Candidate: Graham Vo

Assignment:

- Build a text classification model for news articles using Python, machine learning, and NLP.

- Instructions: Use the Reuters news dataset (https://archive.ics.uci.edu/dataset/137/reuters+21578+text+categorization+collection) to train and test your model.
- Preprocess the data by removing stop words, stemming, and tokenizing the text.
- Use a machine learning algorithm of your choice to train a text classification model.
- Evaluate the performance of your model using appropriate metrics such as accuracy, precision, recall, and F1 score.
- Write a brief report summarizing your approach including how you did preprocessing, EDA, evaluating the performance of your model, and any insights you gained from the analysis.

Brief report:

Reuters Text Classification Report

1. Introduction

The report presents a comprehensive approach for text classification of Reuters articles. The objective of the analysis was to predict the topic of an article based on its content. A machine learning model, specifically the XGBoost algorithm, was used for this task.

2. Data extraction

The initial data was in SGM format and was parsed using BeautifulSoup to extract relevant features. These features included 'TOPICS', 'LEWISSPLIT', 'CGISPLIT', 'OLDID', 'NEWID', 'DATE', 'UNKNOWN', 'PLACES', 'PEOPLE', 'ORGS', 'EXCHANGES', 'COMPANIES', 'TITLE', 'DATELINE', and 'BODY'. The extracted data was loaded into a Pandas DataFrame.

Feature	Description
TOPICS	Yes/no feature indicating whether the article has topics
LEWISSPLIT	Indicates whether data was for training, testing, or not used
CGISPLIT	CGI split label
OLDID	Old ID
NEWID	New ID
DATE	Article's date
UNKNOWN	Unknown information
PLACES	Places mentioned in the article
PEOPLE	People mentioned in the article
ORGS	Organizations mentioned in the article
EXCHANGES	Exchanges mentioned in the article
COMPANIES	Companies mentioned in the article
TITLE	Article's title
DATELINE	Article's dateline
BODY	Article's body

Certainly, here is a detailed breakdown of the steps:

- The first step of the process is to extract the data from the SGM files. This is done using the BeautifulSoup library which is designed to parse HTML and XML documents.
- Extract the data from a single SGM file then store the extracted data. Load them to dataframe.
- Test results: compare data from dataframe and actual sgm file, after testing and modified the code, here is the resulting DataFrame:

Column	Count	Unique Values	Top Value	Frequency of Top Value
TOPICS	21578	3	'YES'	13476
TOPIC_NAME	21578	656	''	10211
LEWISSPLIT	21578	3	'TRAIN'	14668
CGISPLIT	21578	2	'TRAINING-SET'	20856
OLDID	21578	21578	'8914'	1
NEWID	21578	21578	'4001'	1
DATE	21578	21578	'11-MAR-1987 18:04:17.59'	1
UNKNOWN	21578	21577	' \n^! 16#838#'	2
PLACES	21578	1097	'usa'	10878
PEOPLE	21578	218	''	20422
ORGS	21578	68	''	20697
EXCHANGES	21578	58	''	21096
COMPANIES	21578	1	''	21578
TITLE	21578	20030	''	737
DATELINE	21578	8823	''	2535
BODY	21578	18782	''	2535

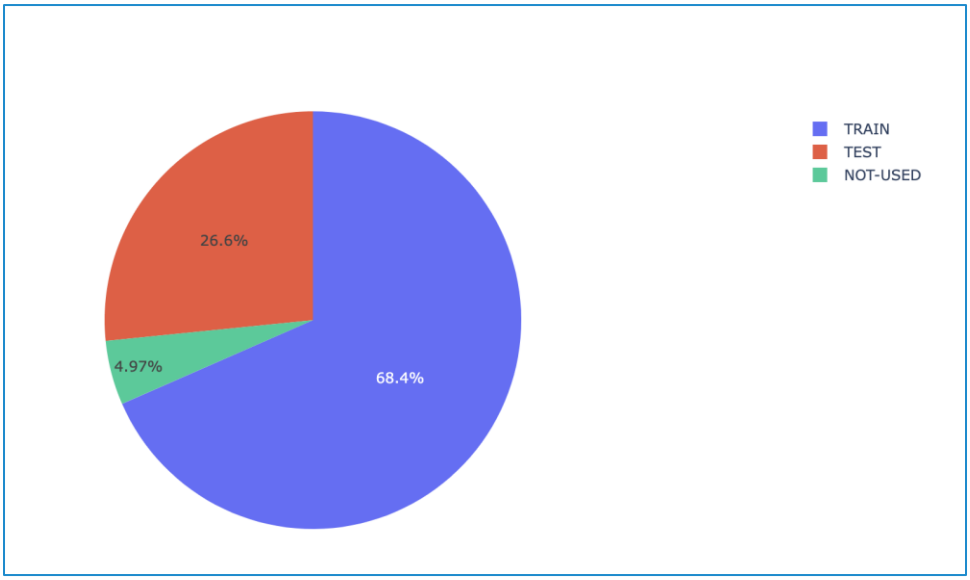
Meanwhile, I have checked Company feature, it contains empty character only, so it will be removed.

