Reuters-21578 Articles topics classification Model

Kedir Ahmed

Who is Kedir

- Kedir is passionated Machine Learning Engineer
- Interested to work in Machine Learning and Data Science projects.
- Has a Masters degree in Computer Science from Ca'Foscari University, Vence Italy
- Has experience working as a Software Engineer.
- Skilled in skilled in Machine Learning, Deep Learning, Statistical Modeling, Data Analysis, and Data Visualization.

Introduction

Background: The Reuters-21578 dataset is one of the most commonly used collections for text classification problems. In this project we work with 21 SGML files from Reuters-21578 dataset, each containing 1000 documents in it.

The files were taken from the Intros ML Coding challenge Data.

Objectives: The main objective of this project is to be able to classify the topics for a given article from Reuters-21578 dataset as **EARN** or **NOT-EARN** class.

In order to tackle the problem, the proposed approach is to use a class classification model in which the 1 class represents the EARN topics and the 0 class represents OTHERS.

Specific objectives

Data wrangling:

- Parse and transform raw data in SGML file format into a pandas dataframe.

Exploratory Data Analysis

- Draw insights from the dataset

Modeling:

- Generate text embedding
- Split the dataset in to train and test
- Train a Single hidden layer feed-forward Neural Network

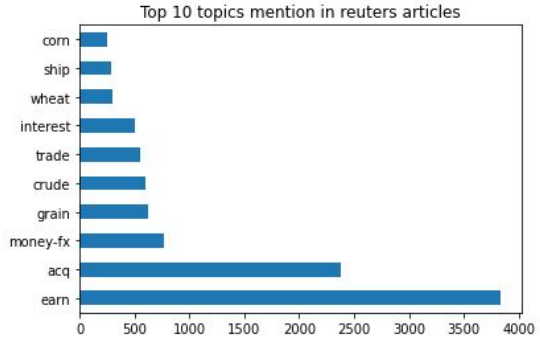
Evaluate the model

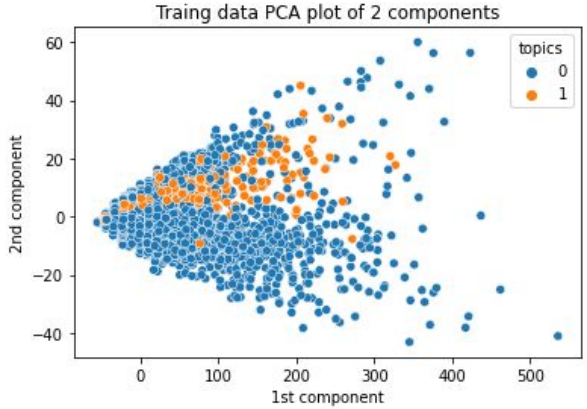
- Apply different evaluation metrics and score the performance of the model.

Data Wrangling

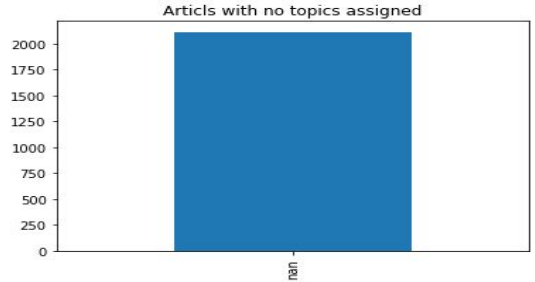
- Parse the raw data in SGML file format into a pandas dataframe.
    - Tools used:
        - Beautiful soap to extract the features of the articles
        - Pandas library to create a pandas dataframe
    - There are two topics tags in the SGML data where the first one tells whether the document has a topic and the other has the topic name. The first one has been renamed as topic_bool to differentiate them.
- Clean the data for unwanted special characters, html tags, extra spaces and so on.

Exploratory Data Analysis (EDA)





- From the above chart you can see that most of the reuter's articles focus on the topics:
  1. EARN(Earnings and Earnings Forecasts)
  2. ACQ(Mergers/Acquisitions) topics.



From the above graph we can see that there are around 2000 documents where it says it has a topic but there is no topic assigned to these documents.

EDA Insights

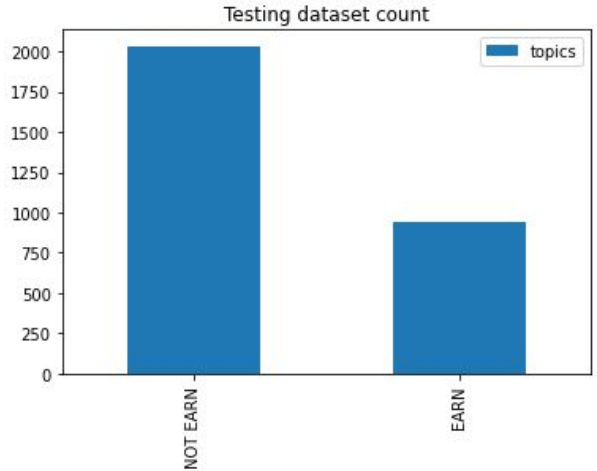
- The above EDA shows us the majority of the documents are categorized under EARN topics which can tell that it is even possible to have a binary class EARN or Others. The main target here is to correctly classify documents which belong to the earn class.
- There are documents where they are supposed to have a topic but they don't. we need to remove this document from the training data. The reverse also happens but we keep the document because they have a topic.
- All documents which belong to the EARN class don't have a mention of Organization. Hence such attributes aren't useful for our target.
- There are articles with a topic but the text feature is not a valid english sentence, e.g. “Bla bla bla …” has been assigned a topic. So, in order to not loss the information we used the title feature in addition to the text feature for better representation of the data.

