NATURAL LANGUAGE UNDERSTANDING
ASSIGNMENT - 01

Name: Eppa Manoj Roll.No: B22AI018

PROBLEM 4: SPORTS OR POLITICS

Github Repo: https://github.com/manojeppa43/NLU_A1

Introduction:

Text classification is a common problem in the field of NLP where the goal is to assign a particular category for a given text.For this task we were supposed to collect data from the internet and build a classifier that reads a text and classifies it as a sports or politics draft. We were supposed to build three different classifiers using any of the feature representation techniques like n-grams, TF-IDF, Bag of words. In this task , I have used Bag of words feature representation and built three classifiers using the following techniques

a) Multinomial Naive Bayes (MNB)
b) Logistic Regression
c) Linear Support Vector Machine

Data Collection:

I surfed the entire internet to find a news dataset involving only 2 classes of labels aka Sports and Politics, but could not find an exact dataset matching it. So I came across this site: https://huggingface.co/datasets/okite97/news-data.

This News Dataset is an English-language dataset containing just over 4k unique news articles scrapped from AriseTv,one of the most popular news television networks in Nigeria.But it also has other labels like business, health along with the politics and sports labels. So I used this data and pre-processed it according to our task.

Data Description:

a) Data Instances:

{
'Title': 'Nigeria: APC Yet to Zone Party Positions Ahead of Convention',
'Excerpt': 'The leadership of the All Progressives Congress (APC), has denied reports that it had zoned some party positions ahead of',
'Category': 'politics',
'labels': 2
}

b) Data Fields:

Title: a string containing the title of a news title as shown above
Excerpt: a string containing a short extract from the body of the news
Category: a string that tells the category of an example (string label)
labels: integer telling the class of an example (label)

c) Total Number of instances in dataset: 4,594
d) Bias in Dataset: This data is biased towards news happenings in Nigeria but the model built using it can as well classify news from other parts of the world with a slight degradation in performance.

Data Cleaning:

So I got the dataset from huggingface, removed the instances labelled as business/health and used only the instances with the labels sports and politics. Then merged the title column and excerpt column into one and named it as “Text” and later removed punctuations,numbers and special characters and converted all of them into lowercase. Then I encoded the category into numeric values and renamed the column as label and saved the entire dataset into a csv file named “cleaned_data.csv”.

Data Pre-Processing:

I tokenize the text using CountVectorizer from scikit-learn to split the text into words and then used Bag of Words representation where we represent each document(here in our case text) as a vector of word counts.This ignores word order but it captures frequency information and split the data into train and test dataset with a test_to_train ratio of 0.2.

Explanation of the Classifier:

1) **Multinomial Naive Bayes (MNB)**:
   Multinomial naive bayes is a probabilistic model that works well for text classification. It predicts the class of a document based on the frequency of words in that document.
   This works in the following way:
   a) It calculates the probability of each class (i.e. sports and politics) in the training data aka prior probabilities
   b) Then for each word in the document/text, it calculates the probability of that word appearing in each class aka posterior probabilities
   c) Then it multiplies these probabilities together and chooses the class with highest probability
   I thought this works well for our data as it is simple and fast to train and it performs well even when the dataset is not very large and mainly it works well with word counts.

Results achieved using Multinomial Naive Bayes Classifier:
Accuracy: 99.0909

Metric	Politics	Sports	Overall(average)
Precision	0.98	1.00	0.99
Recall	1.00	0.98	0.99
F1-score	0.99	0.99	0.99

2) **Logistic Regression:**

Logistic regression is a generalized linear model specifically used for classification. It predicts the probability that a document/text belongs to a certain class(politics and sports in our case)
This works in the following way:

a) First we convert the text into numerical vectors using bag of words representation
b) Then it assigns weight to each word to show important that word is for each class
c) Then using a sigmoid function , it turns the weighted sum of words into a probability between 0 and 1
d) Then chooses the class with highest probability

I thought this works well for our case because by assigning more weight to important words the model can classify a new text easily and it works well with high dimensional data like text because each unique word becomes a feature leading to high dimensional feature space.

Results achieved using Logistic Regression based Classifier:
Accuracy: 98.4090

Metric	Politics	Sports	Overall
Precision	0.99	0.98	0.98
Recall	0.98	0.99	0.98
F1-score	0.98	0.98	0.98

3) **Linear Support Vector Machine:**

Linear SVM is a linear classifier that finds the best hyperplane in high dimensional space that separates two classes, in our case, sports and politics. This works in the following way:

a) First we convert the text into numerical vectors using bag of words representation so each word becomes a feature
b) Then it finds a hyperplane that maximizes the margin between the two classes (margin is the distance between the hyperplane and closest points from each class known as support vectors)
c) Then for a new class it will classify based on which side of the hyperplane does the feature vector of new text lies

I thought this works well for our case because it works well with high dimensional data like text where we have many features as each unique words leads to a new feature and it handles sparse data like most features are zero for a new text efficiently

Results achieved using Linear SVM based Classifier:
Accuracy: 99.0909

Metric	Politics	Sports	Overall
Precision	0.99	0.99	0.99
Recall	0.99	0.99	0.99
F1-score	0.99	0.99	0.99

Observations:

1) Multinomial Naive Bayes and Linear SVM have achieved the highest accuracy of 99.09 percent, logistic regression was slightly lower at 98.40 but it still performed well.
2) Noticed that Multinomial Naive Bayes is fast and simple and works well with the word counts
3) Linear SVM also handles the high dimensional data like word counts efficiently and gives high accuracy but it is more computationally intensive than the multinomial naive bayes.

4) Logistic Regression is actually strong for high dimensional data but it is surprising to me as it underformed compared to the other two methods.

Limitations:

We can see that all the models performed well, but there are some limitations:

1) Dataset bias:

   The dataset is biased toward news from Nigeria so the model may perform slightly worse on news from other countries, but there won't be a big difference as sports and politics are very different fields(i.e. they don't overlap much).

2) Small Dataset:

   The final dataset has only 2198 instances which is relatively small for the text classification

3) Feature Representation:

   As we have used Bag of words for feature representation , it does not capture the word order and context. It works well for small/medium sentences but for complex sentences some meaning might be lost
   For example take this sentence: “PM modi and Amit shah along with president have attended the inauguration of world cup event”, even though the ground truth for this is “sports” the models classify them as politics as it sees many politicians in the sentence.

4) Generalization Issue:

   The model may not generalize well for the real world as it can't handle sarcasm,idioms or complex language since it only relies on word frequencies.

Conclusion:

Even though all the three classifiers we built had performed well it may not work well in the outside world as there were many limitations. We could improve them by using the deep learning models aka neural nets to capture the context which could improve the generalization compared to these methods