



Text Classification

Homework 1 Week 7

Machine Learning and Business Intelligence CS 550

Presented by Imran Noor Saleh ID 19648



Introduction

- Text classification is the process of assigning predefined categories to text data based on their content.
- The objective of our project is to develop a machine learning model that can accurately classify text data into different categories.
- We used various machine learning algorithms and techniques to achieve our objective.
- To prepare the dataset for machine learning, we performed feature extraction and selection techniques. We used bag-of-words, TF-IDF, and word embeddings to represent the text data as numerical features.
- We also performed feature selection techniques to identify the most relevant features that contribute to the classification task.
- We used various machine learning algorithms such as Naive Bayes, Support Vector Machines, and Neural Networks for text classification.



Dataset

The objective is to test the Text Classifier to predict who the real author of Hamlet is, using this dataset.

	Doc	Words	Author
Training	1	W1 W2 W3 W4 W5	C (<u>Christopher Marlowe</u>)
	2	W1 <u>W1</u> W4 W3	C (<u>Christopher Marlowe</u>)
	3	W1 W2 W5	C (<u>Christopher Marlowe</u>)
	4	W5 W6 W1 W2 W3	W (<u>William Stanley</u>)
	5	W4 W5 W6	W (<u>William Stanley</u>)
	6	W4 W6 W3	F (<u>Francis Bacon</u>)
	7	W2 <u>W2</u> W4 W3 W5 <u>W5</u>	F (<u>Francis Bacon</u>)
Test	8 (Hamlet)	W1 W4 W6 W5 W3	?



Manual Calculation

To predict who the real author of Hamlet is, we have the training data and the probability of each to be calculated

$P(C)$: The probability of class C = $3/7$

$P(W)$: The probability of class W = $2/7$

$P(F)$: The probability of class F = $2/7$

$P(W1|C)$: The probability that the word "W1" appears on the 3 class c documents

= $(\text{count}(W1, C) + \underline{1}) / (\text{count}(C) + |V|) = (4+1) / (12+6) = 5/18$

4: how many times the word "W1" appear on the 3 class C documents., 12: how many words in the 3 class C documents, 6: number of vocabulary: (W1 W2 W3 W4 W5 W6)



Manual Calculation Contd.

$P(W1|W)$: The probability that the word "W1" appears on the 3 class W documents

$$= (\text{count}(W1, W) + \underline{1}) / (\text{count}(W) + |V|)$$

$$= (1+1) / (8+6) = 2/14 = 1/7$$

1: how many times the word "W1" appear on the 2 class W documents.

8 : how many words in the 3 class W documents.

6: number of vocabulary: (W1 W2 W3 W4 W5 W6)



Manual Calculation Contd.

$P(W1|F)$: The probability that the word "W1" appears on the 2 class F documents

$$= (\text{count}(W1, F) + \underline{1}) / (\text{count}(F) + |V|)$$

$$= (0+1) / (9+6) = 1/15$$

0: how many times the word "W1" appear on the 2 class F documents.

9: how many words in the 3 class W documents.

6 : number of vocabulary: (W1 W2 W3 W4 W5 W6)



Manual Calculation Contd.

$P(W3|C)$: The probability that the word "W3" appears on the 3 class C documents

$$= (\text{count}(W3, C) + \underline{1}) / (\text{count}(C) + |V|)$$

$$= (2+1) / (12+6) = 3/18 = \frac{1}{6}$$

2: how many times the word "W3" appear on the 3 class C documents.

12 : how many words in the 3 class C documents.

6 : number of vocabulary: (W1 W2 W3 W4 W5 W6)



Manual Calculation Contd.

$P(W3|W)$: The probability that the word "W3" appears on the 3 class W documents

$$= (\text{count}(W3, W) + \underline{1}) / (\text{count}(W) + |V|)$$

$$= (1+1) / (8+6) = 2/14 = 1/7$$

1: how many times the word "W3" appear on the 2 class W documents.

8 : how many words in the 3 class W documents.

6: number of vocabulary: (W1 W2 W3 W4 W5 W6)



Manual Calculation Contd.

$P(W3|F)$: The probability that the word "W3" appears on the 2 class F documents

$$= (\text{count}(W3, F) + \underline{1}) / (\text{count}(F) + |V|)$$

$$= (2+1) / (9+6) = 3/15 = 1/5$$

2: how many times the word "W3" appear on the 2 class F documents.

9: how many words in the 3 class F documents.

6: number of vocabulary: (W1 W2 W3 W4 W5 W6)



Manual Calculation Contd.