3. Exploratory Data Analysis (EDA)

The EDA involved analyzing the distribution of the 'LEWISSPLIT' and 'TOPIC_NAME' variables. 'LEWISSPLIT' indicated whether the data was used for training, testing, or not used, while 'TOPIC_NAME' was the target variable. The number of articles by topic was plotted, and only topics with over 50 samples were included to ensure a balanced dataset. The resulting DataFrame contained 3697 training set instances, 1409 testing set instances, and 312 instances that were not used.

- Splitting dataset into Train/Test/Not-used

LEWISSPLIT	Count
TRAIN	3697
TEST	1409
NOT-USED	312

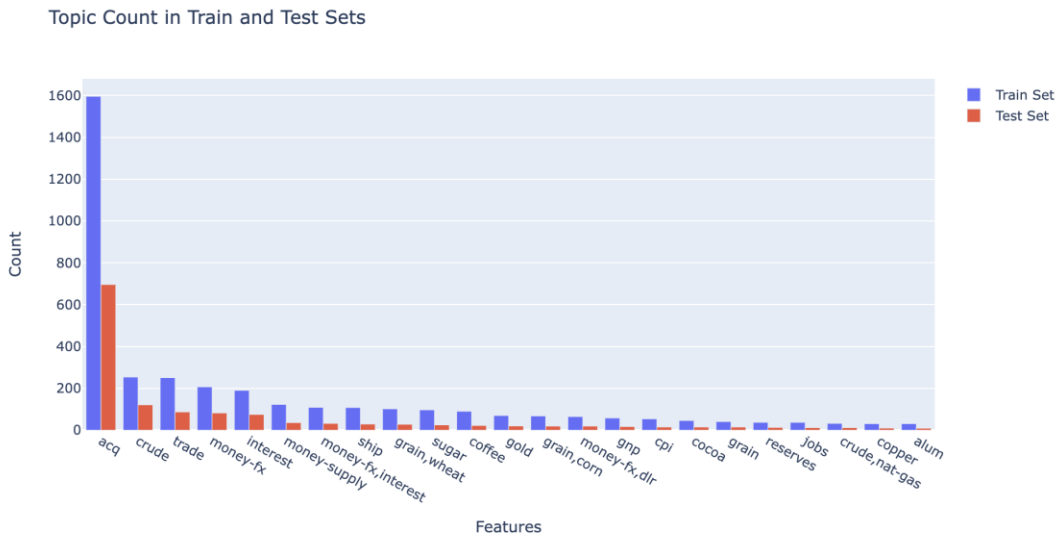


Dataset split

- Top topics with sample counts:

	topic_name	value_count
1	acq	2362
2	crude	408
3	trade	361
4	money-fx	307
5	interest	285
6	money-supply	161
7	ship	158
8	grain,wheat	148
9	sugar	143
10	money-fx,interest	143
11	coffee	116
12	gold	99
13	grain,corn	94
14	money-fx,dlr	92
15	gnp	83
16	cpi	79
17	cocoa	63
18	jobs	55
19	copper	54
20	crude,nat-gas	53

- Distribution of top topics



4. Data Preparation for Modeling

- Because of limited time, I can explore general features and main features: NEWID, TITLE, BODY, TOPIC_NAME, LEWISSPLIT and then use them mainly for next steps

