4N – NEGATIVE NEWS NEURAL NETS

Dan Pavlovič
Darya Pisetskaya
Project Owner

- Kristjan Roosild (Transfer Wise)
Project Description - Problem

- Financial organizations do compliance investigation on customers
- They look for adverse media
- Outsource or in-house
- **Manual check takes a long time**
Project Description - Solution

- Detect if article is adverse media or not
  - Adverse media is crime or crime suspicion news/article about a person or company (corruption, tax evasion, bribery, connection to terrorism, …)
- **Binary classification**
Data

- For training:
  - 801 adverse media (AM) – label 1
  - 397 non-adverse media (NAM) – label 0
  - 396 random – label 0

- For validation:
  - 97 AM
  - 62 NAM/random

- For final test result (client data):
  - Around 100 AM, 50 NAM, 50 random
Model Path

- Naïve Bayes
- **Logistic regression**
  - Word vectorizers
  - Solvers
  - Ensembles
- BERT
- **RoBERTa**
  - Hard voting over windows
  - Soft voting over windows
- **Ensembles**
  - Logistic regression + RoBERTa-All
  - Logistic regression + RoBERTa-Title + RoBERTa-Body
Logistic Regression (LogReg)

- Word Vectorizer
  - Doc2Vec
  - CountVectorizer
  - TfidfVectorizer
    - n-gram = 1
- Solver
  - Liblinear
  - Saga
  - Lbfgs
- Hyperparameter testing
  - Penalty = L2
  - C = 17 (regularization strength)
- Ensembles
  - BaggingClassifier
  - AdaBoostClassifier

\[ \text{sig}(t) = \frac{1}{1 + e^{-t}} \]
Logistic Regression (LogReg) - Results

Accuracy: 0.854, 0.886, 0.937, 0.936
F1: 0.853, 0.884, 0.938, 0.937
Matthews correlation coefficient: 0.709, 0.772, 0.874, 0.872
RoBERTa (RB)

- Improved BERT
- State of the art NLP deep learning model
  - Pre-trained on huge data set
- Requires only fine tuning
- Version: Base (125M params)
- Hyperparameters
  - Learning rate
  - Epoch
- Sliding windows (20% overlap)
  - Hard vote over windows
  - Soft vote over windows

Note: RoBERTa does not pre-train with next sentence prediction (NSP) as shown in picture (only BERT does)
RoBERTa (RB) and BERT

Accuracy

F1

Matthews correlation coefficient

BERT - All  
RB - Title  
RB - TitlePlus  
RB - Body  
RB - All  
RB - All (Soft)
Ensembles (ENS)

Analysis showed models were not making same mistakes

- Ensemble-2 (ENS-2):
  - Logistic regression – All + RoBERTa – All
  - Soft Voting
    - Probabilities
    - Weighted (MCC scores)

- Ensemble-3 (ENS-3):
  - Logistic regression – All + RoBERTa – Title + RoBERTa – Body
  - Soft Voting
    - Probabilities
    - Weighted (MCC scores)
Results – Cross Validation (6 Fold)

Accuracy
- LogReg: 0.936
- RB - All (Hard): 0.95
- RB - All (Soft): 0.957
- ENS-2: 0.96
- ENS-3: 0.958

F1
- LogReg: 0.937
- RB - All (Hard): 0.95
- RB - All (Soft): 0.958
- ENS-2: 0.961
- ENS-3: 0.958

Matthews correlation coefficient
- LogReg: 0.872
- RB - All (Hard): 0.915
- RB - All (Soft): 0.915
- ENS-2: 0.92
- ENS-3: 0.92
Results – Test data

Accuracy

F1

Matthews correlation coefficient

- LogReg
- RB - All (Hard)
- RB - All (Soft)
- ENS-2
- ENS-3
Time Performance

<table>
<thead>
<tr>
<th></th>
<th>LogReg</th>
<th>RB - All (Hard)</th>
<th>RB - All (Soft)</th>
<th>ENS-2</th>
<th>ENS-3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Execution Time</strong></td>
<td>68 ms</td>
<td>9.21 s</td>
<td>9.21 s</td>
<td>9.48 s</td>
<td>12.1 s</td>
</tr>
<tr>
<td><strong>CV-6 Accuracy</strong></td>
<td>93.6%</td>
<td>95%</td>
<td>95.7%</td>
<td>95.7%</td>
<td>96%</td>
</tr>
</tbody>
</table>

Run on test set (159 articles)

Executed in Google Colab

- Logistic regression is significantly faster
- RoBERTa (base) needs around 10GB of GPU memory
Ways to improve

◦ More data better results
  ◦ When gathering additional data we saw improvements for Logistic regression and RoBERTa

◦ Bigger batch size (less overfitting)

◦ RoBERTa large (requires a lot of GPU memory)
Github Repository

https://github.com/dannoc96/4N-Negative-News-Neural-Nets---ML-Project