Assignment 3.1: Text Classification using TF-IDF

Patricia Enrique

Department of Engineering, University of San Diego

AAI-520: Natural Language Processing

Prof. Mokhtari

September 25, 2023

Process

The dataset used for the text classification using TF-IDF is retrieved from Kaggle and consists of 2225 documents from the BBC news website corresponding to stories from one of the five topic categories of sport, business, politics, tech, or entertainment (Sharif, 2018). For this analysis, the news articles are tokenized, the stop words and punctuation are removed, and all words are lowercased before creating a data frame containing the preprocessed text and its corresponding category. A sample of the data frame is summarized in Table 1. The data frame is then split into training and testing datasets before converting the text to a matrix of TF-IDF features using TF-IDF vectorization and classified using three unique classifiers (Scikit Learn, 2023). The classifiers are then evaluated on accuracy, precision, recall, F1 score, Area Under the Receiver Operating Characteristic curve (AUC-ROC), and the confusion matrix.

Table 1

Preprocessed news articles from the BBC news website

ID	Text	Topic
373	wmc says xstrata bid low australian mining f...	Business
754	mogul wilson backing uk rap band tony wilson...	Entertainment
1134	lib dems new election pr chief lib dems appo...	Politics
1383	gadget growth fuels eco concerns technology ...	Tech
1709	new consoles promise big problems making gam...	Tech
2081	november remember saturday newspaper proclai...	sport

Analysis

Support Vector Classification (SVC)

The first classifier used on the data is SVC. The model is fit to the training data and then used to predict the values for the testing dataset. The performance of this model is then evaluated. To optimize the performance, the kernel and regularization parameters are tuned with the default values being radial

basis function (rbf) and 1.0 respectively. The performance of the model with the parameter tuning is summarized in Table 2. The confusion matrix that is resulting from the default parameters is shown in Figure 1.

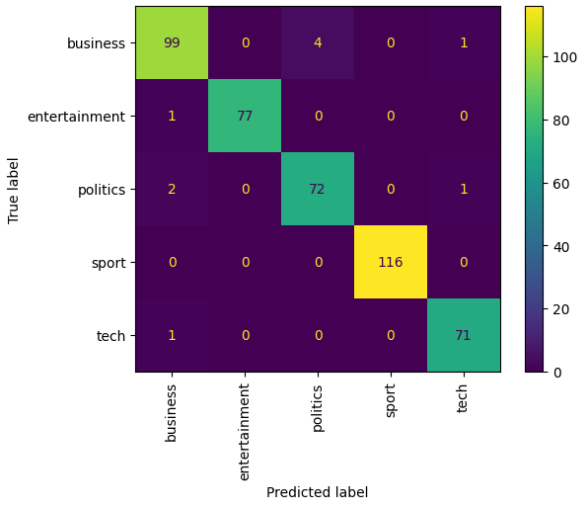
Table 2

SVC model evaluation with parameter tuning

Performance Metric	Kernel			Regularization		
	rbf	poly	sigmoid	1.0	10	50
Accuracy	0.98	0.77	0.98	0.98	0.98	0.98
Precision	0.98	0.88	0.98	0.98	0.98	0.98
Recall	0.98	0.77	0.98	0.98	0.98	0.98
F1 Score	0.98	0.76	0.98	0.98	0.98	0.98
AUC-ROC Score	0.99	0.99	0.99	0.99	0.99	0.99

Figure 1

Confusion matrix resulting from SVC with default parameters



Naïve Bayes Classifier

The Naïve Bayes classifier for multinomial models is implemented by fitting the model to the training data and then predicting the values for the testing dataset. The performance of this model is then

evaluated. To optimize the performance, the additive smoothing parameter is tuned with the default value being 1.0. The performance of the model with the parameter tuning is summarized in Table 3. The confusion matrix that is resulting from the default parameters is shown in Figure 2.

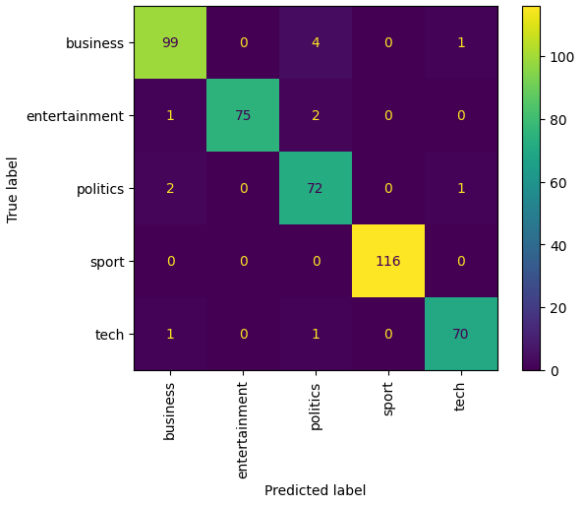
Table 3

Naïve Bayes model evaluation with parameter tuning

Performance Metric	Alpha				
	1	5	10	50	100
Accuracy	0.97	0.95	0.92	0.64	0.53
Precision	0.97	0.95	0.93	0.82	0.49
Recall	0.97	0.95	0.92	0.64	0.53
F1 Score	0.97	0.95	0.92	0.58	0.44
AUC-ROC Score	0.99	0.99	0.99	0.99	0.99

