GALLOGLY COLLEGE OF ENGINEERING
SCHOOL OF COMPUTER SCIENCE
The UNIVERSITY *of* OKLAHOMA

Text Analysis

Text classification

DATA BITE
A bit on Big Data

Presented by: Keerti Banweer

GALLOGLY COLLEGE OF ENGINEERING
SCHOOL OF COMPUTER SCIENCE
The UNIVERSITY *of* OKLAHOMA

DATA BITE
A bit on Big Data

Would you like to be involved in research at the University of Oklahoma?

The purpose of this survey is to understand attendees' knowledge of Python and intended outcomes from attending these sessions. We are trying to gauge the demand of skills, resources, and knowledge of Python programming and associated techniques. This will, in turn, help us understand the success of the workshop and room for improvement.

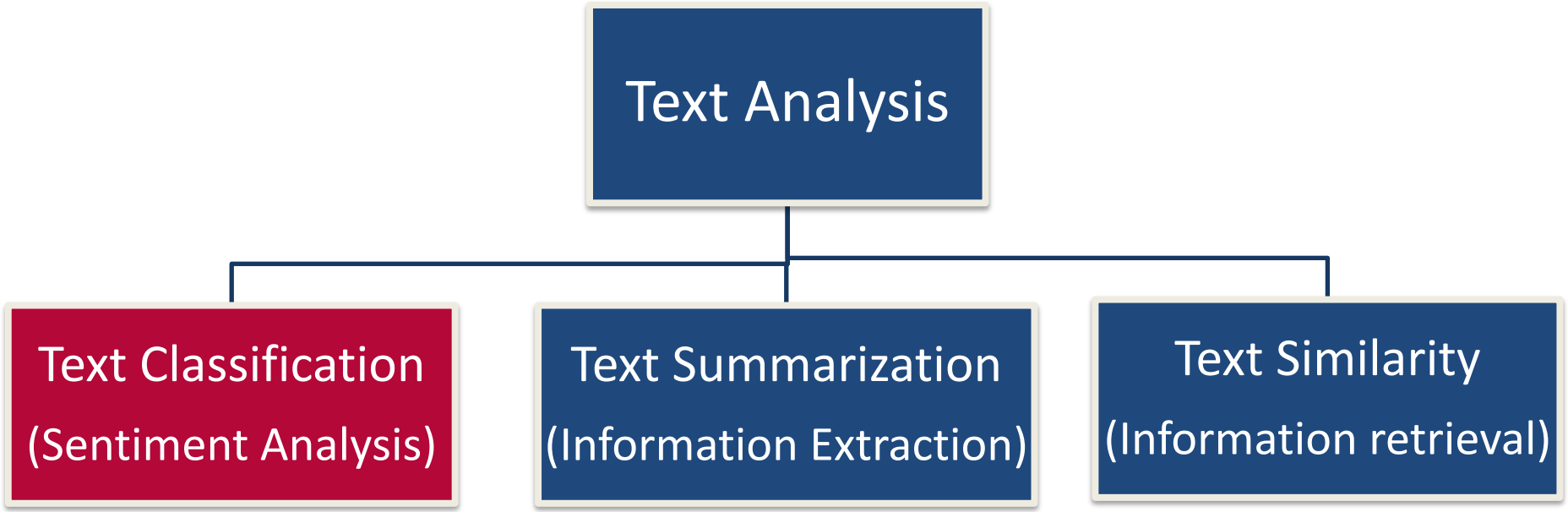
SCAN ME

Overview

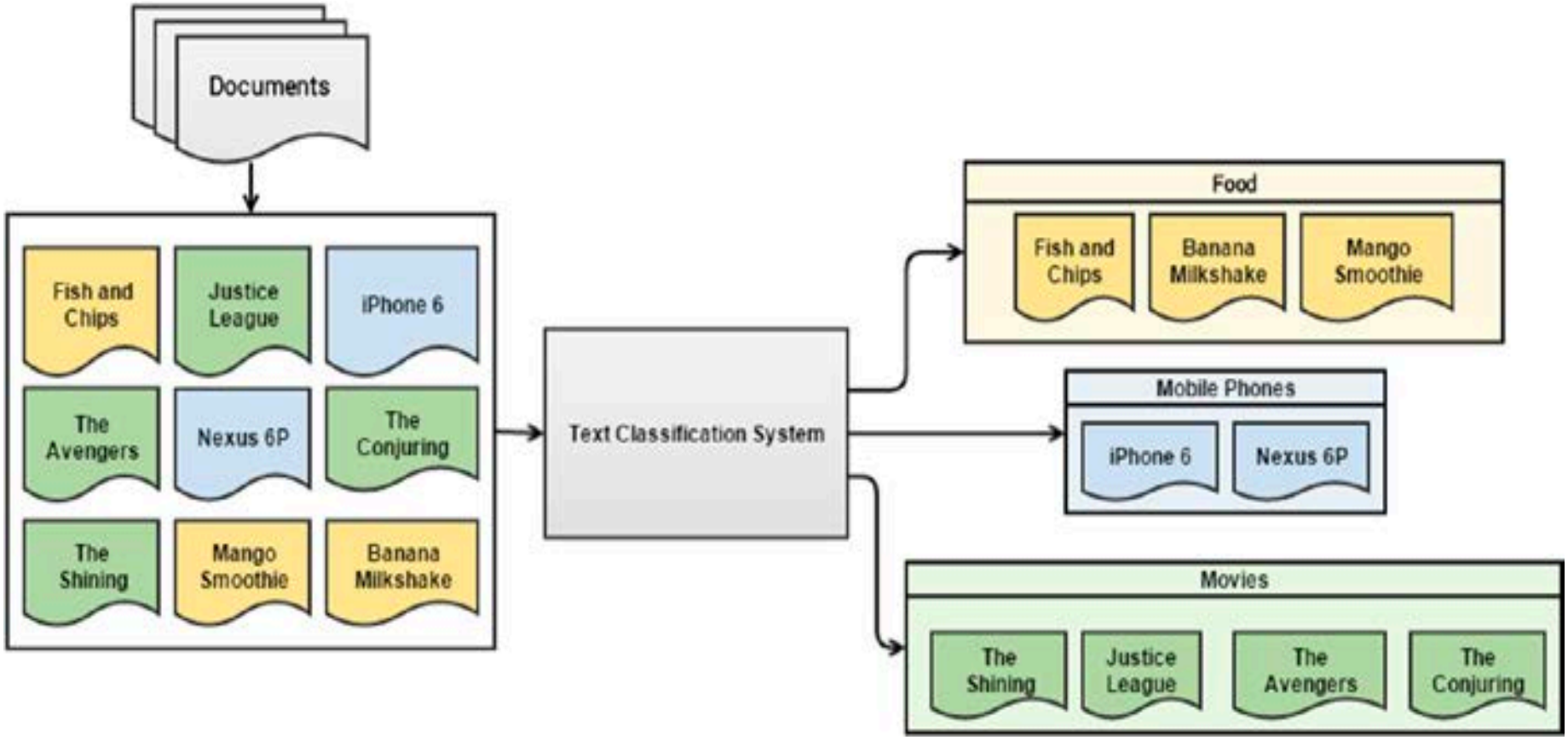
- Text Analysis
- What is Text Classification?
- Machine learning approach
- Types of Text Classification
- Workflow
- Demo

Text Analysis

Analyze complex textual data and extract meaningful patterns and insights from it.



What is Text Classification?



