transformação do BERT
- Camadas de classificação customizadas
- Mecanismos de atenção especializados
- Pipeline de inferência otimizado

4. Implementação

- **Desenvolvimento do Sistema**

O desenvolvimento do sistema seguiu uma abordagem iterativa, com foco na qualidade da classificação e na usabilidade da interface. A implementação incluiu:

- Desenvolvimento do modelo de classificação
- Criação de uma interface web intuitiva
- Implementação do pipeline de processamento
- Otimização do sistema para produção

- **Análise de Desempenho**

O sistema demonstrou resultados promissores na categorização de notícias, com métricas de desempenho significativas. A análise de erros revelou padrões interessantes, como a dificuldade em classificar notícias com múltiplas categorias ou textos muito curtos, sendo esses erros base para melhorar continuamente o sistema. Além de não conseguir receber várias notícias por vez, já que são coletados em tempo real de sites.

- **Aplicação Prática**

A interface web desenvolvida permite que usuários classifiquem notícias em tempo real, oferecendo uma experiência intuitiva e responsiva. O sistema foi testado com diversos tipos de notícias, demonstrando robustez na categorização.

5. Conclusão e discussão

As principais contribuições incluem o desenvolvimento de um classificador eficiente para notícias de futebol, adaptação do BERT para o contexto esportivo brasileiro, criação de uma interface web intuitiva e estabelecimento de um pipeline completo de processamento.

O sistema enfrenta algumas limitações inerentes à natureza da tarefa, como a dificuldade em classificar notícias multitemáticas e a dependência da qualidade do texto de entrada. Além disso, o modelo BERT em português apresenta algumas restrições que podem afetar o desempenho em certos contextos.

6. Referências

1. 1. Souza, F., et al. (2020). Portuguese-BERT: A BERT model for Portuguese language. In Proceedings of the 5th Workshop on Representation Learning for NLP, pages 1-10.

2. Devlin, J., et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT 2019, pages 4171-4186.

3. Fonseca, E., et al. (2021). Evaluating Portuguese Language Models in News Classification Tasks. In Proceedings of the 13th Brazilian Symposium in Information and Human Language Technology, pages 45-52.

4. Silva, R., et al. (2022). Aplicações de Processamento de Linguagem Natural em Notícias Esportivas: Uma Revisão Sistemática. Revista de Inteligência Artificial Aplicada, 15(2), 78-92.

5. Almeida, T. A., et al. (2023). Classificação Automática de Notícias Esportivas: Desafios e Oportunidades. In Anais do XXIV Simpósio Brasileiro de Computação Aplicada, pages 123-135.