Figure 2

Confusion matrix resulting from the Naïve Bayes with default parameters



Gradient Boosting Classifier

The gradient boosting classifier is implemented by fitting the model to the training data and then predicting the values for the testing dataset. The performance of this model is then evaluated. To

optimize the performance, the learning rate and number of boosting stages are tuned with the default values being 0.1 and 100 respectively. The performance of the model with the parameter tuning is summarized in Table 4. The confusion matrix that is resulting from the default parameters is shown in Figure 3.

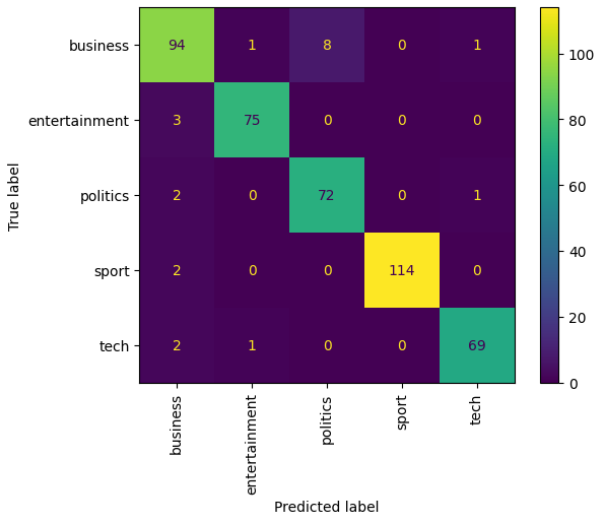
Table 4

Gradient boosting classifier evaluation with parameter tuning

Performance Metric	Learning Rate			Boosting Stages		
	0.1	0.01	0.001	50	100	500
Accuracy	0.96	0.91	0.74	0.95	0.96	0.96
Precision	0.96	0.92	0.82	0.95	0.96	0.96
Recall	0.96	0.91	0.74	0.95	0.96	0.96
F1 Score	0.96	0.92	0.75	0.95	0.96	0.96
AUC-ROC Score	0.99	0.98	0.95	0.99	0.99	0.99

Figure 3

Confusion matrix resulting from the gradient boosting classifier with default parameters



Results

The three classification models obtain excellent results with the default parameters when trained on the TF-IDF transformed training data. From the SVC model evaluation, it can be seen that the regularization parameter does not play a role in the model performance; however, the type of kernel applied does. Both the sigmoid and rbf kernels obtain 98% or above performance for each metric, while the polynomial kernel achieves around 77% in accuracy, recall, and F1 score. The multinomial Naïve Bayes model obtains the best results when the alpha is the default value of 1 and the performance steadily decreases as the parameter value increases. However, the AUC-ROC score remains at 99% for all alpha values. The gradient boosting classifier achieves the worst performance of the three classification models with 96% accuracy, precision, recall, and F1 score. The number of boosting stages selected does not affect the performance and as the learning rate decreases, so does the performance of the model.

The confusion matrices generated by the three classifiers using the default parameters illustrate that the models consistently obtain true positives and true negatives indicating that the actual value and the predicted value are the same. This evaluation is supported by the AUC-ROC score maintaining a value of 99% indicating that the models are efficient at distinguishing between the classes.

Conclusion

SVC, multinomial Naïve Bayes, and gradient boosting classifiers are used to accurately predict the topic category for multiple news articles with the use of TF-IDF vectorization. The performance of these models are evaluated using the accuracy, precision, recall, F1 score, AUC-ROC, and the confusion matrix. All three methods obtain above 96% performance in all metrics with the SVC model slightly outperforming the multinomial Naïve Bayes and gradient boosting classifiers. The classifiers illustrate the simplicity of completing this task with no domain knowledge and little parameter tuning, as the default parameters obtain the optimal performance.

References

Scikit Learn. (2023, September 23). sklearn.feature_extraction.text.TfidfVectorizer https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

Sharif, P. (2018). *BBC News Summary* [Data set]. Kaggle. https://www.kaggle.com/datasets/pariza/bbc-news-summary