626CDS- Lab Project Report

Classification of News Headlines

By

Basmah Misfer AlQahtani
443800986@kku.edu.sa

Maryam Mohmmad Alrashdi
443800993@kku.edu.sa

Submitted to

Associate Prof. Anandhavalli Muniasamy
anandhavalli@kku.edu.sa

King Khaled University
College of Computer Science
Department of Information Systems
Semester II - Year 2022

Table of Contents:

List of Figures:

1. Introduction

Text mining has been increasingly important during the last few years. Users can now access knowledge from a variety of sources, including electronic media, digital media, print media, and many more. Due to the widespread availability of text in a variety of formats, researchers have accumulated a large amount of unstructured data and have discovered several methods in the literature to turn this unstructured data into a defined structured volume, a process known as text categorization. In comparison to short-length text, complete text classification (whole news, enormous documents, lengthy length texts, etc.) is more prevalent.

Text classification datasets are used to sort natural language texts into content categories. Consider categorizing news stories by topic or categorizing book reviews according to a positive or negative response. Text classification is also useful for detecting languages, managing client feedback, and detecting fraud. Machine learning models can automate this process, which is time-consuming when done manually.

For news, categorization is a multi-label text classification issue. The purpose is to allocate a news story to one or more categories.

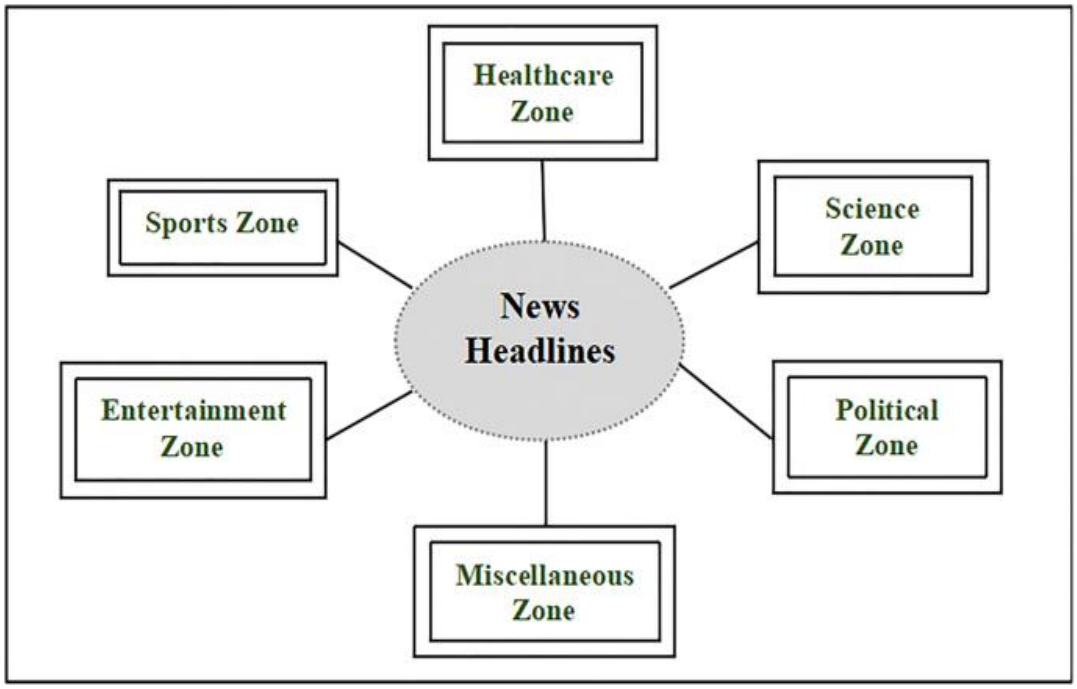

Figure-1: Classification of News Headlines

2. Project Aim

In this project, we aim to train a model that can properly categorize previously unknown news articles into some classes such as business, science and technology, entertainment, and health.

3. Description

The purpose of dividing news into multiple categories is to make the news reader's experience easier. When constructing a model that automatically classifies these news headlines based on their category, it also assists news website owners in the efficiency with which the content is published and classified. So, in this project, we started out to process a dataset including news and its categories, and then two models were used (Naïve Bayes and Logistic Regression) to train this data. As is common, this data was separated into a training and test dataset. Eventually, the performance of these two models was evaluated using accuracy as a matric.

4. Models Used

In this project, two models were used in a dataset for news classification.

4.1. Naïve Bayes

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e., every pair of features being classified is independent of each other.

4.2. Logistic Regression

A classification algorithm is logistic regression. It is used to forecast a result using a set of independent variables. To describe data and explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables, logistic regression is utilized.

5. Dataset Used

In this project, we used a database from Kaggle. This dataset contains headlines, URLs, and categories for 422,937 news stories collected by a web aggregator between March 10th, 2014, and August 10th, 2014.