Source: *Text Analytics With Python : A Practical Real-World Approach To Gaining Actionable Insights From Your Data,* **Second Edition**, Dipanjan Sarkar, 2019, Apress L. P. (ISBN 978-1-4842-4353-4) (eBook ISBN 978-1-4842-4354-1) Publisher Website

Machine Learning Approach

Supervised Learning

- Requires labeled training data
- Each data point corresponds to a class or label
- Learns meaningful patterns for each type of class
  - Sentiment analysis
  - Spam filtering
  - Topic labeling

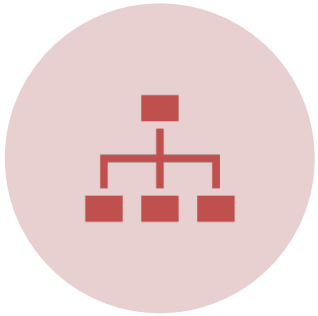
Unsupervised Learning

- Do not require any labeled training data
- Extract meaningful patterns from the data
  - Document Summarization
  - Similarity Analysis
  - Clustering (Useful in Text document categorization)

Types of Classification

BINARY CLASSIFICATION

MULTI-CLASS CLASSIFICATION

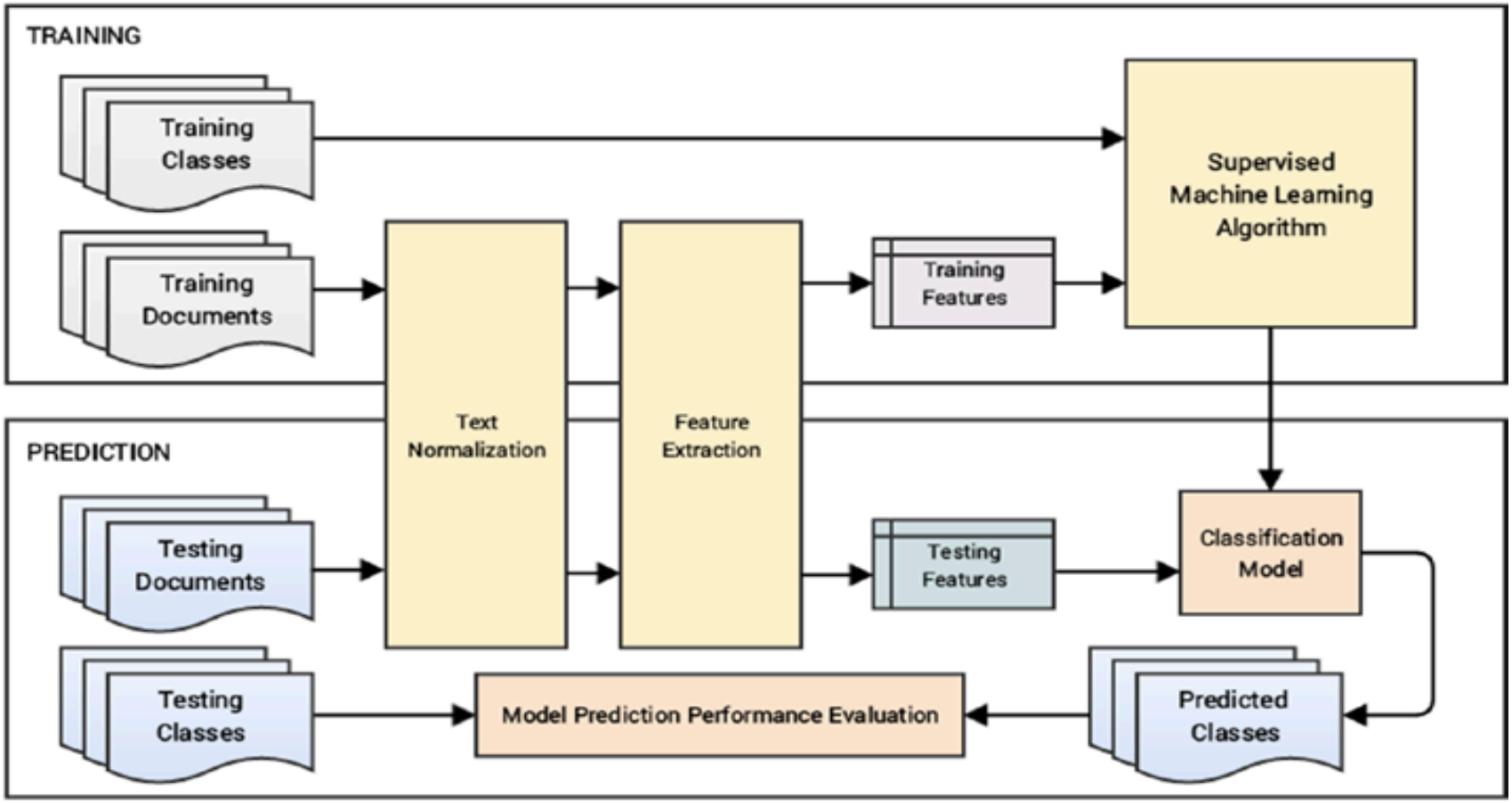


MULTI-LABEL CLASSIFICATION

Text Classification Workflow

- A typical workflow for a text classification system is as follows:
  - Train and test datasets
  - Text Normalization
  - Feature Extraction
  - Model Training
  - Model Prediction

Architecture



Source: *Text Analytics With Python : A Practical Real-World Approach To Gaining Actionable Insights From Your Data,* **Second Edition**, Dipanjan Sarkar, 2019, Apress L. P. (ISBN 978-1-4842-4353-4) (eBook ISBN 978-1-4842-4354-1) Publisher Website

Text Normalization

Tokenization – Segmenting a sentence into words

Named Entity Recognition – like persons, locations or companies

Word vectors – represent the meaning of the word in form of vectors of numbers

Similarity – Comparing words or documents

Bag of Words, TF-IDF, and Advanced word vectorization models

FEATURE EXTRACTION

Bag of Words model

- Convert text document into vectors
- Vector represents the frequency of all the distinct words for specific document
- The frequency of occurrence is the weight of each word

Example

```
from sklearn.feature_extraction.text import CountVectorizer
vectorizer = CountVectorizer(min_df=1, ngram_range=(1,1))
features = vectorizer.fit_transform(corpus)
vectorizer.get_feature_names() ##print the features names


##Corpus is a collection of text:
CORPUS = ['the sky is so beautiful’,
            'sky is blue and sky is beautiful’,
            'the sky is perfectly blue’,
            'i love cream cheese’
            ]
```

Useful link:

https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html#sklearn.feature_extraction.text.CountVectorizer

TF-IDF Model

- Tf – term frequency
- Idf – inverse document frequency
- TF-IDF is a combination of the above two metrics