$P(W4|C)$: The probability that the word "W4" appears on the 3 class C documents

$$= (\text{count}(W4, C) + \underline{1}) / (\text{count}(C) + |V|)$$

$$= (2+1) / (12+6) = 3/18 = 1/6$$

2: how many times the word "W4" appear on the 3 class C documents.

12 : how many words in the 3 class C documents.

6 : number of vocabulary: (W1 W2 W3 W4 W5 W6)



Manual Calculation Contd.

$P(W4|W)$: The probability that the word "W4" appears on the 3 class W documents

$$= (\text{count}(W4, W) + \underline{1}) / (\text{count}(W) + |V|)$$

$$= (1+1) / (8+6) = 2/14 = 1/7$$

1: how many times the word "W4" appear on the 2 class W documents.

8 : how many words in the 3 class W documents.

6: number of vocabulary: (W1 W2 W3 W4 W5 W6)



Manual Calculation Contd.

$P(W4|F)$: The probability that the word "W4" appears on the 2 class F documents

$$= (\text{count}(W4, F) + \underline{1}) / (\text{count}(F) + |V|)$$

$$= (2+1) / (9+6) = 3/15$$

2: how many times the word "W4" appear on the 2 class F documents.

9: how many words in the 3 class F documents.

6: number of vocabulary: (W1 W2 W3 W4 W5 W6)



Manual Calculation Contd.

$P(W5|C)$: The probability that the word "W5" appears on the 3 class C documents

$$= (\text{count}(W5, C) + \underline{1}) / (\text{count}(C) + |V|)$$

$$= (2+1) / (12+6) = 3/18 = 1/6$$

2: how many times the word "W5" appear on the 3 class C documents.

12 : how many words in the 3 class C documents.

6 : number of vocabulary: (W1 W2 W3 W4 W5 W6)



Manual Calculation Contd.

$P(W5|W)$: The probability that the word "W5" appears on the 3 class W documents

$$= (\text{count}(W5, W) + \underline{1}) / (\text{count}(W) + |V|)$$

$$= (2+1) / (8+6) = 3/14$$

2: how many times the word "W5" appear on the 2 class W documents.

8 : how many words in the 3 class W documents.

6: number of vocabulary: (W1 W2 W3 W4 W5 W6)



Manual Calculation Contd.

$P(W5|F)$: The probability that the word "W5" appears on the 2 class F documents

$$= (\text{count}(W5, F) + \underline{1}) / (\text{count}(F) + |V|)$$

$$= (2+1) / (9+6) = 3/15$$

2: how many times the word "W5" appear on the 2 class F documents.

9: how many words in the 3 class F documents.

6: number of vocabulary: (W1 W2 W3 W4 W5 W6).



Manual Calculation Contd.

$P(W6|C)$: The probability that the word "W6" appears on the 3 class C documents

$$= (\text{count}(W6, C) + \underline{1}) / (\text{count}(C) + |V|)$$

$$= (0+1) / (12+6) = 1/18$$

0: how many times the word "W6" appear on the 3 class C documents.

12 : how many words in the 3 class C documents.

6 : number of vocabulary: (W1 W2 W3 W4 W5 W6)



Manual Calculation Contd.

$P(W6|W)$: The probability that the word "W6" appears on the 2 class W documents

$$= (\text{count}(W6, W) + \underline{1}) / (\text{count}(W) + |V|)$$

$$= (2+1) / (8+6) = 3/14$$

2: how many times the word "W6" appear on the 2 class W documents.

8 : how many words in the 3 class W documents.

6: number of vocabulary: (W1 W2 W3 W4 W5 W6)



Manual Calculation Contd.

$P(W6|F)$: The probability that the word "W6" appears on the 2 class F documents

$$= (\text{count}(W6, F) + \underline{1}) / (\text{count}(F) + |V|)$$

$$= (1+1) / (9+6) = 2/15$$

1: how many times the word "W6" appear on the 2 class F documents.

9: how many words in the 3 class F documents.