News categories included in this dataset include business; science and technology; entertainment; and health. Different news articles that refer to the same news item (e.g., several articles about recently released employment statistics) are also categorized together.

The columns included in this dataset are:

- **ID:** the numeric ID of the article
- **TITLE:** the headline of the article
- **URL:** the URL of the article
- **PUBLISHER:** the publisher of the article
- **CATEGORY:** the category of the news item; one of:
  - b: business
  - t: science and technology
  - e: entertainment
  - m: health
- **STORY:** alphanumeric ID of the news story that the article discusses
- **HOSTNAME:** hostname where the article was posted
- **TIMESTAMP:** approximate timestamp of the article's publication, given in Unix time (seconds since midnight on Jan 1, 1970)

6. Workflow Explanation

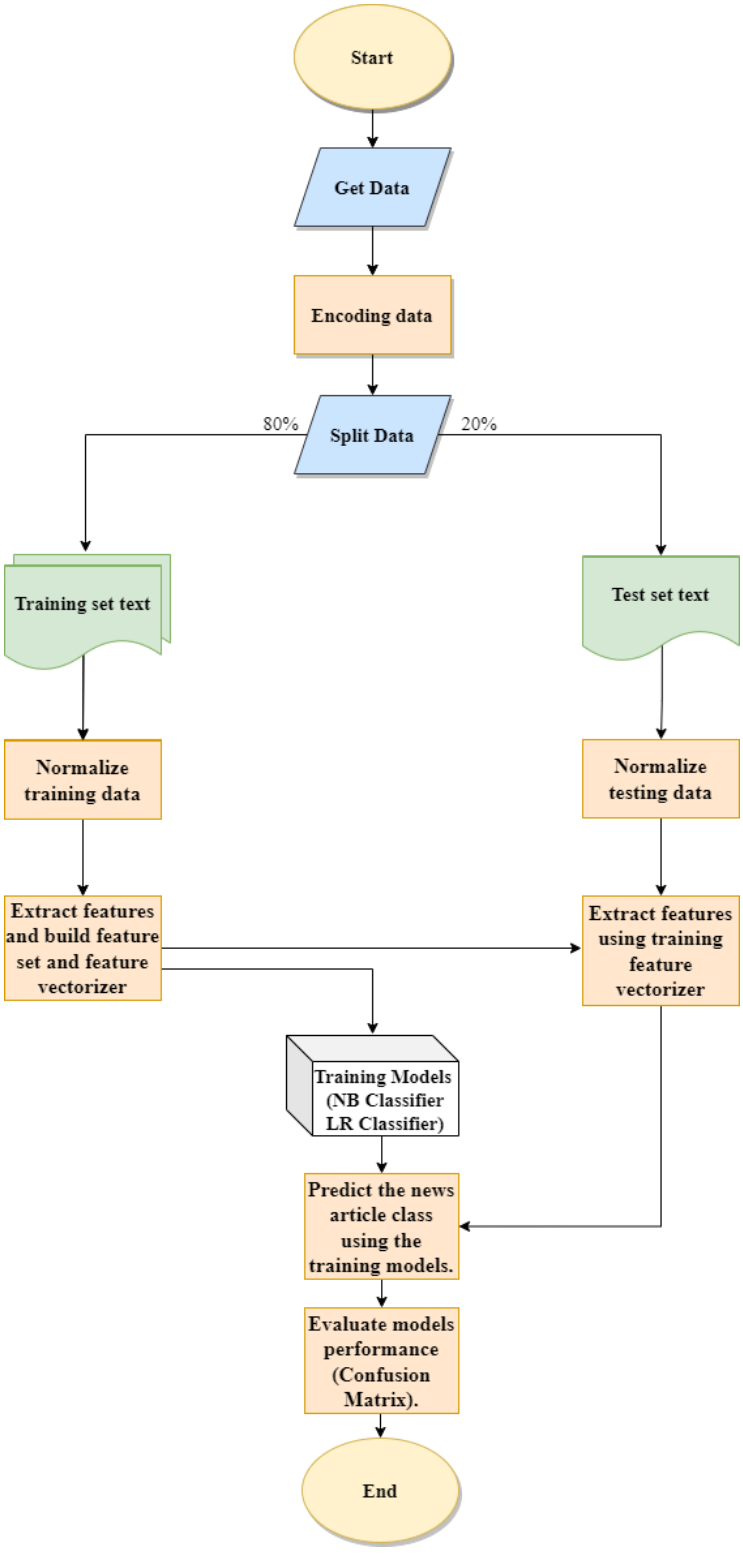

Figure-2: Workflow Explanation

After all preprocessing, the unwanted columns and keeping the only title and category columns are dispensed. Since our target labels in text, we need to convert it to numeric values for that we can use sklearn label encoder <<< class1, class2, class3 --> 1,2,3 >>.

