NLP Recruitment Task

Aleksey Klimchenko

Task’s Objective:

Classify whether or not a news piece describes money laundering activity.

The secondary objective of classifying the type of money laundering activity described was not carried out.

Defining Money Laundering Activity

Determining if ‘money laundering activity’ is being described can be subjective

- Some articles discuss the concept of money laundering, but don’t directly describe any activity
- Other articles reference money laundering activity, but it is not the main subject of that article

Defining Money Laundering Activity

The broad scope of ‘money laundering activity’ had to be narrowed down so that the target variable was well defined:

- News articles were classified as describing “money laundering activity” if the articles’ contents directly discussed a party’s participation in money laundering activity.
- Did not include the discussion of ‘money laundering’ as a whole, or adjacent methods used to launder money by a party

Task Constraints

Scraping articles from new sources was considered, but ultimately decided against

- If issues with the scraping arose, it could become a time sink
- Each scraped article must be vetted to determine whether money laundering activity is described. This can be done while manually collecting news articles
- Articles not describing money laundering activity could be taken from the kaggle BBC News Classification competition's training dataset

The secondary objective of classifying the type of money laundering activity described was not be carried out due to time constraints.

Overview

1. Data Collection, Cleaning, and EDA
2. Models & Results
3. Discussion & Conclusion
4. Future Work

Data Collection

Articles not describing money laundering were randomly sampled from the kaggle BBC News Classification's training dataset

- None of these articles directly described money laundering activity
- Several articles that did not describe money laundering activity, but mentioned the topic or a related activity, were included to prevent bias towards key words

Articles of interest were taken from the BBC website (with a handful of exceptions)

50 articles that described money laundering activity were collected.

Data Cleaning and EDA

Each article was processed to prepare it for modeling:

1. Text made lowercase
2. Non-alphanumeric characters removed
3. All text was tokenized
4. Repeating words removed
5. Stop words removed

Both pre- and post-processing text lengths followed similar distributions.

50 ‘target’ articles were joined with 150 ‘non-target’ articles, resulting in a dataset of 200 articles to be used for model training and testing.

Models & Results

Scikit-Learn Models

The following models were created to classify articles:

- Random Forest
- Logistic Regression
- KNeighborsClassifier
- Decision Tree
- Gaussing NB

Each model had parameters tuned via RandomizedSearch (decision tree) or GridSearch (the rest) using 5-fold cross validation.

Scikit-Learn Model Results

Model	Accuracy	Precision (Target)	Recall (Target)	F1-Score (Target)
Random Forest	0.88	0.7	0.7	0.7
Logistic Regression	0.84	0.62	0.5	0.56
KNN Classifier	0.86	0.67	0.6	0.63
Decision Tree	0.84	0.62	0.5	0.5
Gaussian Naive Bayes	0.88	0.67	0.8	0.73

Neural Network Models & Results

A NN model was created with one or more of these layers:

1. Dense
2. LSTM
3. Conv1D

After 5 epochs, each models' accuracy was 0.8 .

Analyzing Results

The scikit learn models seemed to perform well, achieving accuracies above 0.8 .

However, the lower precision and recall scores show that these models could use some improvement.

RF and GNB models each had an accuracy of 0.88, but the higher f1-score of GNB indicates it is the better model.

The NN models achieved an accuracy of 0.8, but were also overfitting the data.

Conclusion & Discussion

Conclusion

Several models were built for classifying whether a news article describes money laundering activity, thereby achieving the task’s objective.

Discussion

The greatest limitation for this task was the time constraint, which greatly impacted the size of the dataset collected.

Each scikit learn model would likely benefit from a larger dataset, if at least to improve precision and recall.

The NN models' issue of overfitting would also be resolved with a larger dataset.

Confidence in model performance could be improved through cross validation using each model's hypertuned parameters.

Future Work

The secondary objective asked to classify the money laundering activity described in each article into one or more of several categories:

- Allegations
- Accusations
- Charges
- Conviction
- Sentencing

To build models that can do this accurately would require an enormous dataset, for which time was not available. A potential approach would be to combine several categories into one.

Future Work

Synthetic data generation techniques could have been employed to create a larger dataset (if time had permit):

- Back translation
- Synonym Replacement
- Random Insertion/Swap/Deletion
- Albumentation
  - Shuffle Sentences Transform
  - Exclude Duplicate Transform

Sources and Acknowledgements

Article sources:

- https://www.kaggle.com/c/learn-ai-bbc/data
- https://www.bbc.com/

Acknowledgements:

- https://github.com/DiveshRKubal/Data-Science-Use-Cases/tree/master/News%20Classification
- https://www.kaggle.com/tyan001/nlp-classification-news-aggregator