Classification of News Headlines

Presented By:
Basmah AlQahtani
Mariam Alrashdi

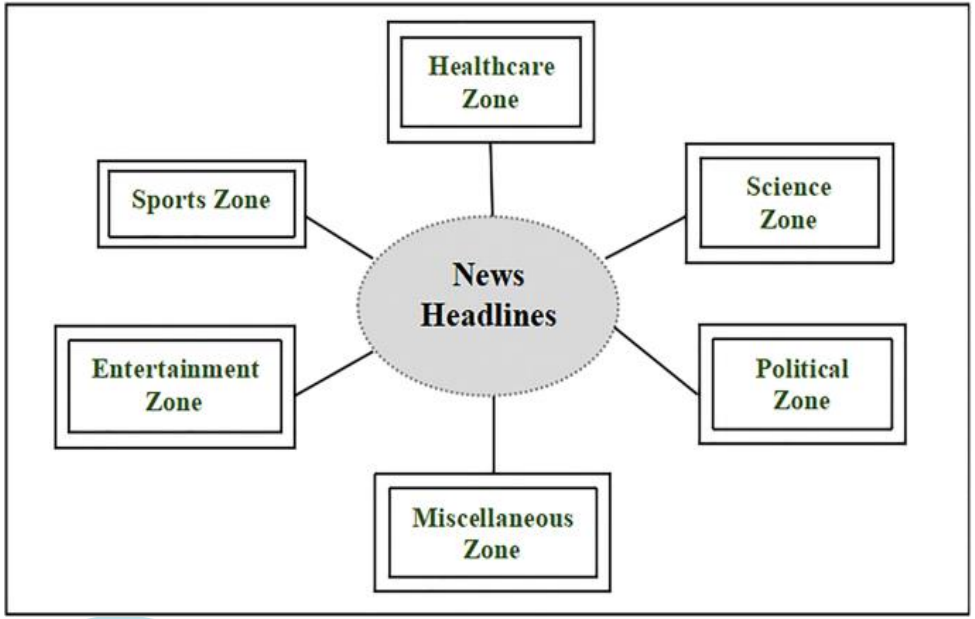
We will discuss ...

1. Introduction
2. Project Aim
3. Description
4. Models Used
5. Dataset Used
6. Workflow Explanation
7. Results & Discussion
8. Conclusion

01

Introduction

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



02

Project Aim

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.

03

Description

Description ...

The purpose of dividing news into multiple categories is to make the news reader's experience easier.

We construct a model that automatically classifies these news headlines based on their category.

Description ...

Preprocessing the dataset

Machine Learning

04

Model Used

Model used ...

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

Naïve Bayes	Logistic Regression
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.	A Classification algorithm 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.

05

Data Used

Data used...

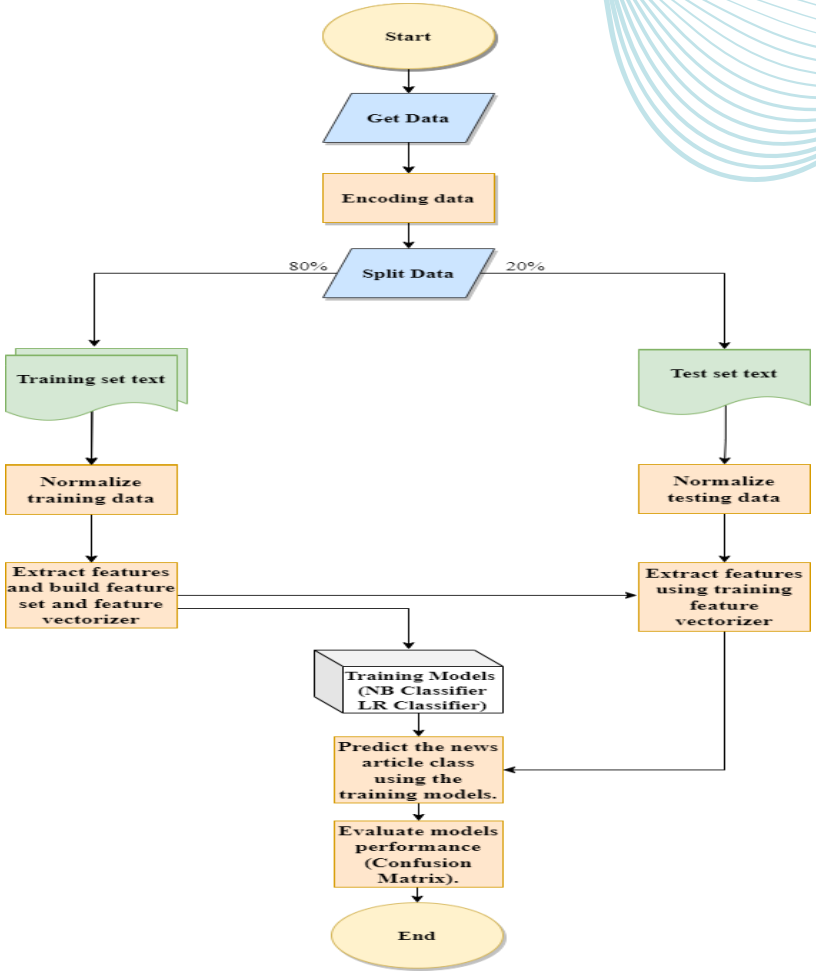
we used a database from Kaggle. This dataset contains headlines, URLs, and categories for 422,937 news stories collected by a web aggregator between March 10th, 2014, and August 10th, 2014.

