



North East University Bangladesh (NEUB)

Department of Computer Science & Engineering

Newspaper Article Headline Category Prediction using Naïve Bayes Algorithm

Course Code: CSE-456

Course Title: Machine Learning Lab

Submitted To

Tasnim Zahan Tithy

Assistant Professor & Head

Dept. of CSE

NEUB, Sylhet – 3100

Submitted By

Md. Jahidul Islam

Reg. No.: 200103020029

Md Ashrafuzzaman Sunny

Reg. No.: 200203020002

Project Title: Newspaper Article Headline Category Prediction using Naïve Bayes Algorithm.

Objectives:

Input: Headlines of Newspaper Articles.

Output: Using Naïve Bayes Algorithm, the headline will be classified into one of several categories (Sports, Religion, Science, Politics, travel, and, etc.).

Methodology:

Dataset: Newspaper articles collected from Kaggle [3].

	category	headline	authors	link	short_description	date
0	CRIME	There Were 2 Mass Shootings In Texas Last Week...	Melissa Jeltsen	https://www.huffingtonpost.com/entry/texas-ama...	She left her husband. He killed their children...	2018-05-26
1	ENTERTAINMENT	Will Smith Joins Diplo And Nicky Jam For The 2...	Andy McDonald	https://www.huffingtonpost.com/entry/will-smit...	Of course it has a song.	2018-05-26
2	ENTERTAINMENT	Hugh Grant Marries For The First Time At Age 57	Ron Dicker	https://www.huffingtonpost.com/entry/hugh-gran...	The actor and his longtime girlfriend Anna Ebe...	2018-05-26
3	ENTERTAINMENT	Jim Carrey Blasts 'Castrato' Adam Schiff And D...	Ron Dicker	https://www.huffingtonpost.com/entry/jim-carre...	The actor gives Dems an ass-kicking for not fl...	2018-05-26
4	ENTERTAINMENT	Julianna Margulies Uses Donald Trump Poop Bags...	Ron Dicker	https://www.huffingtonpost.com/entry/julianna-...	The "Dietland" actress said using the bags is ...	2018-05-26
...
200848	TECH	RIM CEO Thorsten Heins' 'Significant' Plans Fo...	Reuters, Reuters	https://www.huffingtonpost.com/entry/rim-ceo-t...	Verizon Wireless and AT&T are already promotin...	2012-01-28
200849	SPORTS	Maria Sharapova Stunned By Victoria Azarenka I...		https://www.huffingtonpost.com/entry/maria-sha...	Afterward, Azarenka, more effusive with the pr...	2012-01-28
200850	SPORTS	Giants Over Patriots, Jets Over Colts Among M...		https://www.huffingtonpost.com/entry/super-bow...	Leading up to Super Bowl XLVI, the most talked...	2012-01-28
200851	SPORTS	Aldon Smith Arrested: 49ers Linebacker Busted ...		https://www.huffingtonpost.com/entry/aldon-smi...	CORRECTION: An earlier version of this story i...	2012-01-28

Data-Preprocessing:

i) Only keep the 'category' and 'headline' columns. Drop other columns.

```
1 df = df.drop(columns=['authors','link','short_description','date'])
2 df
```

	category	headline
0	CRIME	There Were 2 Mass Shootings In Texas Last Week...
1	ENTERTAINMENT	Will Smith Joins Diplo And Nicky Jam For The 2...
2	ENTERTAINMENT	Hugh Grant Marries For The First Time At Age 57
3	ENTERTAINMENT	Jim Carrey Blasts 'Castrato' Adam Schiff And D...
4	ENTERTAINMENT	Julianna Margulies Uses Donald Trump Poop Bags...
...
200848	TECH	RIM CEO Thorsten Heins' 'Significant' Plans Fo...
200849	SPORTS	Maria Sharapova Stunned By Victoria Azarenka I...
200850	SPORTS	Giants Over Patriots, Jets Over Colts Among M...
200851	SPORTS	Aldon Smith Arrested: 49ers Linebacker Busted ...
200852	SPORTS	Dwight Howard Rips Teammates After Magic Loss ...

