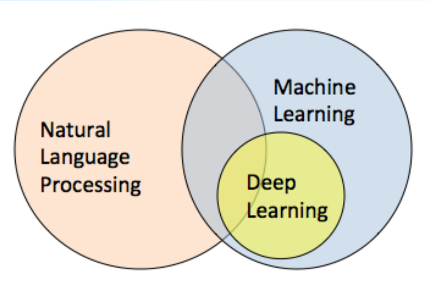


Arabic Sentiment Analysis Using LSTM

Faculty of Graduate Studies for Statistical Research– Cairo University

Under the Supervision of:

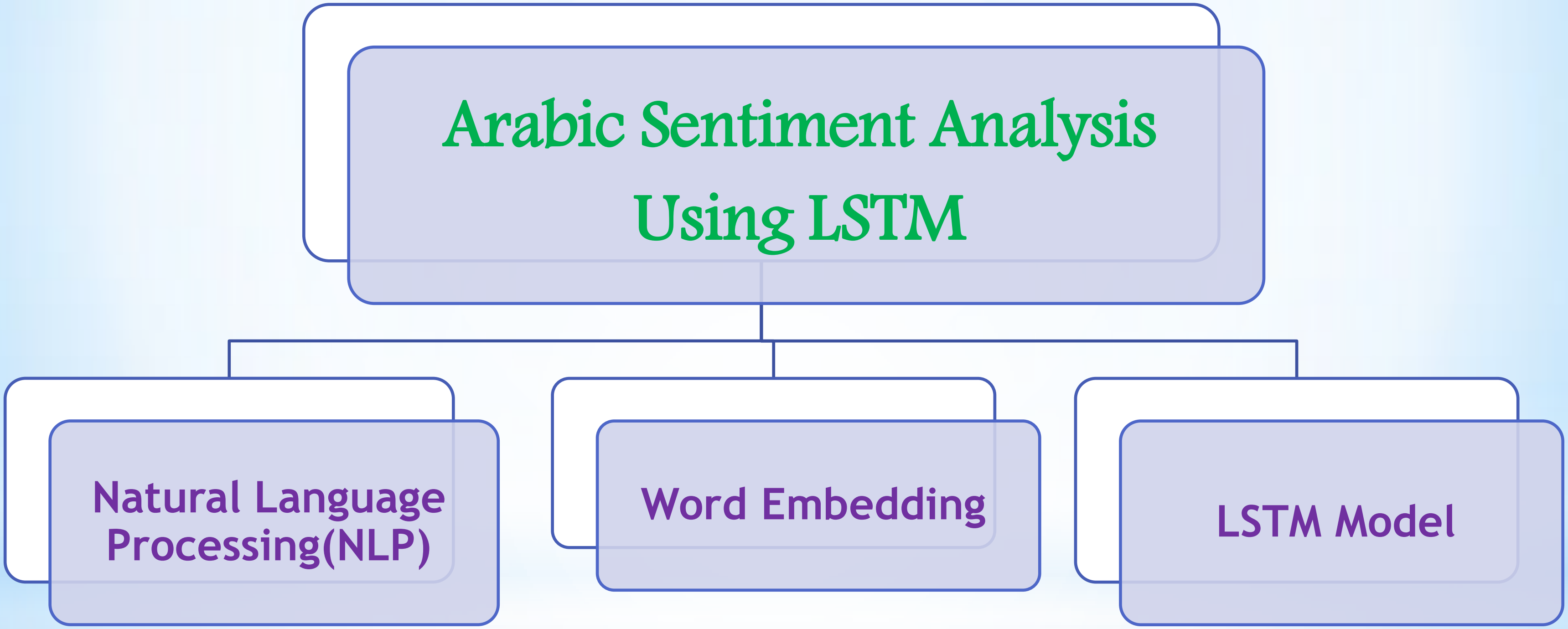
Associate Prof .Ammar Mohamed

Rania Abd EL Monam Kora

13 /7/2019

Outline

1. Introduction.
2. Sentiment analysis.
3. Characteristics of the Arabic language.
4. Challenges of colloquial Arabic language.
5. Word Embedding.
6. Deep learning techniques (LSTM).
7. LSTM code



Arabic Sentiment Analysis Using LSTM

Natural Language Processing(NLP)

Word Embedding

LSTM Model

Introduction

- Social media are considered an excellent source of information and can provide opinions, thoughts, and insights toward various important topics.

- Social media is generating a large volume of sentiment rich data in the form of tweets, status updates, blog posts, comments, reviews, etc.