Example

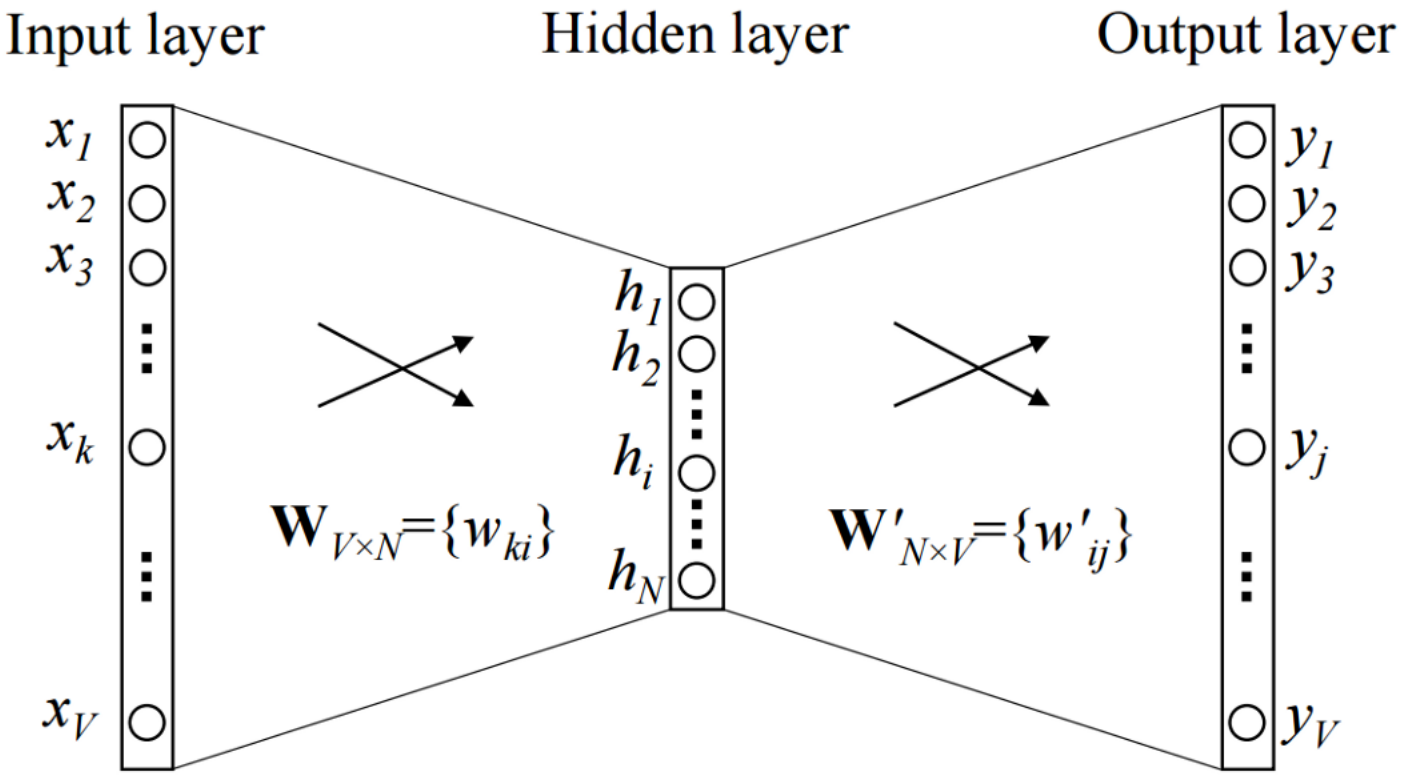
```
from sklearn.feature_extraction.text import
TfidfTransformer

transformer = TfidfTransformer(norm='l2’,
                               smooth_idf=True,
                               use_idf=True)