Modeling

- Textual data embedding
    - Generate text embedding for **text** and **title** features of the documents.
    - The text feature is a collection of paragraphs and the following steps has been done to generate embedding vector for the text.
        1. Apply NLTK paragraph tokenizer and get list of sentences
        2. Pass the sentences through BERT text embedding model and get the vector representation of each sentences of length (1,768).
        3. Sum the vectors of all sentences and represent the text by the sum vector.
    - The same has been done for the title feature.
    - Generated final embedding for a given article by concatenating the title vector and the text vector.
    - The target variable, the topics feature has been binarized in to 1 and 0 where 1 represents EARN class and 0 represents OTHERS.

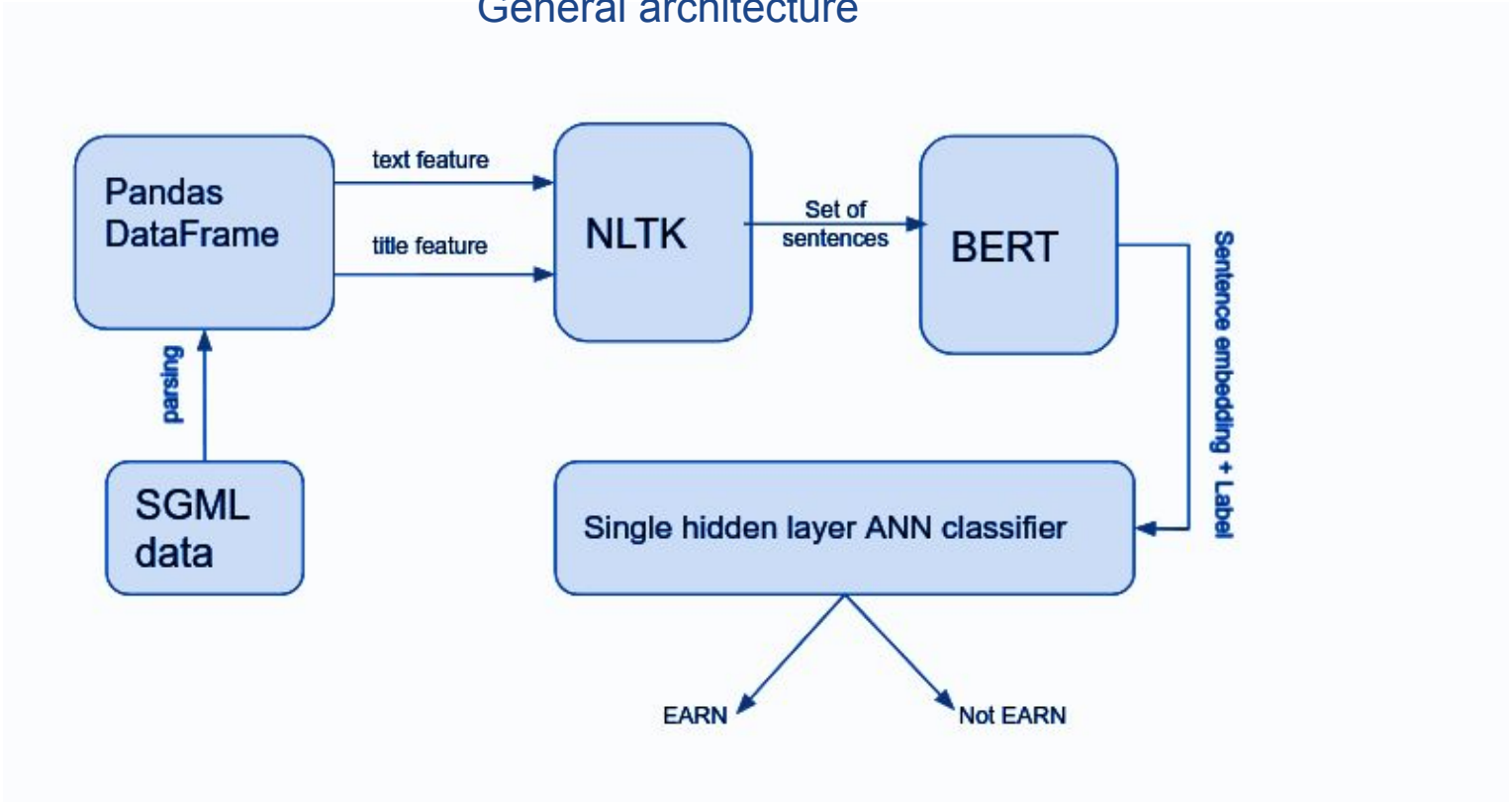
Dataset split

- Based on the recommended Train Test split approach (from the Readme of the Reuters-21578), the Modified Apte split method has been applied.
  - Why Modified Apte split?
    We require each training document to have at least a topics category.



Train a Single hidden layer feed-forward Neural Network

General architecture



Evaluate the model

- The evaluation metrics used for scoring the performance of the model are:
  - Precision and Recall
  - ROC AUC curve
  - Balanced Accuracy

  Why?
- Precision and Recall from the basic confusion metrics could tell the overall performance.
- As is is a Binary class classification problem, we can get benefit of Balanced accuracy because It takes into account the accuracy of both classes.
  In addition:
  ROC AUC curve is not affected by class imbalance.

- Testing the model on the testing dataset results

precision_0	precision_1	recall_0	recall_1	auc	balanced_accuracy
0.992982	0.947639	0.974902	0.985059	0.97998	0.97998

Conclusion

- The proposed approach classifies a given document as EARN and NOT EARN with score of above 95% for all evaluation metrics used.

- As a future work, I would recommend to use a multilabel classification approach which takes more topics in to account.