	NEWID	TITLE	BODY	TOPIC_NAME	LEWISSPLIT
5096	6716	EC WARNS CONGRESS ON NEW TRADE BILL	The European Community (EC) has\nwarned the U....	trade	TRAIN
5097	4493	SAFEGUARD SCIENTIFIC <SFE> IN EQUITY DEAL	Safeguard Scientifics Inc\nsaid it made a 2.5 ...	acq	TRAIN
5098	17700	COMPUTER ASSOC. <CA> BOLSTERS HAND AGAINST IBM	Computer Associates International Inc's\n800 m...	acq	TEST
5099	1739	CTS <CTS> AND DYNAMICS <DYA> REACH ACCORD	CTS Corp and Dynamics Corp of\nAmerica reached...	acq	TRAIN
5100	15995	JACOR <JCOR> TO BUY TWO DENVER RADIO STATIONS	Jacor Communications Inc said it\nagreed to bu...	acq	TEST
5101	8020	INVESTMENT FIRM HAS 7.1 PCT OF CYCLOPS <CYL<	Halcyon Investments, a New York risk\narbitrag...	acq	TRAIN
5102	10736	MIYAZAWA EXPECTS DOLLAR REBOUND SOON - SPOKESMAN	Japanese Finance Minister Kiichi Miyazawa\nexp...	money-fx,dlr	TRAIN
5103	1347	U.S. WINE EXPORTS ROSE 15 PER CENT LAST YEAR	Exports of American wine rose 14.9\nper cent l...	trade	TRAIN
5104	924	SWISS CAPITAL EXPORTS RISE IN JANUARY	Swiss capital exports rose to 4.64\nbillion fr...	trade	TRAIN
5105	5118	VENEZUELA TO LEND UP TO 12.5 MLN BARRELS OF OIL	Venezuela will lend Ecuador up to 12.5\nmln ba...	crude	TRAIN

- A new column, 'article', was created by combining the 'TITLE' and 'BODY' of the articles. This column was then cleaned of extraneous elements. The text in the 'article' column will be transformed into embeddings using the OpenAI 'text-embedding-ada-002' model, which returned a list of embeddings for each article. These embeddings were stored as features in a DataFrame.

	NEWID	TOPIC_NAME	LEWISSPLIT	article
5103	1347	trade	TRAIN	.S. WINE EXPORTS ROSE 15 PER CENT LAST YEAR\n\nExports of American wine rose 14.9\nper cent last year to a total of 7.2 million gallons, the Wine\nInstitute said.\n The San Francisco-based promotional group said 95 per cent\nof the exported wine was from California and the top three\nmarkets were Canada, Japan and Britain.\n Japan, which edged out Britain as the leading importer of\nAmerican wines in 1985, continued in second place.\n Canadian imports rose 3.2 per cent to 2.6 million gallons;\nJapan, 9.9 per cent to 1.2 million gallons, and Britain, 20.2\nper cent to 962,360 gallons.\n A spokesman for the institute said the declining value of\nthe dollar and strong promotional efforts on the part of the\nCalifornia wine industry contributed significantly to the\nincreased shipments.\n The 1986 wine exports were valued at 34.7 mln dlrs, up from\n27.6 mln dlrs the previous year.
5104	924	trade	TRAIN	SWISS CAPITAL EXPORTS RISE IN JANUARY\n\nSwiss capital exports rose to 4.64\nbillion francs in January after 2.54 billion in December and a\nyear earlier 3.64 billion, the Swiss National Bank said.\n New bond issues accounted for 4.12 billion of the total\nafter December's 2.15 billion, and credits 525.1 mln after\n389.9 mln.\n In January 1985, before the National Bank ended the\ndistinction between notes and bonds, bond issues totalled 1.66\nbillion francs, notes 1.39 billion and credits 597.5 mln.
5105	5118	crude	TRAIN	VENEZUELA TO LEND UP TO 12.5 MLN BARRELS OF OIL\n\nVenezuela will lend Ecuador up to 12.5\nmln barrels of crude oil to help it meet its export commitments\nand its domestic energy demand, Ecuadorean Energy and Mines\nMinister Javier Espinosa said today in a statement.\n Ecuador was forced to suspend exports after the pipeline\nconnecting its jungle oil fields with the Pacific Ocean port of\nBalao was damaged last week by an earthquake.\n Venezuela would lend 50,000 barrels per day of crude for a\ntotal of up to to 7.5 mln barrels to help Ecuador meet export\ncommitments, Espinosa said. Also, Venezuela will sell the crude\nand provide the foreign exchange earnings to Ecuador, he said.\n Ecuador would repay Venezuela in crude once it resumed its\nexports after repairing its pipeline to Balao, a task that\nwould take an estimated five months.\n Venezuela is lending Ecuador five mln barrels of crude for\nrefining in this country to meet domestic demand. Ecuador would\nrepay that loan with crude once the oil pipeline is repaired.\n Both countries are the only Latin American members of the\nOrganisation of Petroleum Exporting Countries (OPEC).\n Ecuador was exporting about 140,000 bpd before the\nearthquake, Energy Ministry officials said. Its total output\nwas around 260,000 bpd.