# fit to data and then transform it
tfidf_matrix =transformer.fit_transform(corpus)
```

Advanced word vectorization models

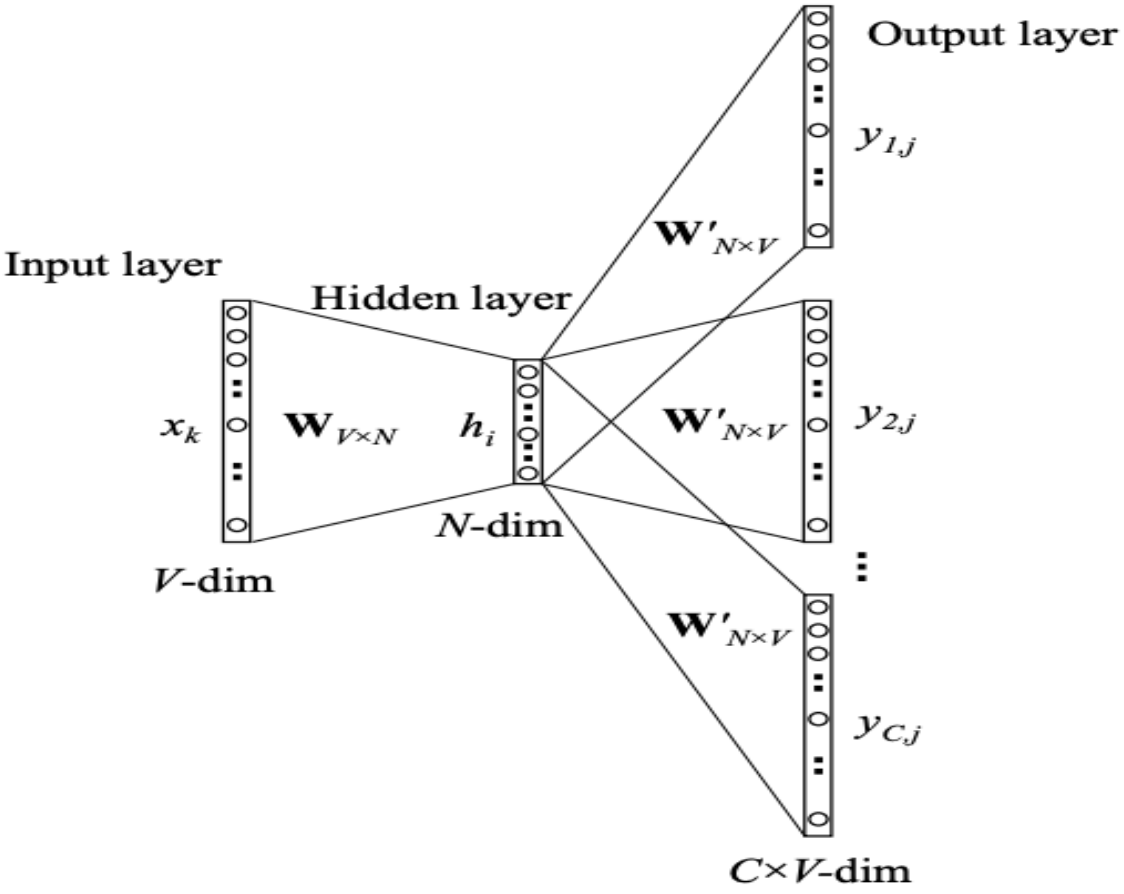
- Word2Vec – learn word embeddings
- It is a neural network-based implementation that learns distributed vector representations of words