6: number of vocabulary: (W1 W2 W3 W4 W5 W6)



Manual Calculation Contd.

$$P(C|d8) : P(C) * P(W1|C) * P(W4|C) * P(W6|C) * P(W5|C) * P(W3|C)$$

$$= ((3/7) * (5/18) * (1/6) * (1/18) * (1/6) * (1/6))$$

$$= 0.00003061924, \text{ approx } 0.00003$$

$$= 3/7: \text{ prior : } P(C)$$

There are 5 words in d8 : W1 W4 W6 W5 W3

Each word "W1" has $P(W1|C) = 5/18$, The word "W4" has $P(W4|C) = 3/18 = 1/6$, The word "W6" has $P(W6|C) = 1/18$, The word "W5" has $P(W5|C) = 3/18 = 1/6$, The word "W3" has $P(W3|C) = 3/18 = 1/6$



Manual Calculation Contd.

$$P(W|d8) = P(W) * P(W1|W) * P(W4|W) * P(W6|W) * P(W5|W) * P(W3|W)$$

$$= (2/7 * 2/14 * 2/14 * 3/14 * 3/14 * 2/14)$$

$$= 0.00004 = 2/7: \text{prior} : P(W)$$

There are 5 words in d8 : W1 W4 W6 W5 W3

Each word "W1" has $P(W1|W) = 2/14$, The word "W4" has $P(W4|W) = 2/14$, The word "W6" has $P(W6|W) = 3/14$, The word "W5" has $P(W5|W) = 3/14$, The word "W3" has $P(W3|W) = 2/14$



Manual Calculation Contd.

$$P(F|d8) = P(F) * P(W1|F) * P(W4|F) * P(W6|F) * P(W5|F) * P(W3|F)$$

$$= (2/7) * (1/15) * (3/15) * (2/15) * (3/15) * (3/15)$$

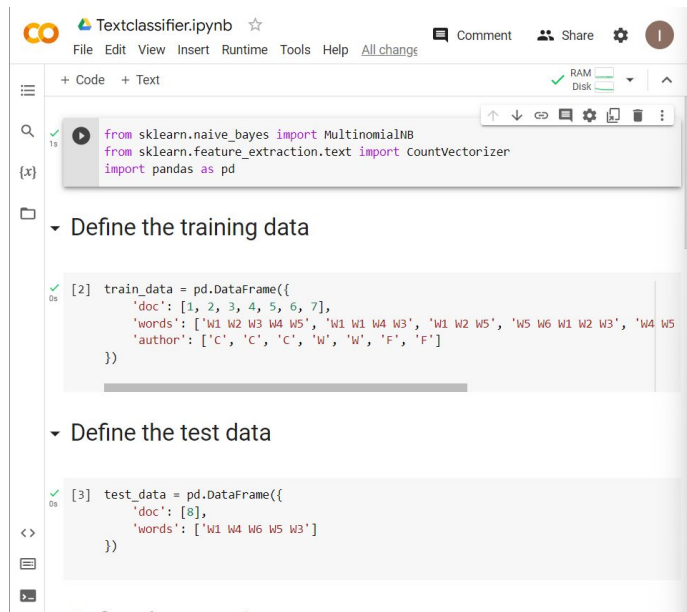
$$= 0.00002 = 2/7 : \text{prior} : P(F)$$

There are 5 words in d8 : W1 W4 W6 W5 W3

Each word "W1" has $P(W1|F) = 1/15$, The word "W4" has $P(W4|F) = 3/15$, The word "W6" has $P(W6|F) = 2/15$, The word "W5" has $P(W5|F) = 3/15$, The word "W3" has $P(W3|F) = 3/15$

Document 8 should belong to class W because it has the highest probability calculation.