5. Embedding text to vectors

- Create embedding vectors, we can apply traditional method like word2vec, fasttext, doc2vec, bert,.. a lot of method but in the reality getting embedding vector using openAI with model text-embedding-ada-002

- Then save embedding vectors, newID to json file

	NEWID	prompt_tokens	embedding_feature0	embedding_feature1	embedding_feature2	embedding_feature3	embedding_feature4	embedding_feature5
5101	8020	166	0.000809	-0.029200	0.013126	-0.035182	-0.016485	0.011382
5102	10736	153	-0.038325	-0.031647	-0.012786	-0.010574	-0.005314	0.010869
5103	1347	231	-0.001636	-0.034642	0.008016	-0.031295	-0.003553	-0.004320
5104	924	149	-0.016515	-0.019256	0.005274	-0.028622	0.004542	0.008976
5105	5118	305	-0.024442	-0.017076	0.012521	-0.021542	-0.011928	-0.002946

6. Model Training and Evaluation

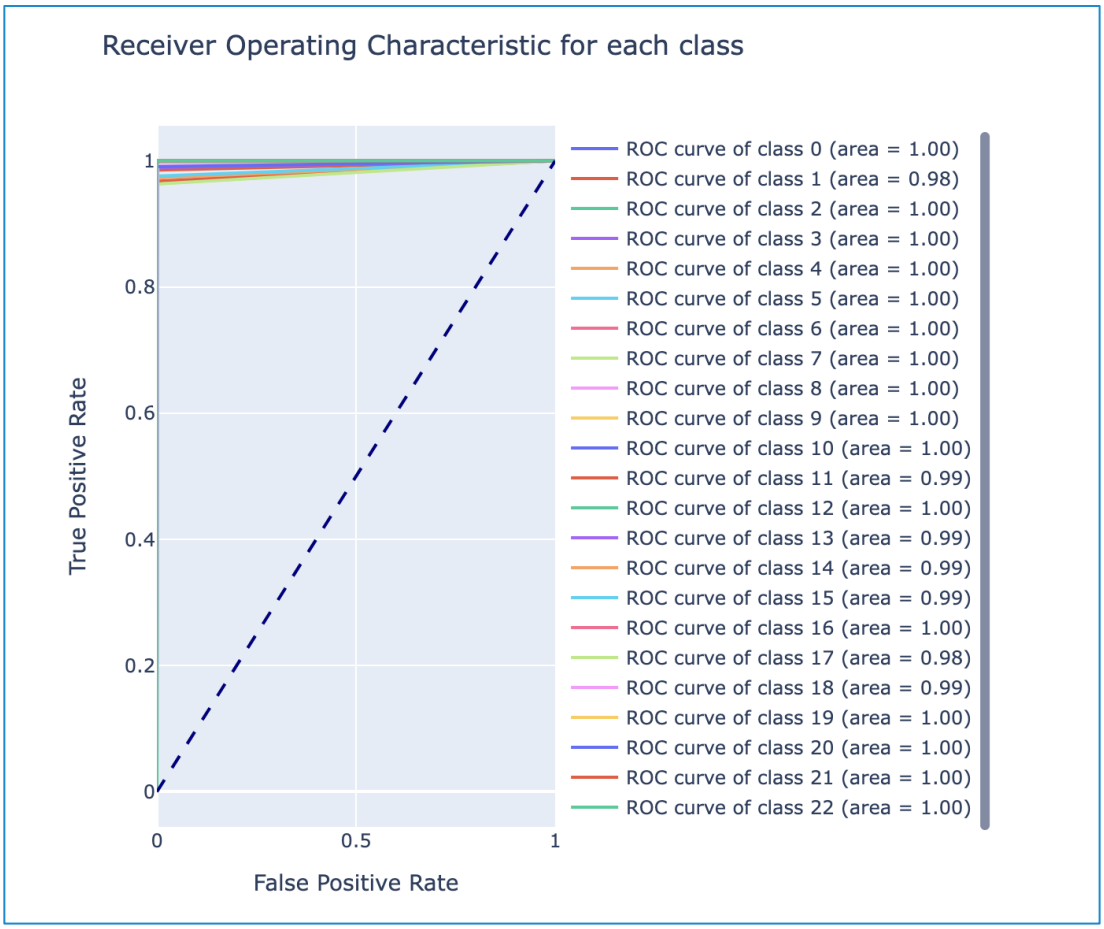
The XGBoost algorithm, a powerful gradient boosting algorithm, was used for the classification task. The model was trained on the embedding features with 'TOPIC_NAME' as the target variable. The performance of the model was evaluated using both training and test datasets.