Source: Rong, X. (2014). word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*.

CBOW model


Input layer
Hidden layer
Output layer
x_1
x_2
x_3
x_k
x_V
h_1
h_2
h_i
h_N
y_1
y_2
y_3
y_j
y_V
$\mathbf{W}_{V\times N}=\{w_{ki}\}$
$\mathbf{W}'_{N\times V}=\{w'_{ij}\}$

Skip gram model



Multinomial Naïve Bayes and Support Vector Machines

CLASSIFICATION ALGORITHMS

Multinomial Naïve Bayes

- Scikit-learn library provides MultinomialNB

Support Vector Machines

Supervised learning algorithm

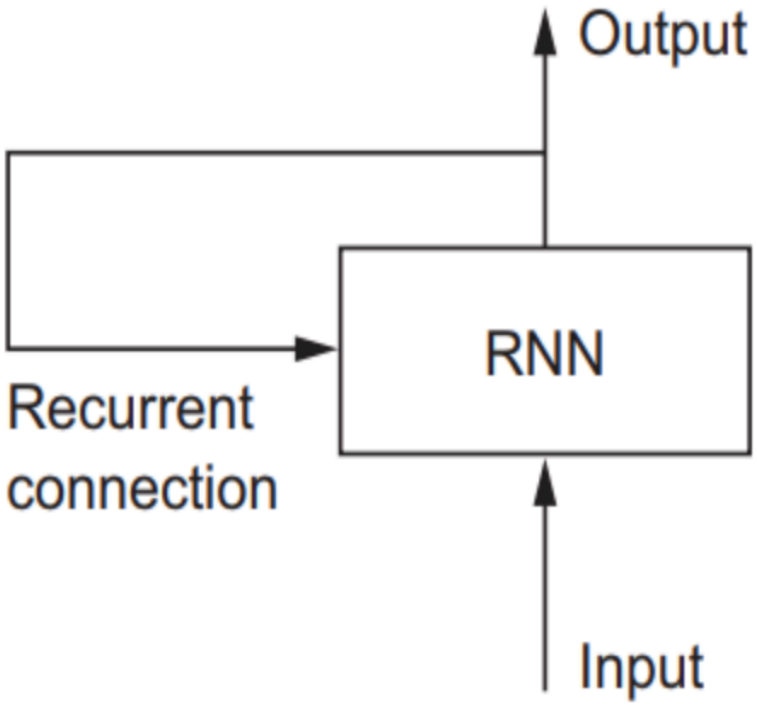
SVM constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outliers detection