Programming Solution

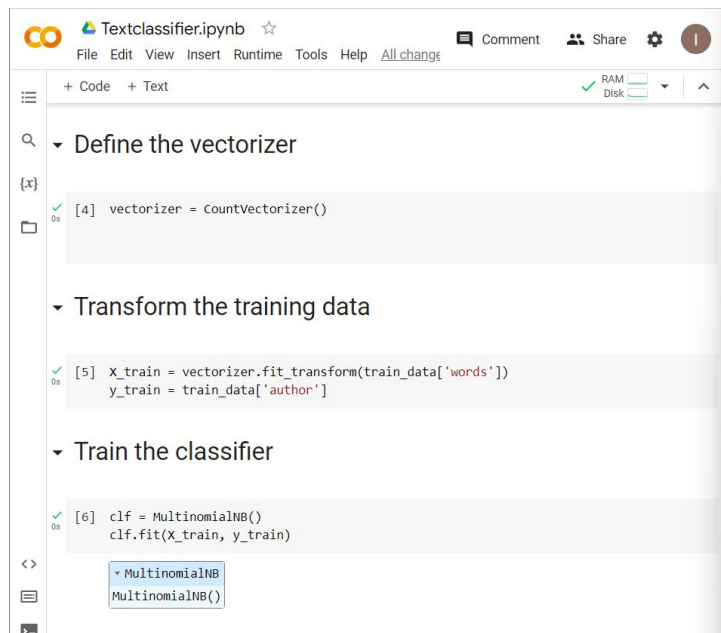


The image shows a Jupyter Notebook interface for a file named 'Textclassifier.ipynb'. The interface includes a top menu bar with options like File, Edit, View, Insert, Runtime, Tools, Help, and a status bar showing RAM and Disk usage. The notebook is divided into three sections: 'Define the training data', 'Define the test data', and 'Define the vectorizer'. The first section contains a code cell with imports for MultinomialNB, CountVectorizer, and pandas. The second section contains a code cell that creates a DataFrame for training data with columns for document ID, words, and author. The third section contains a code cell that creates a DataFrame for test data with columns for document ID and words.

```
from sklearn.naive_bayes import MultinomialNB
from sklearn.feature_extraction.text import CountVectorizer
import pandas as pd

[2] train_data = pd.DataFrame({
    'doc': [1, 2, 3, 4, 5, 6, 7],
    'words': ['W1 W2 W3 W4 W5', 'W1 W1 W4 W3', 'W1 W2 W5', 'W5 W6 W1 W2 W3', 'W4 W5', 'W1 W2 W3 W4 W5'],
    'author': ['C', 'C', 'C', 'W', 'W', 'F', 'F']
})

[3] test_data = pd.DataFrame({
    'doc': [8],
    'words': ['W1 W4 W6 W5 W3']
})
```



The image shows a Jupyter Notebook interface for a file named 'Textclassifier.ipynb'. The interface includes a top menu bar with options like File, Edit, View, Insert, Runtime, Tools, Help, and a status bar showing RAM and Disk usage. The notebook is divided into three sections: 'Define the vectorizer', 'Transform the training data', and 'Train the classifier'. The first section contains a code cell that creates a CountVectorizer object. The second section contains a code cell that uses the vectorizer to transform the training data into X_train and y_train. The third section contains a code cell that creates a MultinomialNB classifier and fits it to the training data. A dropdown menu is visible below the code cell, showing the class name 'MultinomialNB' and the method 'fit(X_train, y_train)'.

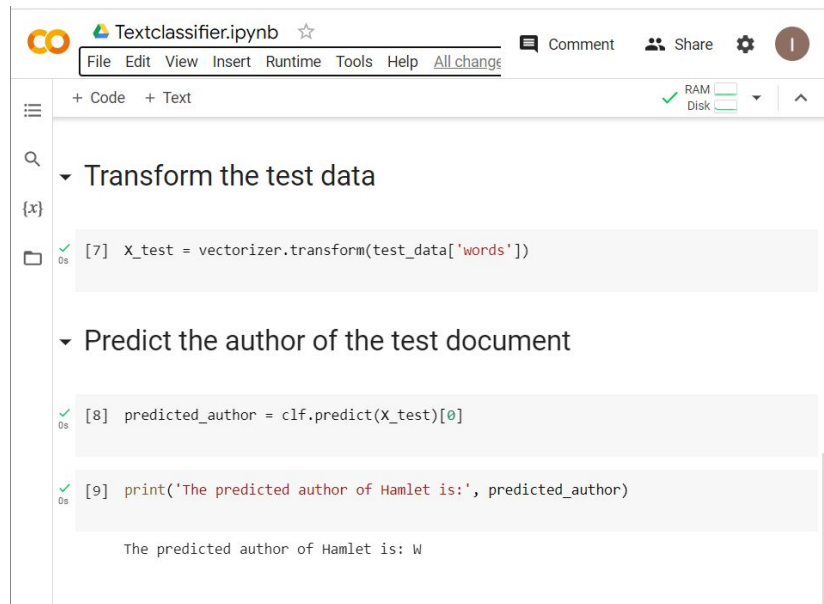
```
[4] vectorizer = CountVectorizer()

[5] X_train = vectorizer.fit_transform(train_data['words'])
    y_train = train_data['author']

[6] clf = MultinomialNB()
    clf.fit(X_train, y_train)
```

Programming Solution Contd.

So, the manually calculated and programmed solution is the same for this problem.



The screenshot shows a Jupyter Notebook titled "Textclassifier.ipynb". The interface includes a menu bar (File, Edit, View, Insert, Runtime, Tools, Help, All changes), a toolbar with icons for code and text, and a sidebar with a search icon and a file explorer icon. The notebook content is organized into sections: "Transform the test data" and "Predict the author of the test document".

```
[7] X_test = vectorizer.transform(test_data['words'])
```

```
[8] predicted_author = clf.predict(X_test)[0]
```

```
[9] print('The predicted author of Hamlet is:', predicted_author)
```

The predicted author of Hamlet is: W