Classificação de sentimento em reviews de filmes

Gabriel de Oliveira Pontarolo, GRR20203895

1 Introdução

O relatório exibe os resultados dos experimentos na tarefa de classificação de sentimentos em reviews de filmes. Foram feitos dois testes principais, um utilizando modelos clássicos de classificação com o *word2vec* como encoder, e outro utilizando modelos disponibilizados pela plataforma *HuggingFace*.

2 Experimentos

2.1 Word2Vec

Inicialmente, foi feito um experimento com a biblioteca *PyCaret* para obter uma estimativa sobre quais classificadores teriam o melhor desempenho. Para esses testes, foi utilizado *VectorSize* de 300, *WindowSize* de 15 e a média no *Word2Vec*. Os quatro melhores modelos encontrados foram:

Modelo	**Acurácia**	**AUC**	**Recall**	**Precision**
Logistic	0.843	0.917	0.847	0.841
SVM	0.839	0.0	0.849	0.832
LDA	0.838	0.914	0.849	0.832
Ridge	0.832	0.0	0.793	0.861

Com base nisso, foram selecionados os modelos de regressão logística, SVM com kernel linear e análise discriminante linear. Foi feito também o *ensemble* dos modelos, utilizando a estratégia de *Hard Voting*.

Os parâmetros para cada modelo foram escolhidos com base na documentação da biblioteca utilizada na implementação, *SciKit Learn*. Fiz testes variando mais os parâmetros e utilizando a função de *Grid Search*, porém a diferença nos valores era pior ou menor do que 0.01.

A diferença mais significativa foi obtida ao aumentar o *VectorSize* para 500, que subiu em mais de 1% nas quatro métricas avaliadas. Acima disso, as melhorias eram muito pequenas em comparação com o tempo de execução dos modelos e do encoder.

Seguem os melhores resultados obtidos, para o *VectorSize* de 700 e *WindowSize* de 10 :

- **Logistic Regression**
  Penalty: L2
  C: 10.0
  Solver: Newton-Cholesky

	Neg	**Pos**
Neg	10619	1881
Pos	1787	10713

  Accuracy: 0.853
  Precision: 0.851
  Recall: 0.857
  F1: 0.854

- **Linear SVM**
  Kernel: Linear
  C: 10.0
  Penalty: L2

	Neg	**Pos**
Neg	10618	1882
Pos	1799	10701

  Accuracy: 0.853
  Precision: 0.850
  Recall: 0.856
  F1: 0.853

- **Linear Discriminant Analysis**
  Solver: SVD

	Neg	Pos
Neg	10523	1977
Pos	1782	10718

Accuracy: 0.850
Precision: 0.844
Recall: 0.857
F1: 0.851

- **Ensemble**
  Voting: Hard

	Neg	Pos
Neg	10619	1881
Pos	1788	10712

Accuracy: 0.853
Precision: 0.851
Recall: 0.857
F1: 0.854

2.2 Hugging Face

Em seguida, foram feitos testes utilizando modelos grandes pré-treinados disponibilizados pela API da plataforma *HuggingFace.* Cada um dos modelos testados disponibiliza o próprio encoder, então o *Word2Vec* não foi utilizado nestes testes. Em todos eles também foi utilizado o número máximo de tokens em 512, pelas limitações dos modelos. Seguem os resultados dos três melhores modelos encontrados e também o *ensemble* por *hard voting* deles:

- **nlptown/bert-base-multilingual-uncased-sentiment:**

	Neg	Pos
Neg	19330	5455
Pos	1950	23265

Accuracy: 0.852
Precision: 0.810
Recall: 0.923
F1: 0.863

Como a saída desse modelo consiste de uma nota de 1 a 5, uma análise negativa foi considerada como uma nota menor que 3, e positiva como maior ou igual a 3.

- **JamesH/Movie_review_sentiment_analysis_model:**

	Neg	Pos
Neg	23320	1465
Pos	1340	23875

Accuracy: 0.944
Precision: 0.942
Recall: 0.947
F1: 0.945

- **LiYuan/amazon-review-sentiment-analysis:**

	Neg	Pos
Neg	20565	4220
Pos	1930	23285

Accuracy: 0.877
Precision: 0.847
Recall: 0.923
F1: 0.883

Também foi necessário converter a saída desse modelo de uma nota numérica para uma análise positiva/negativa. Foi usado o mesmo threshold do **bert-base.**

- **Ensemble**
  Voting: Hard

	Neg	Pos
Neg	23875	910
Pos	3600	21615

Accuracy: 0.910
Precision: 0.960
Recall: 0.857
F1: 0.906

3 Conclusão

Baseando-se apenas nas métricas avaliadas, para o caso dos modelos clássicos o melhor resultado foi obtido com a regressão logística. Entretanto a diferença para os outros modelos é mínima, sendo mais correto dizer que os três modelos apresentam um desempenho muito similar. É interessante notar também que todos eles tendem a um plateau de 85% nas quatro métricas, mesmo com tentativas de *fine tuning* nos parâmetros.

Para os modelos pré-treinados, é notável que o melhor desempenho foi obtido pelo modelo *JamesH/Movie_review_sentiment_analysis_model* , com quase 95% de acurácia. Mesmo utilizando a técnica de *ensemble*, o resultado obtido foi pior do que utilizando o modelo sozinho.