6.1. Model Training:

a. Normalize training data

- Remove English stop words from the sentences.
- Remove punctuations from sentences.
- Stemming the words using the NLTK library.
- Clean numbers from sentences.
- Apply normalization to the dataset.

b. Extract features and build feature set and feature vectorizer

Feature Extraction Techniques:

The collection of text documents is converted to a matrix of token counts using count vectorize that produces a sparse representation of the counts.

1. Define feature vectorization.

- Define the sklearn pipeline.
  - **Step 1.** user counter vectorizer to vectorize each sentence based on vocabulary
  - **Step 2.** use the TF-IDF feature extraction method.
- Extract features from the training dataset and applies the sklearn pipeline.

c. Use supervised learning algorithm Naïve Bayes & Logistic Regression to build a predictive model

Model 1. Naïve Bayes classifier

- Train the model using extracted features and target label set
- Model performance evaluation on train set
- Analyze the results

Model 2. Logistic Regression classifier

- Train the model using extracted features and target label set.
- Model performance evaluation on train set
- Analyze the results

6.2. Model Testing:

a. Normalize testing data

- Apply normalization to the test dataset.

b. Extract features using training feature vectorizer

- Extract features from the test dataset

c. Predict the news article class using the training model.

- Take predictions from trained model naive Bayes classifier.
- Take predictions from trained model logistic regression.

d. Evaluate model performance.

- Performance analysis on naive Bayes model.
- Performance analysis on the logistic regression model.

7. Results & Discussion

1. Naïve Bayes:

To perform the model in the training dataset accuracy: 0.9287022652285203 while on the test dataset accuracy: 0.9208844278206524.

We can see here that there is no difference as the model performed well in the classification of the news as we can see other measures of the model's performance are very close.

	precision	recall	f1-score	support
accuracy	0.9287022652285203			
b	0.90	0.92	0.91	92774
e	0.95	0.97	0.96	121975
m	0.97	0.86	0.91	36511
t	0.91	0.91	0.91	86675
accuracy			0.93	337935
macro avg	0.93	0.91	0.92	337935
weighted avg	0.93	0.93	0.93	337935

Figure-3: Naïve Bayes model performance on the training dataset

Naive Bayes Model Performance Analysis

accuracy 0.9208844278206524

	precision	recall	f1-score	support
b	0.89	0.91	0.90	23193
e	0.94	0.97	0.96	30494
m	0.97	0.84	0.90	9128
t	0.90	0.90	0.90	21669
accuracy			0.92	84484
macro avg	0.93	0.90	0.91	84484
weighted avg	0.92	0.92	0.92	84484

Figure-4: Naïve Bayes model performance on the test dataset

2. Logistic Regression:

To perform the model in the training dataset accuracy: 0.9534259546954296

While you are in the test dataset accuracy: 0.9407935230339473

We can see here that there is very little difference, but the model performed well in the news classification as we can see other metrics of the model's performance very closely.

accuracy 0.9534259546954296

	precision	recall	f1-score	support
b	0.93	0.94	0.94	92774
e	0.97	0.98	0.98	121975
m	0.97	0.93	0.95	36511
t	0.94	0.94	0.94	86675
accuracy			0.95	337935
macro avg	0.95	0.95	0.95	337935
weighted avg	0.95	0.95	0.95	337935

Figure-5: Logistic Regression model performance on the training dataset

Logistic Regression Model Performance Analysis

accuracy 0.9407935230339473

	precision	recall	f1-score	support
b	0.92	0.92	0.92	23193
e	0.96	0.98	0.97	30494
m	0.96	0.91	0.93	9128
t	0.93	0.92	0.92	21669
accuracy			0.94	84484
macro avg	0.94	0.93	0.94	84484
weighted avg	0.94	0.94	0.94	84484

Figure-6: Logistic Regression model performance on the test dataset

8. CONCLUSION

In the end, we can say that both Naïve Bayes and Logistic Regression models are suitable to be used for classifying news by category, although in the recent period the classification of short news remains a challenge, these models handle classification well so if it is developed it will be classifying even short news by it.

9. REFERENCES

- Naive Bayes Classifiers. (2022, Feb 02). Retrieved from geeksforgeeks: https://www.geeksforgeeks.org/
- News Aggregator Dataset. (2014). Retrieved from https://www.kaggle.com/
- Rana, M. I., S. K., & Akbar, U. M. (2014, December). News classification based on their headlines: A review. Retrieved from researchgate: https://www.researchgate.net/
- SINGH, D. (2021, December 27). Text Classification of News Articles. Retrieved from analytics vidhya: https://www.analyticsvidhya.com

10. Appendix: Program Code File

https://drive.google.com/file/d/1ohyD7pzz7ngll2b7glY0T0PZJOtOK0-h/view?usp=sharing