Dataset	Accuracy
Train	99.51%
Test	88.93%

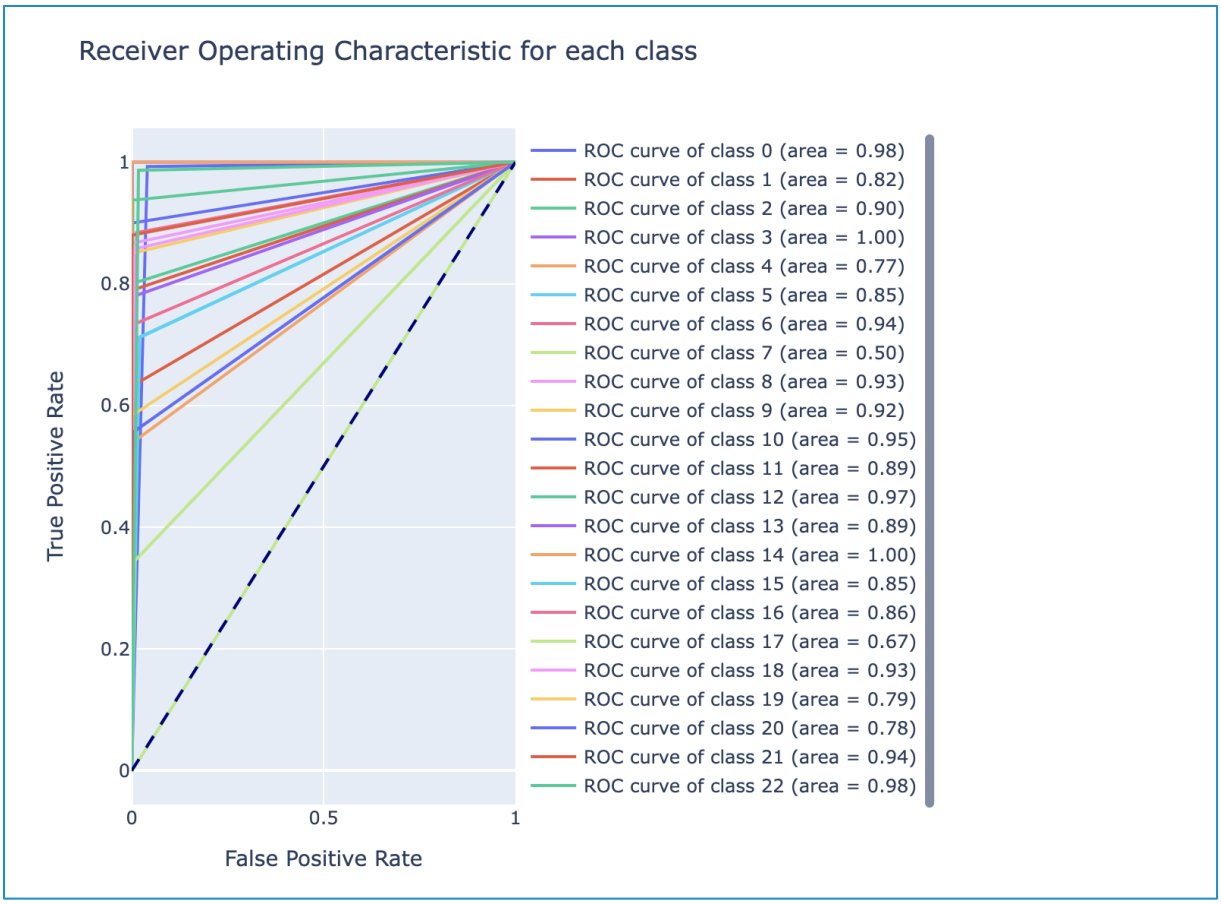
In addition to accuracy, other performance metrics such as precision, recall, and F1-score were calculated for both the test and training sets.