200853 rows × 2 columns

ii) Merge common categories, like, 'ARTS & CULTURE' and 'CULTURE & ARTS'.

```
1 # Replace 'arts & culture' with 'culture and arts'
2
3 df['category'] = df['category'].replace('ARTS & CULTURE', 'CULTURE & ARTS')
4 df['category'] = df['category'].replace('PARENTING', 'PARENTS')
5 df['category'] = df['category'].replace('THE WORLDPOST', 'WORLDPOST')
6
7 print(df['category'])
8
9 len(df.category.unique())
```

iii) convert all the text in the 'headline' column to lowercase characters.

```
1 #convert all headlines to lowercase
2
3 df['headline'] = df['headline'].str.lower()
4 df
5 df.head(10)
```

	category	headline
0	CRIME	there were 2 mass shootings in texas last week...
1	ENTERTAINMENT	will smith joins diplo and nicky jam for the 2...
2	ENTERTAINMENT	hugh grant marries for the first time at age 57
3	ENTERTAINMENT	jim carrey blasts 'castrato' adam schiff and d...
4	ENTERTAINMENT	julianna margulies uses donald trump poop bags...
5	ENTERTAINMENT	morgan freeman 'devastated' that sexual harass...
6	ENTERTAINMENT	donald trump is lovin' new mcdonald's jingle i...
7	ENTERTAINMENT	what to watch on amazon prime that's new this ...
8	ENTERTAINMENT	mike myers reveals he'd 'like to' do a fourth ...
9	ENTERTAINMENT	what to watch on hulu that's new this week

iv) Using stop_words Library, remove common words from the **training set** like 'the', 'is', 'at', etc.

```
1 #list of common words to be removed
2
3 import re
4 #Remove common words from the 'train_set'
5
6 for word in stop_words:
7     X_train = X_train.str.replace(r'\b' + re.escape(word) + r'\b', '', regex=True).str.strip()
8
9 X_train
```

```
7916      photos capture brutal devastation  california ...
62456      woman bitten police dog  slept challenging ...
27599      suicide bomb blast hits nato convoy  kabul, ki...
67362      police arrest stabbing suspect  'psycho' foreh...
139069      amazing restaurant bathrooms  america
...
194442      hindu wedding planners thrive  united states
65615      ohio hospital performs first uterus transplant...
77655      story  know  women  helped overturn doma
56088      orlando attack could mark  shift  gay muslims
38408      nunes finishes rousey  ufc 207, garbrandt deth...
Name: headline, Length: 186793, dtype: object
```

iv) Using stop_words Library, remove common words from the **test set** like ‘the’, ‘is’, ‘at’, etc.

```
1 #remove common words from test set
2
3 for word in stop_words:
4     X_test = X_test.str.replace(r'\b' + re.escape(word) + r'\b', '', regex=True).str.strip()
5
6 X_test
```



```
37896      ' going  women' march washington  daughters...
28243      13 times latinos refused stay silent trump' ...
65110      evangelical christians hand donald trump win ...
104989              temporarily physical disability
193964              takes village
...
143236              ode elasticity mother heart
170388              let' talk sex
172992      thanksgiving leftovers: 5 ways get diet back...
2182      beyoncé announces $100,000 scholarships hbcu...
160756              creating violent teens?
Name: headline, Length: 14060, dtype: object
```

v) Split data into a training set (**93%**) and test set (**7%**).

```
1 #splitting data into training and test set
2 X = df['headline']
3 Y = df['category']
4
5 from sklearn.model_selection import train_test_split
6 X_train, X_test, y_train, y_test = train_test_split(X, Y, random_state=100, test_size=0.07, shuffle=True)
```



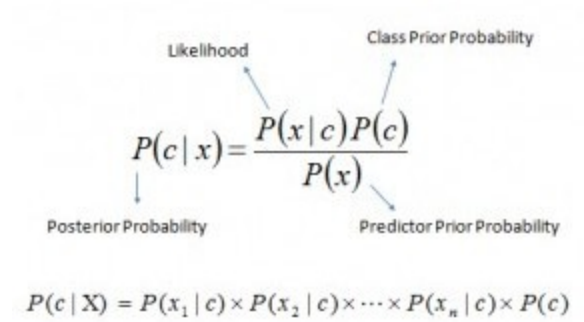
```
1 y_test
```



```
37896      WOMEN
28243      LATINO VOICES
65110      POLITICS
104989      WOMEN
193964      WELLNESS
...
143236      PARENTS
170388      WELLNESS
172992      BLACK VOICES
2182      EDUCATION
160756      PARENTS
Name: category, Length: 14060, dtype: object
```

Model Training:

Using **Naïve Bayes** Algorithm, using this formula.



