CMT316: Applications of Machine Learning: Natural Language Processing and Computer Vision

Part 2

Roshan Roy

C2037620

Question 1)

When constructing the text classification model, the dataset served as the starting point, in which we organised and loaded the news articles into a Pandas DataFrame. The data was then categorised into their distinct classes such as business, entertainment, politics, sport, and tech, to ensure uniformity and consistency in subsequent analyses. Each text file was individually read, and its content, along with the corresponding category label, was recorded.

Each text file underwent a series of preprocessing steps, including lowercase conversion, punctuation removal, tokenization, stopword removal, and lemmatization. These steps collectively standardised the text, eliminating noise, and extracting meaningful information. This preliminary step was crucial in preparing the raw textual data for ensuing feature extraction and model training, ensuring the text data was optimised for accurate and effective classification.

The following phase involved splitting the dataset into training, validation, and test sets. By employing a stratified approach during this split was essential to maintain a balanced representation of categories in each subset. Features, represented by news articles, were separated from their corresponding category labels, creating distinct sets for the model training, validation, and testing phases. This separation was vital to assess the model's performance on unseen data accurately and for additional experimentation in refining the model architecture.

The choice of features became a pivotal step in the text classification, and three methods were used: Absolute Word Frequency (Bag-of-Words) , N-grams, and TF-IDF(Term Frequency-Inverse Document Frequency). Absolute Word Frequency was captured through the utilisation of the CountVectorizer, a technique that quantifies the occurrences of individual words in each document. N-grams, specifically bi-grams, were leveraged using the same CountVectorizer, allowing the model to capture relationships between pairs of adjacent words. TF-IDF, implemented through the TfidfVectorizer, assigned weights to terms based on their importance within a document relative to the entire amount. Each of these methods enhanced the model's ability to understand and generalise from the text.

After feature extraction, the chi-squared (chi2) method was applied for feature selection. This method aided in identifying the most informative features by selecting those most likely to be independent of each other. This step was crucial as it ensured that the model focused on the most discriminative aspects of the data.

Finally, I utilised the validation set to evaluate multiple classifiers for the different choice of features. The four classifiers used were - Support Vector Machine (SVM), Random Forest, Naive Bayes, and Logistic Regression - were initialised, trained on the entire training set, and evaluated on the validation set. The performance metrics, including accuracy, confusion matrix, and classification report, were computed for each classifier on the validation set. The best classifier was then used to predict the test set new articles.

Question 2)

Beginning with the use of Pandas for loading and preprocessing news articles offers a structured and efficient approach. Its DataFrames provide a tabular representation, facilitating seamless integration with natural language processing libraries. The built-in functionalities of Pandas enable easy data cleaning, transformation, and manipulation; in our case, creating different categories, preprocessing the data, and establishing the Train-Val-Test split. This approach ensures a systematic and organised workflow, especially when dealing with various categories and sizable volumes of news articles.

The preprocessing steps, including lowercase conversion, punctuation removal, tokenization, stopword removal, and lemmatization, were strategically chosen to ensure standardised, noise-free, and meaningful text representations. Lowercasing maintained consistency, punctuation removal reduced noise, tokenization laid the groundwork for subsequent analyses, stopword removal enhanced efficiency, and lemmatization consolidated word variations.

In optimising feature extraction for text classification, Absolute Word Frequency, N-grams, and TF-IDF methods were strategically chosen. Absolute Word Frequency is a straightforward metric quantifying the occurrences of individual words. N-grams, a more complex feature, retain contextual relationships by considering sequential word combinations. TF-IDF assesses the significance of a term within a document set, providing a normalised representation of word importance. The chi-squared method refines the feature set by selecting the most informative and independent features, enhancing the model's discriminative capabilities. This comprehensive approach ensures a nuanced representation of language use, emphasising both individual word prevalence and contextual dependencies.

The validation set played a pivotal role, ensuring a stratified split to maintain class balance and facilitate diverse model training. Furthermore, the validation set became instrumental in evaluating multiple classifiers with the choice of features, contributing to an iterative model selection process. The figures below display the accuracy of different classifiers with the various choices of features.

	SVM	Random Forest Classifier	Naïve Bayes Classifier	Logistic Regression Classifier
Accuracy	85.80	92.44	94.51	94.22

Figure 1: Classifier Accuracy Comparison using Absolute Word Frequency and Chi-Squared Feature Selection

	SVM	Random Forest Classifier	Naïve Bayes Classifier	Logistic Regression Classifier
Accuracy	85.67	93.94	94.89	94.22

Figure 2: Classifier Accuracy Comparison using n-grams and Chi-Squared Feature Selection

	SVM	Random Forest Classifier	Naïve Bayes Classifier	Logistic Regression Classifier
Accuracy	89.33	95.07	96.07	95.51

Figure 3: Classifier Accuracy Comparison using TF-IDF and Chi-Squared Feature Selection

	SVM	Random Forest Classifier	Naïve Bayes Classifier	Logistic Regression Classifier
Accuracy	87.08	94.66	96.15	94.38

Figure 4: Classifier Accuracy Comparison using a combination and Chi-Squared Feature Selection

We can clearly determine that Naïve Bayes is the most ideal classifier with all the choice of features. The validation set's use in selecting the best classifier based on performance underscored the importance of this intermediate evaluation step, mitigating overfitting risks and paving the way for robust performance on the independent test set.

