Universidade Federal de Goiás
Instituto de informática
Profa Nádia Félix Felipe da Silva

Relatório (Kaggle - Competição 2)

Aluno: Thiago Achcar Trevisan
Diciplina: Inteligência Computacional

Fevereiro
2023

Universidade Federal de Goiás
Instituto de Informática
Disciplina: Inteligência Computacional

Relatório

Segundo Relatório da participação do Aluno Thiago Achcar Trevisan do Curso Engenharia de Computação da Universidade Federal de Goiás, como requisito parcial para Aprovação da Disciplina Inteligência Computacional.

Aluno: THIAGO ACHCAR TREVISAN

Professor: NADIA FELIX FELIPE DA SILVA

Fevereiro
2023

Conteúdo

1 Resumo

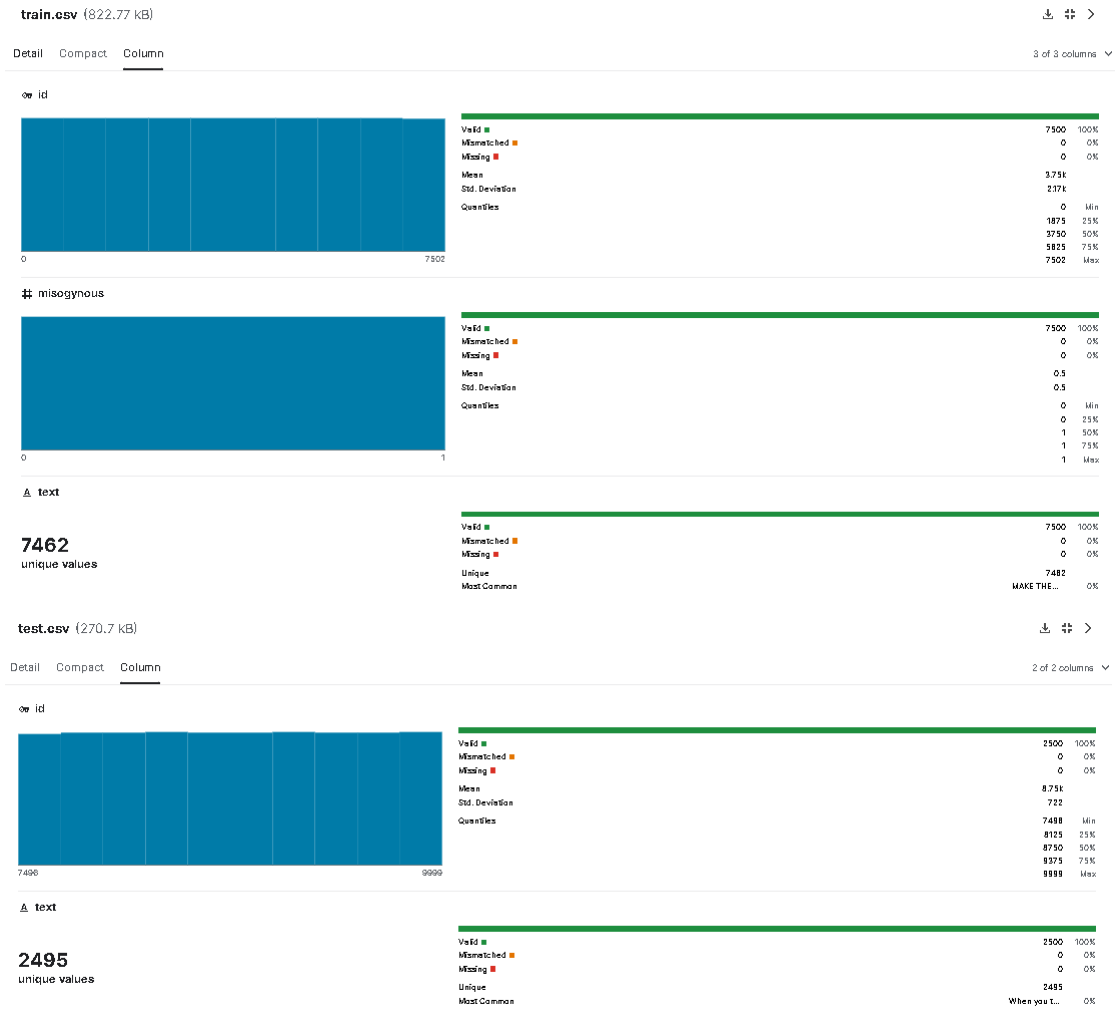
A misoginia, caracterizada pela hostilidade, preconceito e discriminação contra mulheres, é um problema grave em diversas áreas da sociedade, incluindo as redes sociais. A detecção de textos misóginos pode ser uma ferramenta importante para combater esse tipo de comportamento e promover um ambiente mais inclusivo e respeitoso.

