UNIVERSIDADE FEDERAL DA PARAÍBA
CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL - CENTRO DE INFORMÁTICA

DISCENTE: VINICIUS VIERI BEZERRA DE LIMA
DOCENTE: YURI DE ALMEIDA MALHEIROS BARBOSA

PROCESSAMENTO DE LINGUAGEM NATURAL: RELATÓRIO DE PROJETO

LIBERAL x CONSERVADOR: CLASSIFICANDO ORIENTAÇÕES IDEOLÓGICAS ATRAVÉS DE ARTIGOS JORNALÍSTICOS EM SUBREDDITS COM USO DE LONG SHORT-TERM MEMORY

Santa Rita, PB
Outubro de 2024

SUMÁRIO

LISTA DE FIGURAS

1. APRESENTAÇÃO DO PROBLEMA

Com o crescente uso de plataformas de mídia social como fontes de debate e divulgação de opiniões, entender a orientação ideológica das informações publicadas online tornou-se uma questão de grande relevância. Redes sociais, como o Reddit, tem se consolidado como espaços de troca de ideias, onde usuários compartilham e discutem uma ampla variedade de tópicos. Dentro dessas plataformas, subcomunidades denominadas subreddits são criadas para abrigar discussões específicas, incluindo debates de cunho político.

Neste contexto, a identificação automática de orientações ideológicas em textos publicados nas redes sociais é um desafio interessante para a área de Processamento de Linguagem Natural (PLN). A classificação de textos políticos como liberais ou conservadores, com base em suas características linguísticas, pode fornecer insights sobre as dinâmicas discursivas presentes nas plataformas, bem como auxiliar na análise de polarizações políticas. No entanto, essa tarefa é complexa devido à diversidade de estilos e vocabulários utilizados pelos usuários, além da presença de textos neutros ou ambíguos.

O presente projeto busca abordar essa problemática utilizando técnicas de PLN para classificar automaticamente a orientação ideológica de textos postados nos subreddits r/Liberal e r/Conservative. A partir da aplicação de um modelo de Long Short-Term Memory (LSTM), será avaliado se é possível detectar diferenças significativas e computáveis entre os textos dessas duas comunidades, proporcionando uma análise automatizada da ideologia presente nos conteúdos.

2. OBJETIVOS

Objetivo Geral

Desenvolver um modelo de Long Short-Term Memory (LSTM) capaz de classificar artigos postados em subreddits políticos como liberais ou conservadores, com base nas características textuais extraídas dos dados.

Objetivos Específicos

- **Pré-processamento e tratamento dos dados**: Unificar os arquivos JSON dos subreddits r/Liberal e r/Conservative, transformando-os em um dataset consolidado no formato Parquet, contendo as colunas de artigos e rótulos. Implementar técnicas de PLN, como a remoção de stopwords e a aplicação de expressões regulares (Regex), para limpar e preparar os textos para o modelo de aprendizado de máquina.

- **Implementação do modelo LSTM**: Desenvolver um modelo de rede neural recorrente baseado em LSTM, adequado para a classificação de sequências textuais.

- **Análise de desempenho**: Avaliar a acurácia do modelo utilizando métricas adequadas, como acurácia e perda (loss), para verificar a eficácia do LSTM na distinção entre textos liberais e conservadores. Comparar os rótulos previstos pelo modelo com os rótulos reais do dataset, buscando padrões de erro e possíveis sobreposições de conteúdo.

3. DADOS UTILIZADOS E PRÉ-PROCESSAMENTO DOS DADOS

Os dados utilizados neste estudo foram extraídos do *Reddit Ideology Database*, disponibilizado por Ravi et al. (2023) na plataforma *Papers with Code*. O conjunto de dados consiste em 45.088 artigos postados nos subreddits r/Liberal e r/Conservative, sendo 22.544 artigos de cada subreddit. Esses subreddits foram escolhidos por representarem comunidades com posicionamentos políticos não tão opostos, o que torna o problema de classificação ideológica relevante pelo desafio.

O dataset fornecido estava originalmente em formato JSON, contendo textos brutos associados a rótulos binários que indicam a orientação ideológica do artigo (0 para liberal e 1 para conservador). Para facilitar o manuseio dos dados e melhorar a eficiência do pipeline de pré-processamento, os arquivos foram convertidos para o formato Parquet, que oferece maior desempenho e compactação.