- Analyzing these data can be very useful for understanding and better decision making. So there is a need to automate this, various sentiment analysis techniques are widely used.

Sentiment Analysis (SA)

- SA (also known as opinion mining or review mining or Sentiment mining).

- SA is a field of Natural Language Processing (NLP).

- SA is the process of determining whether a piece of writing is positive ☺ or negative ☹.

- SA involves classifying opinions in text into categories like "positive" or "negative" or "neutral" or more (e.g., positive, neutral, negative, very positive and very negative).

- SA can be divided into three levels, namely document level, sentence level, and aspect level (also known as word- or feature- level).

- SA helps in achieving various goals like observing public mood regarding political topics, market intelligence, new product sales prediction, the measurement of customer satisfaction and many more. Hence, sentiment analysis plays a significant role in our daily decision making process.

Sentiment Analysis for Arabic language

SA on the Arabic Language is still limited and is considered a challenging work due to several reasons:

- Arabic language has very complex morphological structures.

- The Most Arab users use colloquial Arabic instead of Modern Standard Arabic (MSA).

(colloquial Arabic varies from region to another region, from country to other countries, even from even state to another.)

- The lack of available Arabic corpora.

Characteristics of the Arabic language

- ❑ Arabic is Semitic language, and it's written from right to left.
- ❑Arabic language is one of six official languages of the United Nations.
- ❑ Arabic is the official language of 21 countries, and it's the major language in several areas of the world.
- ❑ Arabic language is classified into three types;
- ✓ Classical Arabic (book of Islam "AL-Qur'an")
- ✓ Modern Standard Arabic (formal communications, television, radio, news, education)
- ✓ Colloquial Arabic (shopping, chatting or in their homes.)

Challenges of colloquial Arabic language

Analyzing tweets composed in Arabic is a particularly challenging task due to:

❑ Unstructured language (informal language).

❑ Spelling inconsistencies.

❑ Slang words. Such as (كدا, ولو , كويس, يلا)

❑ Idiomatic expressions.

Such as (الطيور على اشكالها تقع, التكرار يعلم الحمار, بعد ما شاب ودوه الكتّاب,
هاك الشبل من ذاك الأسد, شحات ونزهي, طباخ السم بيدوقه)

❑ Colloquial expressions.

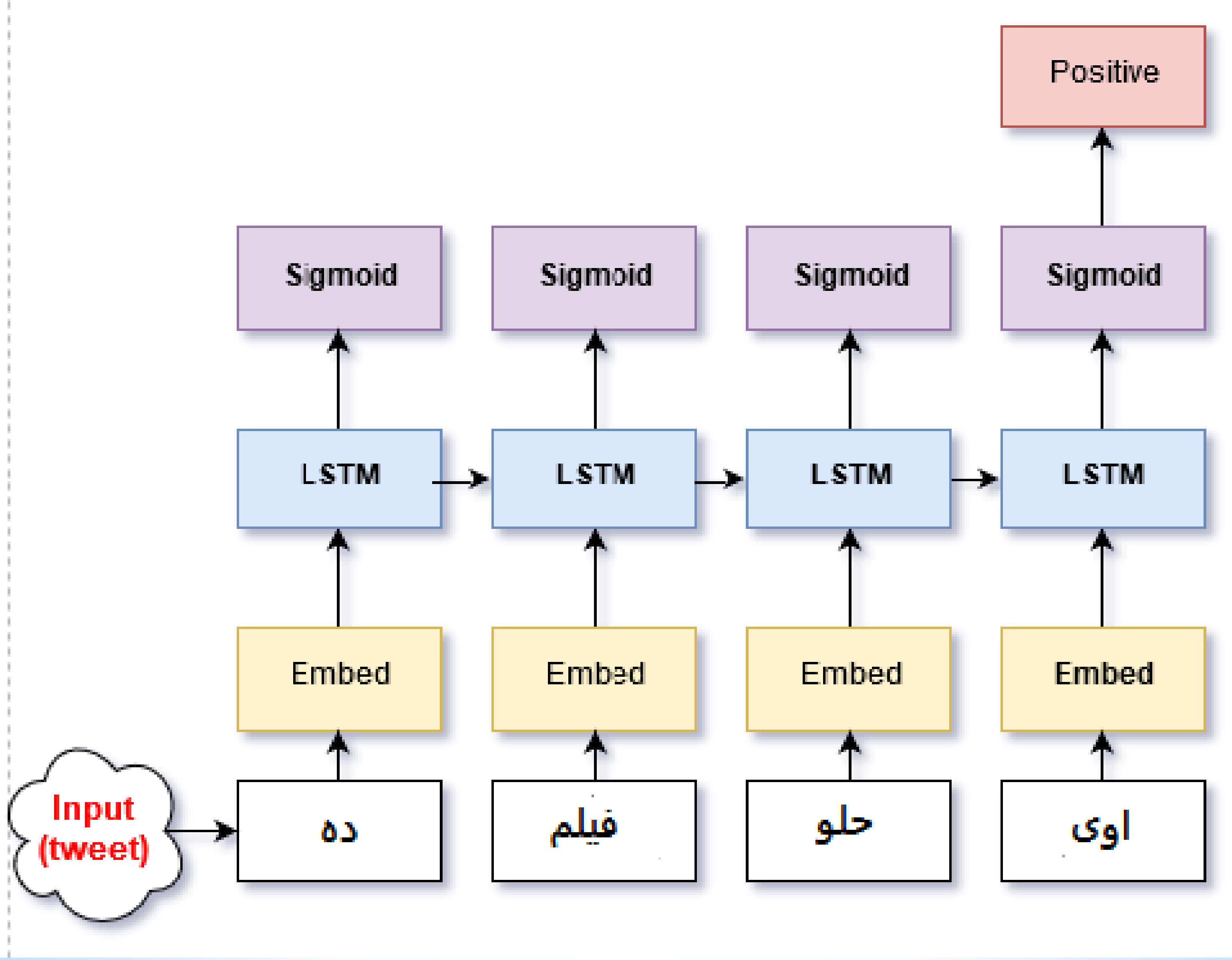
Such as (ناس مابتجيش الا بالعين الحمرا, مفيش فايدة, في المشمش, يا خبر أبيض,
يا سلام, زي السمن على العسل, يسرق الكحل من العين, زي القطط تاكل وتنكر)