O dataset córpus é composto por textos em inglês, coletados de redes sociais e anotados manualmente, estes textos foram classificados como conteúdo misógino (rótulo 1) ou sem presença de misoginia (rótulo 0).

Neste relatório, será apresentado o algoritmo LSTM (Long Short-Term Memory) aplicado à detecção de textos misóginos no dataset especificado. Serão apresentados alguns detalhes sobre a implementação do algoritmo, desde o pré-processamento dos dados até a geração do arquivo de submissão com as previsões.

2 Descrição do Conjunto de dados

O conjunto de dados utilizado neste estudo é composto por um conjunto de treino com 7500 amostras e um conjunto de teste com 2500 amostras, ambos armazenados em arquivos CSV.



3 Descrição de atividades

O modelo de algoritmo LSTM deste trabalho é uma versão adaptada do código 'Classificação de Sequencias com LSTM' presente no slide 'Uma introdução às Recurrent Neural Networks: Visão geral, Implementação, e Aplicação'(Nádia Félix) disponibilizado para o aluno durante o decorrer da disciplina.

Os dados de texto foram convertidos em sequências de números inteiros e foram utilizados os métodos de Tokenizer e pad_sequences para truncação e preenchimento das sequências de entrada.

O modelo LSTM foi implementado utilizando a biblioteca Keras do TensorFlow. O modelo consiste em uma camada de Embedding para transformar a sequência de entrada em uma representação de embedding de palavras, seguida por uma camada LSTM e uma camada Dropout para prevenir overfitting. Por fim, é adicionada uma camada totalmente conectada com um único nó para a previsão. Foi utilizada a loss function binary_crossentropy para a classificação binária, e a métrica accuracy para avaliar a acurácia do modelo.

O modelo foi treinado com três epochs e um batch_size de 64. As previsões foram realizadas no conjunto de teste e convertidas em 1 ou 0 para a criação do arquivo de submissão.

4 Análise dos Resultados

O modelo LSTM apresentou uma acurácia de aproximadamente 87% no conjunto de dados para treinamento. Já o arquivo de submissão gerado apresentou um score de 0.78443 no Kaggle, onde foi submetido.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 2000, 32)          160000

 lstm (LSTM)                 (None, 100)               53200

 dropout (Dropout)           (None, 100)               0

 dense (Dense)               (None, 1)                 101

=================================================================
Total params: 213,301
Trainable params: 213,301
Non-trainable params: 0
_________________________________________________________________
None
Epoch 1/3
118/118 [==============================] - 31s 194ms/step - loss: 0.6119 - accuracy: 0.6611
Epoch 2/3
118/118 [==============================] - 15s 127ms/step - loss: 0.4090 - accuracy: 0.8208
Epoch 3/3
118/118 [==============================] - 11s 96ms/step - loss: 0.3156 - accuracy: 0.8704
79/79 [==============================] - 2s 22ms/step
```

5 Trabalhos Futuros

O algoritmo LSTM mostrou-se ligeiramente eficiente na detecção de textos misóginos, apresentando uma acurácia satisfatória. Podendo ser ajustado posteriormente para obtenção de melhores acurácias dependendo do tamanho do conjunto de dados entre outras especificidades.

Bibliografia

Dyonnatan. (2022). Classificação de texto - Identificação de misoginia.
Kaggle. https://kaggle.com/competitions/competicao-dois-ic.

Brownlee, J. (2019). A Gentle Introduction to the Long Short-Term Memory Network. Machine Learning Mastery.

Nádia Félix. Uma introdução às Recurrent Neural Networks: Visão geral, Implementação, e Aplicação.

https://en.wikipedia.org/wiki/Long_short-term_memory