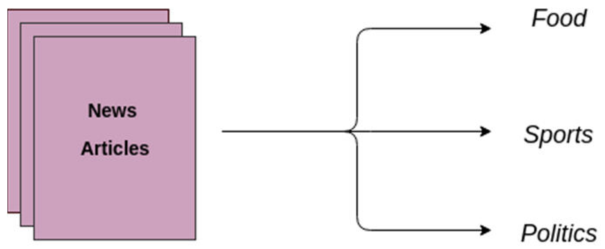
# Text Classification

# About the Module

- ❑ Text Classification Task
- ❑ Dataset Preparation
- ❑ Feature Extractor
- ❑ Classification Approaches

# Text Classification Task



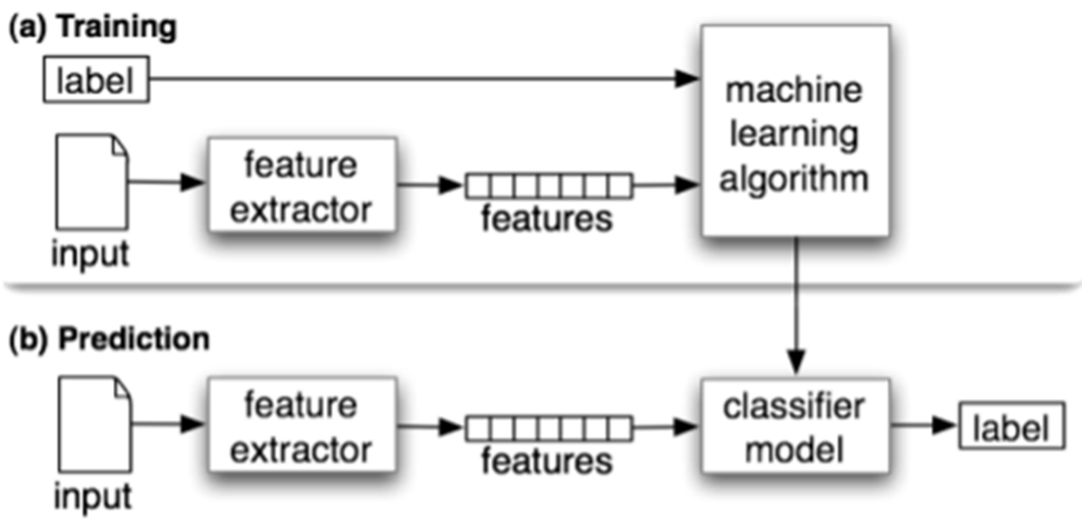

- Technique to systematically classify text object (document or sentence) in a fixed category
- Helpful in organizing, information filtering, and storage purposes

Examples

- Sentiment Analysis
- Email Spam Classification
- Author Identification from Articles
- News Topic Classification

Analytics Vidhya
Learn everything about analytics

# Text Classification Task




Analytics Vidhya
Learn everything about analytics

# Dataset Preparation

Text Cleaning

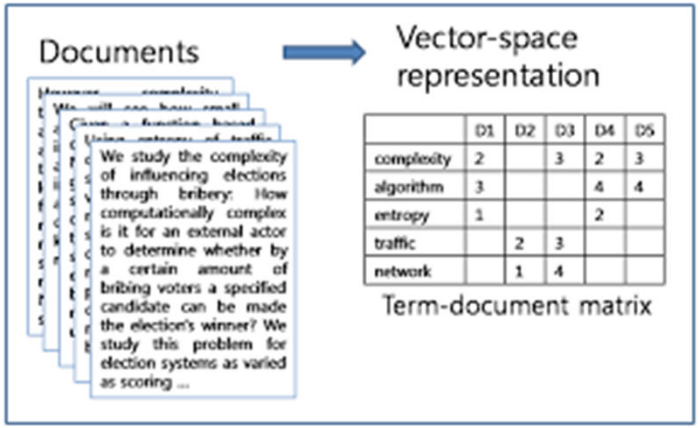
- Removal of Stop words
- Keyword Lemmatization / Stemming
- Removal of Punctuations

Target : Label Encoding

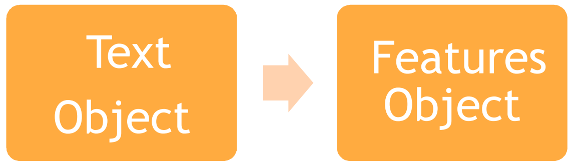
Train Test Validation Split

# Feature Extractor

- Count Features
- TF IDF Features
- Word Embedding Features
- Meta Features
- Topic Models as Features




| | D1 | D2 | D3 | D4 | D5 |
|---|---|---|---|---|---|
| complexity | 2 | | 3 | 2 | 3 |
| algorithm | 3 | | | 4 | 4 |
| entropy | 1 | | | 2 | |
| traffic | | 2 | 3 | | |
| network | | 1 | 4 | | |




Analytics Vidhya
Learn everything about analytics

# Classification Models

Rule Based
- Hand Crafted Rules

Probability Based
- Naïve Bayes

Learning Based
- Logistic Regression
- State Vector Machines
- Ensemble Models

Deep Learning Based
- Convolutional Neural Networks
- Recurrent Neural Networks
- Hybrid Deep Neural Networks

# Rule Based Text Classification

- Prepare rules to classify the text

Example Rules :

I. Classify text objects based on number of words

II. Classify text objects on the presence of certain words

III. Classify text objects based on grammar rules and part of speech tags

- Accuracy can be high if rules are highly refined

- Maintenace and Building these rules is expensive

# Naïve Bayes Text Classification

- Classification based on Bayesian theorem, Using Prior probabilities to classify new text

$$p(A|B) = \frac{p(B|A) * p(A)}{p(B)}$$

- P(A | B) : the likelihood of event A occurring given that B is true
- P(B | A) : the likelihood of event B occurring given that A is true
- P(A) and P(B) are the independent probabilities of observing A and B

**Example : Detecting if an email is spam / not spam**

- P(spam | w1,w2,w3)) = (P(spam) * P(w1,w2,w3)| spam)) / P(w1,w2,w3))

Analytics Vidhya
Learn everything about analytics

# Other Classification Models

***K-Nearest Neighbors***

- Finds the minimum distance of the given text document in the entire data space
- Assigns the label with majority voting

***Logistic Regression***

- Finds the likelihood of a given text document to lie between 0 and 1

***State Vector Machine ( SVM)***

- Particularly good for very sparse data in very high dimensional spaces

# Ensemble Methods

- Simple models often suffers from Bias and Variance errors

Bias : How much does the model is far away from the actual truth
Variance : How much does the model output change with different training data

- High bias leads to Underfitting, high variance leads to overfitting

Ensemble Models
- Bagging : Extra Trees Classifiers, Random Forests
- Results are averaged in bagging

- Boosting : XgBoost, Lightgbm, Catboost
- Results are sequentially improved in boosting

# Enhancing Text Classification Pipeline

Handling dataset imbalance

Improved Text Cleaning

Improved Feature Engineering
- NGrams as Features
- Part of Speech Tags as Features
- Grammar Relations as Features
- Topic Models as Features

Model Tuning

Model Stacking