The diagram shows the Naive Bayes formula with labels for its components:

$$P(c | x) = \frac{P(x | c) P(c)}{P(x)}$$

Labels and arrows:

- Likelihood** points to $P(x | c)$
- Class Prior Probability** points to $P(c)$
- Posterior Probability** points to $P(c | x)$
- Predictor Prior Probability** points to $P(x)$

$$P(c | X) = P(x_1 | c) \times P(x_2 | c) \times \dots \times P(x_n | c) \times P(c)$$

Step – 1: Find the probability of each class occurrence.

```
1  #finding probability of each class out of total classes
2
3  class_probability = {}
4
5  for key,value in class_count.items():
6      class_probability[key] = value/total_classes
7
8  class_probability
```

```
{'POLITICS': 0.16298255287939056,
 'WELLNESS': 0.08873458855524564,
 'ENTERTAINMENT': 0.07991198813660041,
 'PARENTS': 0.06288244206153336,
 'TRAVEL': 0.049327330253275015,
 'STYLE & BEAUTY': 0.048213798161601346,
 'HEALTHY LIVING': 0.033052630451890594,
 'FOOD & DRINK': 0.031200312645548817,
 'QUEER VOICES': 0.031194959125877306,
 'WORLDPOST': 0.03113607040949072,
 'BUSINESS': 0.02955142858672434,
 'COMEDY': 0.025702247942910067,
 'SPORTS': 0.02436386802503306,
 'BLACK VOICES': 0.022522257258034296,
 'HOME & LIVING': 0.020744888727093628,
 'WEDDINGS': 0.018196613363455804,
 'WOMEN': 0.01753277692418881,
 'IMPACT': 0.017313282617656977,
 'DIVORCE': 0.017024192555395546,
 'CRIME': 0.01687429400459332,
 'MEDIA': 0.014020868019679538,
```


Step - 2: Merging all the headlines of each category separately into single strings. Each string holds all the words associated with that particular category.

```
1  #storing all the words in each class separately in a dictionary
2
3  filtered_classes = {}
4  new = pd.concat([y_train, X_train], axis=1)
5
6  for key,value in class_count.items():
7      s = ''
8      x = new.loc[df['category'] == key, 'headline']
9      for item in x:
10         s += item
11         s = s.split()
12         filtered_classes[key] = s
13
14  filtered_classes['SPORTS']
```

```
['floyd',
 'mayweather',
 'jr.',
 'girlfriend:',
 'miss',
 'shantel',
 'jackson',
 'wears',
 'revealing',
 'dress',
 'fight',
 '(video)chloe',
 'kim',
 'said',
 ''hangry''
```

Step - 3: Find the total number of unique words in the training dataset.

```
1 #finding no. of total unique words in the
2 total_words = []
3 for key, value in filtered_classes.items():
4     total_words += value
```

```
1 len(total_words)
```

1081416

```
1 total_unique_words = set(total_words)
2 len(total_unique_words)
```

258210

Step - 4: Apply Naïve Bayes Formula to the training dataset, using the parameters calculated above

```
1 # Applying Naive Bayes formula to the training dataset
2 def findAnswer(item):
3     probabilites = {}
4     test_string = item
5     test_list = test_string.split()
6
7     probability_list = []
8
9     for i in test_list:
10         for key, value in filtered_classes.items():
11             count = 0
12             for item in value:
13                 if i==item:
14                     count += 1
15             total_class_words = len(value)
16             probabilites[key] = (count+1)/(total_class_words + len(total_unique_words))
17         probability_list.append(probabilites)
18
19     answer = []
20     for key, value in class_probability.items():
21         a = value
22         for idx, val in enumerate(test_list):
23             a *= probability_list[idx][key]
24         answer.append(a)
25
26     return answer.index(max(answer))
27
```

Model Evaluation:

Find the accuracy using **500 test** data.

We implemented the Naïve Bayes Algorithm using full raw code, without any optimizations or library functions. There might be some mistakes in the implementation, which is why we received a low accuracy of **22.8 %**. The testing was done using only **500 test data**, due to the slowness of the algorithm and lack of proper hardware. Insha Allah, we will try to improve the algorithm to get better accuracy.

Technologies: Python, Pandas, Scikit-Learn, stop-words.

```
1  #finding accuracy using 500 test data
2
3  count = 0
4  for i in range(500):
5      if x[i]==y[i]:
6          count += 1
7
8  count
9
10 accuracy = (count/500)*100
11 accuracy
```

22.8

References:

1. <https://github.com/Jahid234/Newspaper-Article-Headline-Category-Prediction>
2. <https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/>
3. **Dataset** - <https://www.kaggle.com/datasets/rmisra/news-category-dataset>