Pré-processamento dos dados

O pré-processamento dos dados foi uma etapa fundamental para preparar os textos para o treinamento do modelo de aprendizado de máquina. Inicialmente, expressões regulares (Regex) foram aplicadas para remover caracteres indesejados, como símbolos e pontuações que não contribuem para a análise semântica dos textos. Em seguida, stopwords, como artigos, preposições e conjunções, foram removidas, visto que essas palavras não acrescentam valor significativo para a classificação ideológica.

Após a limpeza dos dados, os textos foram tokenizados e vetorizados, processo pelo qual cada palavra foi substituída por um identificador numérico único, permitindo que os artigos fossem convertidos em sequências de números. Para garantir que todas as sequências tivessem o mesmo comprimento, aplicou-se a técnica de padding, preenchendo as sequências menores com zeros, para padronizar o tamanho dos textos. Essa etapa de pré-processamento foi essencial para que os dados pudessem ser adequadamente utilizados pelo modelo Long Short-Term Memory (LSTM), que requer dados estruturados em forma de sequências numéricas de tamanho consistente.

4. METODOLOGIA

Este estudo emprega um modelo de Long Short-Term Memory (LSTM) para a classificação automática de artigos postados nos subreddits r/Liberal e r/Conservative, com o

objetivo de identificar padrões ideológicos presentes nos textos. A metodologia aplicada segue um fluxo que abrange desde o pré-processamento dos dados até a avaliação do modelo.

4.1. Técnica utilizada

O LSTM foi escolhido devido à sua habilidade de capturar dependências temporais em sequências de dados, o que é particularmente útil para análise de texto. Diferentemente de redes neurais tradicionais, as LSTMs possuem mecanismos internos, chamados de gates, que permitem armazenar informações relevantes por períodos mais longos, o que é essencial para lidar com contextos mais amplos em textos políticos. O modelo foi implementado utilizando a biblioteca Keras, rodando sobre o TensorFlow, e configurado para classificar os textos em duas categorias: liberal (0) e conservador (1). As principais camadas do modelo incluíram uma camada de entrada de embedding para representar as palavras, seguida de camadas LSTM e uma camada densa totalmente conectada para a predição final. A função de ativação sigmóide foi aplicada na camada de saída, apropriada para classificação binária. Durante o treinamento, utilizou-se a função de perda binary cross-entropy e o otimizador Adam, devido à sua eficiência em lidar com grandes quantidades de dados e sua capacidade de ajustar os parâmetros do modelo de forma robusta.

4.2. Experimento para avaliar a técnica utilizada

O experimento foi conduzido dividindo-se o dataset em duas partes: 80% dos dados foram utilizados para o treinamento do modelo, enquanto os 20% restantes foram reservados para testes e validação. Essa divisão foi realizada de forma aleatória, garantindo que os conjuntos de treino e teste fossem representativos das duas classes ideológicas. Durante o treinamento, o desempenho do modelo foi monitorado através da acurácia e da função de perda. Para prevenir o sobreajuste, foi incorporada a técnica de dropout, que consiste em desativar aleatoriamente uma fração dos neurônios durante o treinamento, promovendo uma melhor generalização do modelo. O modelo foi treinado em 5 épocas, com lotes de dados (batches) ajustados para maximizar a eficiência computacional e a convergência da função de perda. Ao final do processo de treinamento, o modelo foi avaliado no conjunto de teste, e métricas como acurácia e loss foram calculadas para determinar o desempenho da LSTM na tarefa de classificação ideológica. Abaixo se encontra um exemplo de input com output esperado pelo modelo:

Figura 1: Textos comparativos de rotulagem correta e incorreta do modelo

“paris mayor anne hidalgo criticized fox news during an interview for falsely reporting on muslim "no-go zones" in the city after terror attacks, saying the channel's claims were lies. hidalgo stated that paris planned to sue fox news for the damage caused. although fox issued corrections, hidalgo argued that the misinformation had economic consequences, as it made americans fearful of visiting paris. she accused the network of stigmatizing a portion of the population. additionally, the article briefly mentions planned parenthood leaving a federal program for family planning services.”

Classe verdadeira: 1
Classe predita: 1

“alternate headline great depression concerns are gaining attention, and now even barack obama seems among the worried. according to a cnn poll, nearly half of americans expect a significant economic collapse within the next year. obama’s approval rating has dropped to 48%, and more americans are fearing a potential great depression. the poll indicates that pessimism about the economy has grown since obama took office, with a significant increase in the number of people who believe the country is heading towards a major depression. the drop in approval is particularly notable among republicans and independents. the economy remains a critical issue for voters heading into the 2012 election, with 51% identifying it as their top concern.”