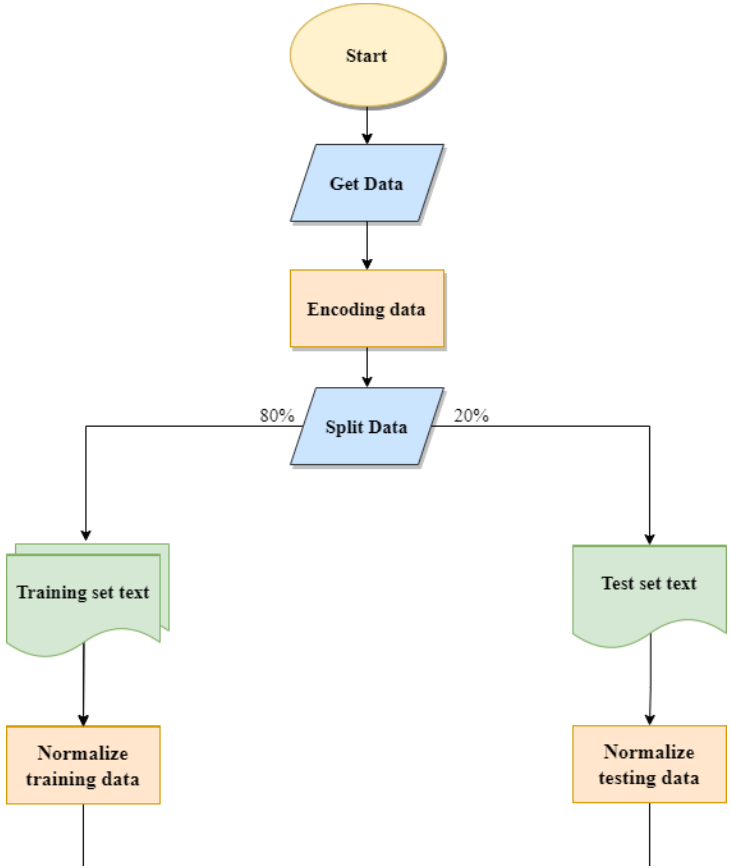
The columns included in this dataset are:

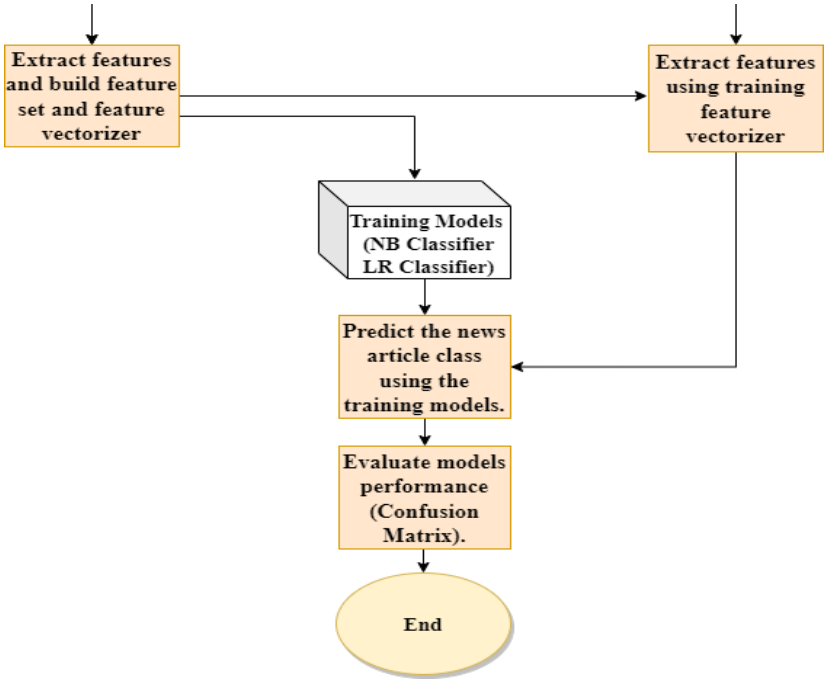
- **ID:** the numeric ID of the article
- **TITLE:** the headline of the article
- **URL:** the URL of the article
- **PUBLISHER:** the publisher of the article
- **CATEGORY:** the category of the news item; one of:
    - b: business
    - t: science and technology
    - e: entertainment
    - m: health
- **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)

06

Workflow Explanation


Start
Get Data
Encoding data
80%
Split Data
20%
Training set text
Test set text
Normalize training data
Normalize testing data
Extract features and build feature set and feature vectorizer
Extract features using training feature vectorizer
Training Models (NB Classifier LR Classifier)
Predict the news article class using the training models.
Evaluate models performance (Confusion Matrix).
End


Start
Get Data
Encoding data
Split Data
80%
20%
Training set text
Test set text
Normalize training data
Normalize testing data


Extract features and build feature set and feature vectorizer
Extract features using training feature vectorizer
Training Models (NB Classifier LR Classifier)
Predict the news article class using the training models.
Evaluate models performance (Confusion Matrix).
End

07

Results & Discussions

Results …

1- Naïve Bayes:

Naïve Bayes model performance on the training dataset:

accuracy 0.9287022652285203

	precision	recall	f1-score	support
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

Results …

1- Naïve Bayes:

Naïve Bayes model performance on the test 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

Results …

2- Logistic Regression:

Logistic Regression model performance on the training dataset:

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

Results …

2- Logistic Regression:

Logistic Regression model performance on test 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


Conclusion

Thank you

Basmah AlQahtani
443800986@kku.edu.sa
Mariam AlRashdi
443800993@kku.edu.sa
10 May 2022