Dataset	Precision	Recall	F1-score
Train	99.53%	99.51%	99.52%
Test	88.61%	88.93%	88.35%

- ROC in train set:



- ROC in test set



+ The ROC curves are being plotted for each class separately. For each class, the class is treated as the positive class and all other classes are treated as the negative class.

+ The AUC is calculated for each class and is displayed in the legend of the plot next to the corresponding ROC curve.

+ There are a diagonal line from (0,0) to (1,1). This line represents the ROC curve of a random classifier (AUC = 0.5). A good classifier has its ROC curve much higher than this line (towards the top left corner of the plot).

+ The resulting plot shows multiple ROC curves, one for each class. By comparing the AUCs, you can see which classes your model is better at predicting. For example, in this outputs, the model seems to perform best for class 14 (AUC=1.0), and worst for class 7 (AUC = 0.4993).

7. Conclusion

The analysis achieved promising results in classifying the Reuters articles into their respective topics. The use of OpenAI's text embeddings facilitated the transformation of the text data into a format suitable for machine learning, contributing significantly to the model's performance.

Despite the model's high performance, there may be potential areas for improvement. For example, the model could be further fine-tuned or other models could be explored for better performance. Furthermore, more advanced text embedding techniques or the incorporation of additional contextual information into the embeddings could potentially improve the model's performance.

Overall, the analysis demonstrated the potential of machine learning in classifying news articles, which has wide-ranging applications in news agencies, publishing companies, and beyond.

8. Future work

Because of the limited time, I am still working in my current company as a leader of AI team. I am leading the team to finish releases of AI Evonik project in this sprint. So, I am super busy. Meanwhile, I have explored some main steps as above instruction. Future work, we can focus on the following topics to improve the accuracy of models and data quality:

- Data Cleaning: Although we're already cleaning the text data (removing extraneous elements), there could be room for more robust preprocessing. This could include removing or normalizing punctuation, removing stop words (common words that do not contribute much information), lemmatization (reducing words to their base or root form), and handling of typos.

- Feature Engineering: While the text embeddings capture a significant amount of information, there might be other features that could improve the model. For example, the length of an article, the number of unique words, or the frequency of certain words or phrases could all be informative. Also, meta-features like the day of the week or the time the news was published might be useful.

- Handling Imbalanced Classes: If the classes in the target variable are imbalanced, it could hinder the model's performance. Techniques like over-sampling the minority class, under-sampling the majority class, or using SMOTE (Synthetic Minority Over-sampling Technique) could be employed to balance the classes.

- Text Augmentation: Text data augmentation involves creating new synthetic data examples that add diversity to the existing data. Techniques for text augmentation include synonym replacement, random insertion, random deletion, and sentence shuffling. This could potentially improve the model's performance by providing it with a more diverse range of examples to learn from.

- Hyperparameter Tuning: Although "text-embedding-ada-002" might already be the best embedding model, the performance of the XGBoost model could potentially be improved by tuning its hyperparameters. This could involve adjusting parameters like the learning rate, maximum depth of the trees, and the number of estimators among others.

- Ensemble Methods: Using a combination of models instead of a single model could potentially improve performance. This could involve training multiple models and combining their predictions in some way (e.g., taking the average, voting, etc.). This approach is known as ensemble learning and it's often effective in reducing overfitting and improving the model's robustness and generalization ability.

- Use of Domain Knowledge: If there's any domain-specific knowledge available about the news topics, it could be used to create additional features that could help improve the model's performance.