Long Short-Term Memory: Allows access to previous information to used at later time

LSTM

RNNs

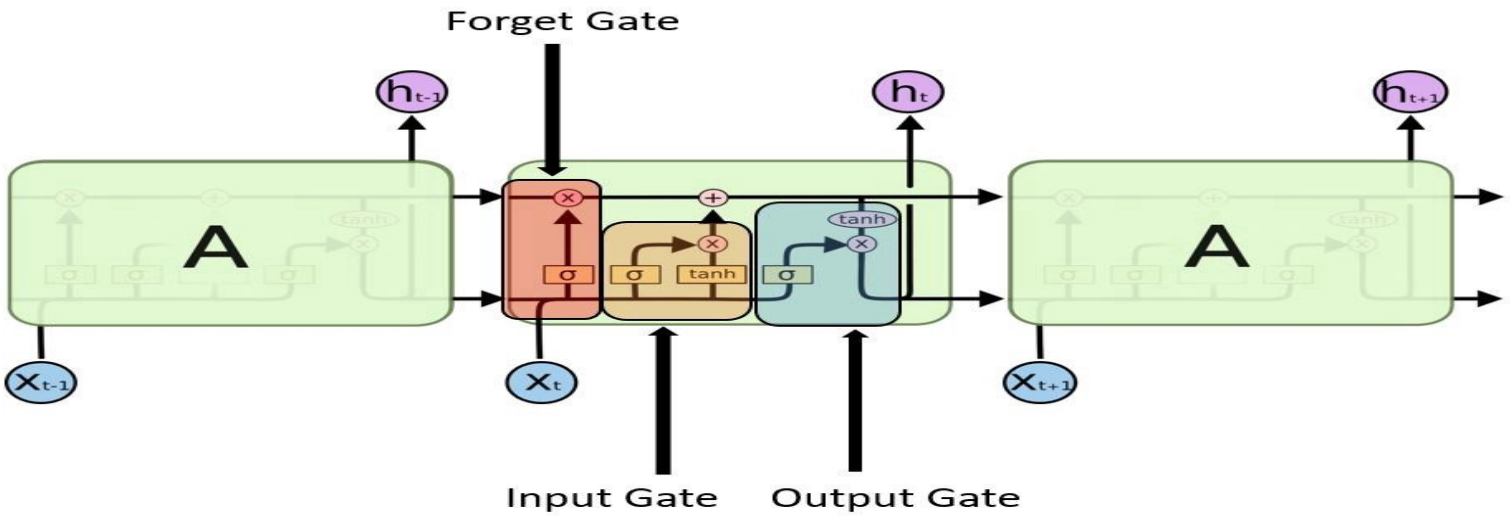
- Recurrent Neural Networks
- Is a generalization of feedforward neural network that has an internal memory
- RNNs can use their internal state (memory) to process sequences of inputs.



Source: https://www.manning.com/books/deep-learning-with-python

LSTMs

- It remembers the previous data



Source: https://towardsdatascience.com/understanding-rnn-and-lstm-f7cdf6dfc14e

Demo

- Implement Feature Extraction using TF-IDF
- Text classification using existing packages in Scikit learn
- Sentiment analysis of imdb dataset using Tensorflow Keras
- Text categorization using Clustering

GALLOGLY COLLEGE OF ENGINEERING
SCHOOL OF COMPUTER SCIENCE
The UNIVERSITY *of* OKLAHOMA

Questions?

Thank you!

GALLOGLY COLLEGE OF ENGINEERING
SCHOOL OF COMPUTER SCIENCE
The UNIVERSITY *of* OKLAHOMA

DATA BITE
A bit on Big Data

Would you like to be involved in research at the University of Oklahoma?

The purpose of this survey is to understand attendees' knowledge of Python and intended outcomes from attending these sessions. We are trying to gauge the demand of skills, resources, and knowledge of Python programming and associated techniques. This will, in turn, help us understand the success of the workshop and room for improvement.

SCAN ME