Classe verdadeira: 0
Classe predita: 1

5. RESULTADOS

Partindo do ponto anterior, foi observado que o modelo consegue lidar bem com a maioria dos textos. Contudo, como exemplificado, ele ainda é falho para classificar dados que possuam assuntos difíceis de detectar facilmente como sendo apenas de um lado. O tema de aborto e direitos reprodutivos provou ainda ser uma dificuldade para o LSTM, que analisou o contexto geral e classificou o artigo como liberal, quando o mesmo tinha cunho conservador. No geral, a acurácia do modelo foi satisfatória pela sua simplicidade. Abaixo temos um resultado geral com gráfico de acurácia e matriz de confusão.

Figura 2: Gráficos de acurácia e perda(loss) de treino e validação

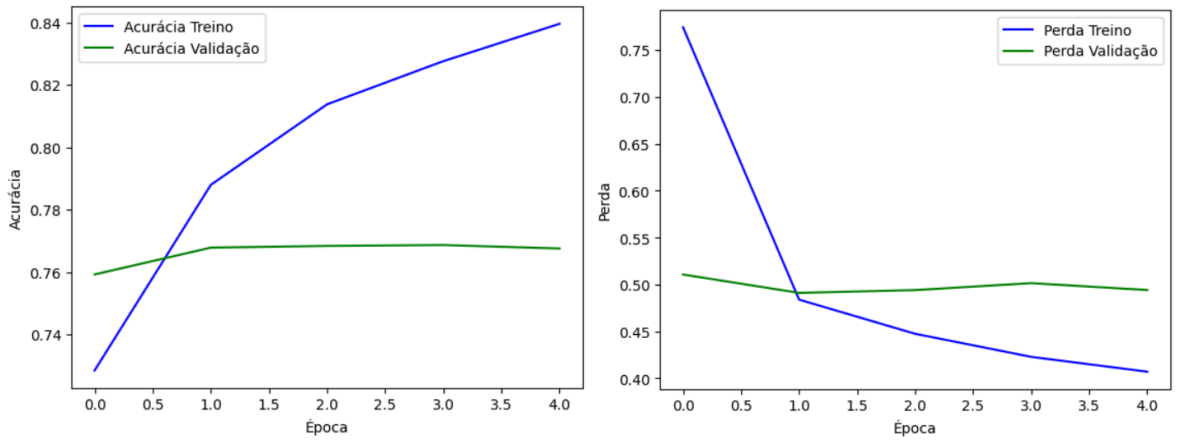
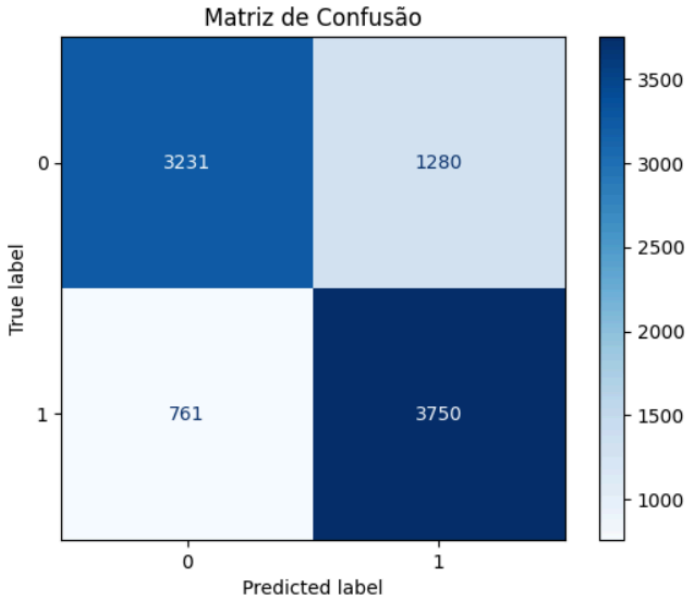


Figura 3: Matriz de confusão do modelo



Como sugestão de melhoria, destaca-se o uso de máquinas de vetor de suporte(SVM) para uma separação através de truque de kernel, o que pode ser eficaz para separar os dados sobrepostos de maneira eficiente através de múltiplas dimensões.

6. REFERÊNCIAS

RAVI, Alok et al. Classifying the ideological orientation of user-submitted texts in social media. Reddit Ideology Database. 2022. Dataset. Licença: CC BY 4.0.