CS 614 – Applications of ML

Project 3 – Natural Language Processing (Multiclass Classification)

Submitted by: Manil Shrestha

Project: Reuters News Classification

1. Pitch:

State the circumstances an organization or sets of users would be willing to fund the proposed ML application based on their perceived value. Value could be financial, or any other type of outcome organizations or users need to obtain (e.g., institutional image, customer satisfaction, increased quality, or efficiency).

Organizations such as news agencies, content management platforms, or digital marketing firms could be interested in funding a multiclass news classification ML application, as it could streamline their content organization, improve recommendation algorithms, and enhance user experience by providing tailored content. Additionally, this ML application could potentially increase efficiency by automating manual content categorization tasks, thus freeing up human resources for other value-added activities.

2. Data source:

Indicate with a link where you obtained the data. If you generated the data yourself, please provide a link to the code of the used approach.

The Reuters-21578 text categorization dataset is a widely recognized collection of documents that initially appeared on the Reuters newswire in 1987. Compiled and indexed with categories by personnel from Reuters Ltd. and the Carnegie Group, Inc., this dataset is easily accessible to machine learning enthusiasts via the NLTK library.

3. Model and data justification:

Justify why you chose a specific model to learn from the selected data. If you learned that a given more would be suitable for a type of data from a publication, please provide a link to the source. Note you still have to justify it with your own words (as always, limit to three sentences).

The initial step involved preprocessing the data: tokenizing the news using nltk.word_tokenize, extracting TF-IDF features from these tokens, and applying Singular Value Decomposition (SVD) to reduce the feature set from 35658 to 5000, while preserving 95% of data variance. Following this, a fully connected neural network comprising two hidden layers activated by a ReLU function was trained using stochastic gradient descent and cross entropy loss. While the model's performance was subpar with the initial 35k+ features, the application of dimensionality reduction led to a light enhancement in results; additionally, I utilized a simpler sklearn model, specifically the OneVsRestClassifier and Linear SVC, to further evaluate the model.

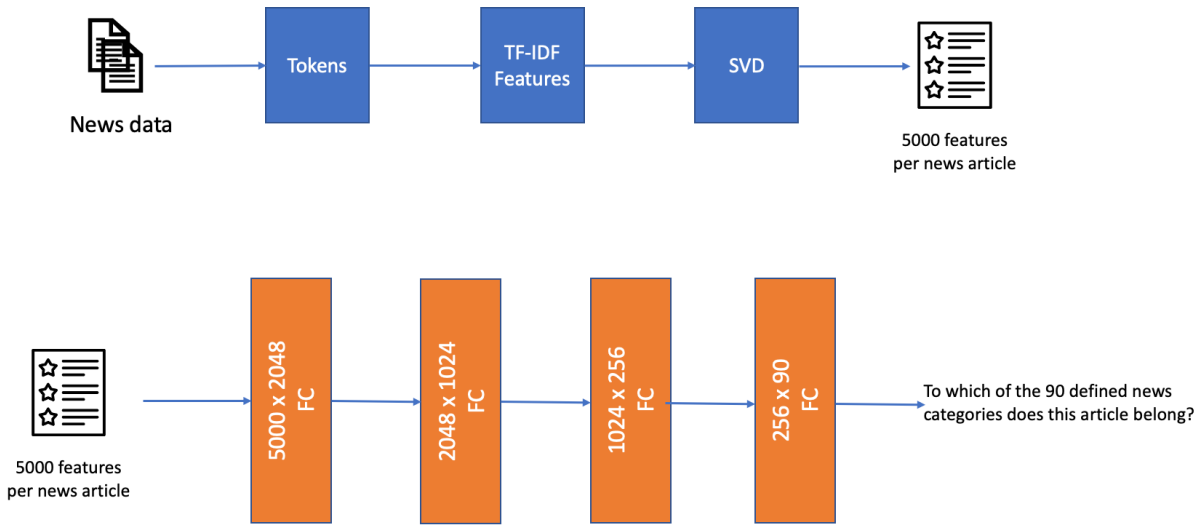

Figure 1: Preprocessing and NN architecture

4. Commented examples:

Indicate the input where trained model is applied, the output and whether it is as expected or any observations you may have.