Question 3)

	Absolute Word Frequency	N-grams	TF-IDF	Combination of choice of features
Accuracy	95.73	96.18	95.96	85.39
Macro-averaged precision	95.75	96.13	96.31	87.38
Macro-averaged recall	95.67	96.16	95.85	84.70
Macro-averaged F1	95.66	96.12	96.05	84.93

Figure 5: Classifiers Evaluation Metrics

The N-grams feature representation consistently demonstrated the highest performance across accuracy, macro-averaged recall, and F1 score on the original dataset, outperforming both Absolute Word Frequency and TF-IDF. While the combination of features, though achieving a decent accuracy of 85.39%, exhibited scores lower that the features individually. This may be caused by introduction of noise or conflicting information from different feature sets. Furthermore, the synergy between features might not always be optimal, and the model could struggle to discern relevant patterns, leading to a decrease in overall performance.

Question 4)

The final code and analysis could be enhanced by considering the possibility of overfitting, especially when dealing with complex models or extensive feature sets. It's important to scrutinise the model's performance on both the training and validation sets to identify signs of overfitting, which may manifest as a high training accuracy but lower performance on unseen data.

This could be improved by incorporating and comparing the different k-fold cross-validation makes when assess model performance across various data splits, enhancing generalisation insights. Additionally, utilising the validation set not just for classifier selection but also for hyperparameter tuning ensures optimal model configuration.

Finally, I conducted preliminary assessments using the models to classify new texts with varying lengths, ranging from entire sentences to individual words. These initial tests provided insights into how the models responded to different text lengths. However, for conclusive and robust results, a more comprehensive and in-depth testing regime, encompassing a broader range of text lengths, should have been employed.

```
Testing New Texts for Absolute Word Frequency:
Test String 1: 'the latest phone has the new 120MP camera which is able to zoom into the moon' - Predicted Category: tech
Test String 2: 'The latest financial reports indicate positive growth in the stock market sector.' - Predicted Category: business
Test String 3: 'A new blockbuster Lego movie is set to hit the theaters this weekend.' - Predicted Category: entertainment
Test String 4: 'World leaders engage in heated debates over key issues in the upcoming election.' - Predicted Category: politics
Test String 5: 'The football team QPR celebrated a remarkable victory in the championship against Leicester.' - Predicted Category: sport
Test String 6: 'gaming' - Predicted Category: tech
Test String 7: 'President' - Predicted Category: business
Test String 8: 'prime minister' - Predicted Category: politics
Test String 9: 'watching a new volleyball game' - Predicted Category: sport


Testing New Texts for TF-IDF:
Test String 1: 'the latest phone has the new 120MP camera which is able to zoom into the moon' - Predicted Category: sport
Test String 2: 'The latest financial reports indicate positive growth in the stock market sector.' - Predicted Category: business
Test String 3: 'A new blockbuster Lego movie is set to hit the theaters this weekend.' - Predicted Category: entertainment
Test String 4: 'World leaders engage in heated debates over key issues in the upcoming election.' - Predicted Category: politics
Test String 5: 'The football team QPR celebrated a remarkable victory in the championship against Leicester.' - Predicted Category: sport
Test String 6: 'gaming' - Predicted Category: sport
Test String 7: 'banking' - Predicted Category: sport
Test String 8: 'President' - Predicted Category: sport
Test String 9: 'prime minister' - Predicted Category: politics
Test String 10: 'watching a new volleyball game' - Predicted Category: sport


Testing New Texts for n-gram features:
Test String 1: 'the latest phone has the new 120MP camera which is able to zoom into the moon' - Predicted Category: tech
Test String 2: 'The latest financial reports indicate positive growth in the stock market sector.' - Predicted Category: business
Test String 3: 'A new blockbuster Lego movie is set to hit the theaters this weekend.' - Predicted Category: entertainment
Test String 4: 'World leaders engage in heated debates over key issues in the upcoming election.' - Predicted Category: politics
Test String 5: 'The football team QPR celebrated a remarkable victory in the championship against Leicester.' - Predicted Category: sport
Test String 6: 'gaming' - Predicted Category: tech
Test String 7: 'banking' - Predicted Category: business
Test String 8: 'President' - Predicted Category: business
Test String 9: 'prime minister' - Predicted Category: politics
Test String 10: 'watching a new volleyball game' - Predicted Category: sport
```

Figure 6: New Text Classification

For example, an in-depth analysis of how the models react to unfamiliar words or words that fit until multiple categories.

Extra Credit:

https://github.com/roy-roshan/CMT316-Coursework-1.git