❑ Emoticons. Such as (🙂, 😀 , 😢 , 😅 , 😂)

❑The tendency to repeat letters in writing to convey feelings.

Such as(كتيرررررررررررر)

LSTM code



Sentiment Analysis with LSTM

1-Load Data (corpus)

- Consist of : Colloquial Arabic language.
- Text + label (pos., neg.,neutral).
- Excel sheet.

Libraries can be used in python

- ✓ pandas
- ✓ Numpy
- ✓ xlrd
- ✓ keras
- ✓ Matplotlib

2- preprocessing (Handling data)

- Tokenization
- Normalization
- Stop words
- Stemming
- Part of speech

- Tokenization is breaking the sentence into words and punctuation, and it is the first step to processing text.

Includes:

- ✓ Sent tokenize
- ✓ Word tokenize

- Normalization is the process that consist of:

  - Remove any punctuation such as (" ",' ').
  - Remove any short vowels such as(ٌّـِ ُ).
  - Remove non-letters such as (% # $).
  - Replace آ،أ،إ with bare alif (ا).
  - Remove Tatweel ___such as (كثيـــــــــر) becomes   (كتير).
  - lastly replace final such as: ("ى" with "ي", "ة" with   "ه" , "ء" with" ا ", "ؤ" with "و", "ئ " or "ىء "with "ي") .

- Stop words: include words that do not of themselves confer much semantic value.

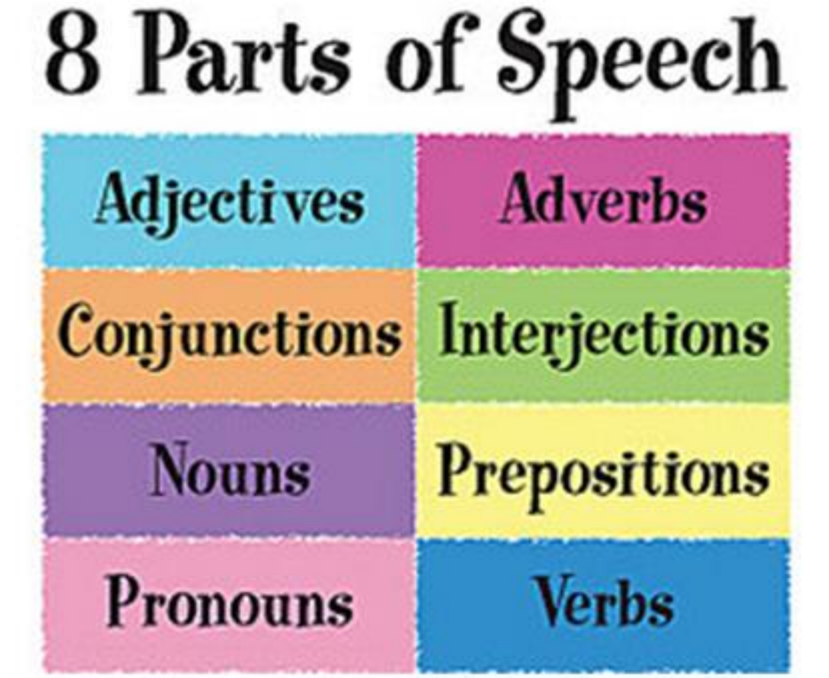
Any text may contain stop words like :