The following commented example is the result of OneVsRestClassifier:

```
index=98

print(train_documents[index])

tfidf_vectorised_train_documents = tfidfvectorizer.fit_transform(train_documents)
predictions = classifier.predict(tfidf_vectorised_train_documents)

mlb.inverse_transform(np.array([predictions[index]]))
```

```
N.Y. BANK DISCOUNT BORROWINGS NIL IN WEEK
  The eight major New York City banks
  did not borrow from the Federal Reserve in the week ended
  Wednesday March 25, a Fed spokesman said.
      It was the second half of a two-week bank statement period
  that ended on Wednesday. The banks did not borrow in the first
  week of the period.
```

```
[('interest', 'money-supply')]
```

Figure 2 Commented example demonstrating the classification of news to category

In this demonstration, I processed a single sample (index 98) from the training dataset and ran it through the classifier. The prediction was initially in a multi-label binary format, which I then converted to a more human-readable form using inverse transformation. The news article, centered around 'bank discount,' aligns well with the predicted category, confirming the effectiveness of our model.

5. Testing:

Provide a confusion matrix with one or more metrics and comment the results.

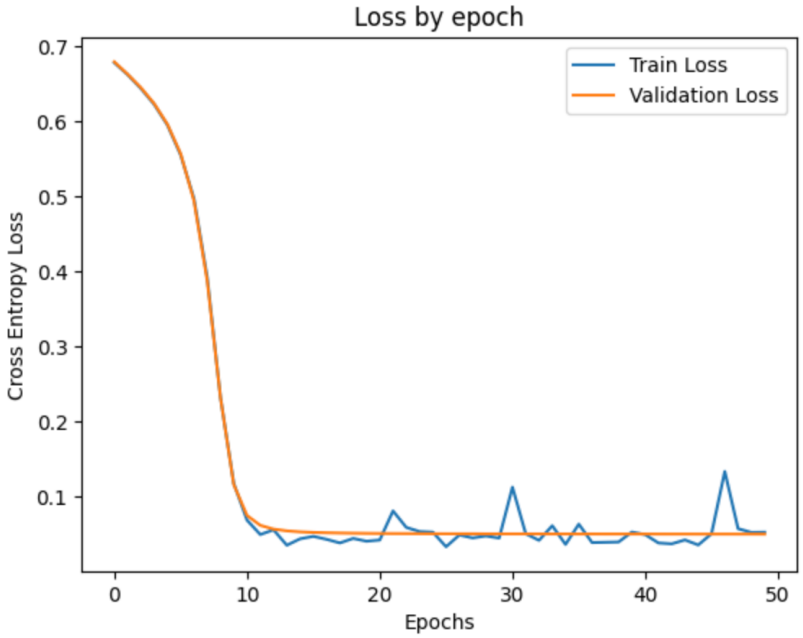

Figure 3: Cross entropy loss by training epoch

Metric	Macro	Micro
Accuracy	0.8099	-
Precision	0.6074	0.8467
Recall	0.3702	0.7970
F1-measure	0.4405	0.8211

We can see that macro recall is significantly low, which means the model is not able to correctly identify a substantial number of actual positive instances of the class. In multiclass problem, it generally happens because the model may struggle to learn patterns for the under-represented classes, resulting in low recall. The low recall is impacting the low Macro F1 as well.

6. Code and instructions to run it:

Provide a link to the code and any required instructions to run it. Please include some testing examples so we can quickly experience what you experienced with the model.

Here is the link to Jupyter Notebook which has been uploaded in the github:
https://github.com/ManilShrestha/AppliedMLProjects/blob/main/Reuters-NewsClassification.ipynb

You can recreate the training and experiments by following the steps:

1. Make sure all the dependencies are installed in your machine.
2. Run all the cells in the notebook, since data is downloaded via NLTK library, no need to specify data directories.