NitroNLP

Classifying News into Satire and Non-satire

Teodora Diaconescu
Cristina Damov
Iancu Ivasciuc

Task

We had to classify news articles between

Non-satire

Financial Times: SUA, Japonia, Australia și India vor lansa un sistem de monitorizare a pescuitului ilegal în regiunea Indo-Pacific

and

Satire

Lupul singuratic din PNL pleacă cu coada între picioare, ca să nu supere maghiarii din Cluj, după ce a urlat: AaaaaaaU!

Dataset

- Each entry in the dataset consists of a title, content and, a corresponding label
- Unbalanced – the number of satirical news was double that of non-satirical news
- Variable content size – from a paragraph to multiple news sections
- We observed that, in general, entries with a long content were satirical news

Data Preprocessing

- Removed links
- Removed emails
- Removed hashtags / we tried to replace twitter hashtags with keyword 'Persoana'
- Removed stopwords
- Removed punctuation
- Normalized dashes
- Normalized quotation marks

Metrics

$$\text{BalancedAccuracy} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$$

Precision and Recall

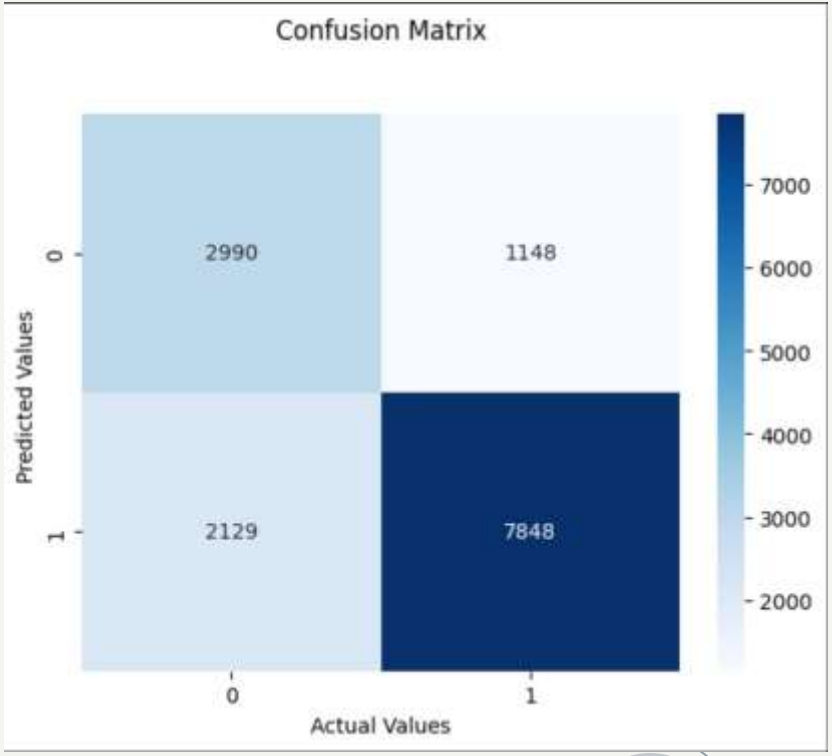
- precision measures the accuracy of positive predictions made by the model

  $$\text{Precision} = \frac{TP}{TP+FP}$$

- recall measures the ability of the model to identify all relevant instances

  $$\text{Recall} = \frac{TP}{TP+FN}$$

Confusion Matrix



BERT (Bidirectional Encoder Representations from Transformers)

- Pretrained: deep learning based on context from a vast corpus of text
- Bidirectionality: the unique ability to understand word context by considering both preceding and succeeding text
- Aplicability: state of the art performance on NLP tasks such as text classification, question answering or machine translation

Transfer Learning

Transfer learning allows the adaptation or reuse of a network model that has been trained for a specific task using a very large dataset to perform a new, related task for which only a small datasets available

Models

- Pretrained Bert, no data preprocessing - 65%
- Pretrained Bert, only titles, no data preprocessing - 81%
- SVM, only titles, data preprocessing - 89%
- SVM, no data preprocessing - 90%

References

1. Text Classification with BERT in PyTorch
2. RoBERT