('من', 'هى', 'فى', ' الى', 'انت' , ' اما')

Stop words can be filtered from the text to be processed.

- Parts Of Speech (POS) is a way to describe the function of words.
- POS is the meaning of relationship with adjacent and related words in a phrase, sentence, or paragraph.

The identification of words as noun, verb, adjective, adverb, pronoun, preposition, conjunction, and interjection.



8 Parts of Speech

Adjectives	Adverbs
Conjunctions	Interjections
Nouns	Prepositions
Pronouns	Verbs

- Stemming is the process of reducing inflected words to their word stem.

Includes:

✓ Stemmend word

✓ Stemmend sent

Example:

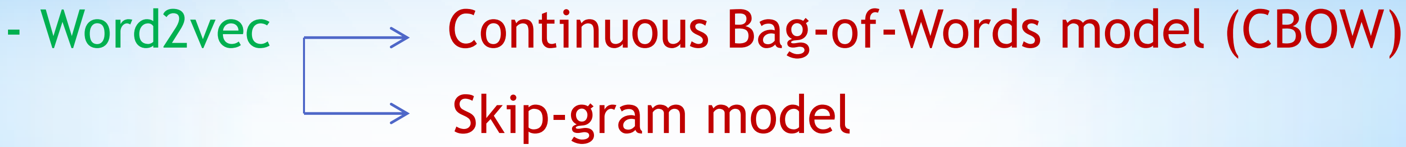
Root (المنظمات)- >نظم , Stem (المنظمات)- >منظم

3- Word Embedding

In very simplistic terms, Word Embedding are the texts converted into vectors.

The different types of word embeddings can be broadly classified into two categories:

- Frequency based Embedding
  - Count Vector
  - TF-IDF Vector
  - Co-Occurrence Vector
- Prediction based Embedding



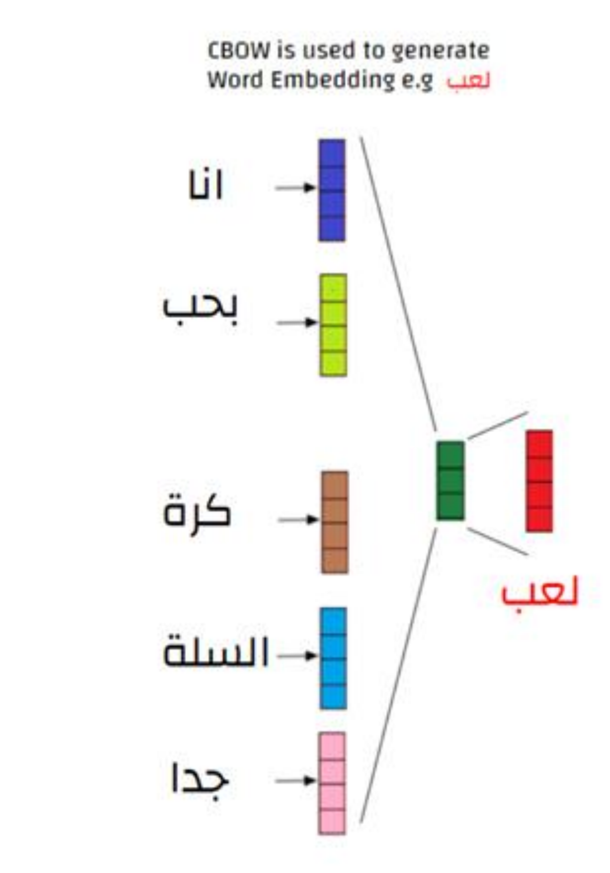
- Word2vec

Continuous Bag-of-Words model (CBOW)

Skip-gram model

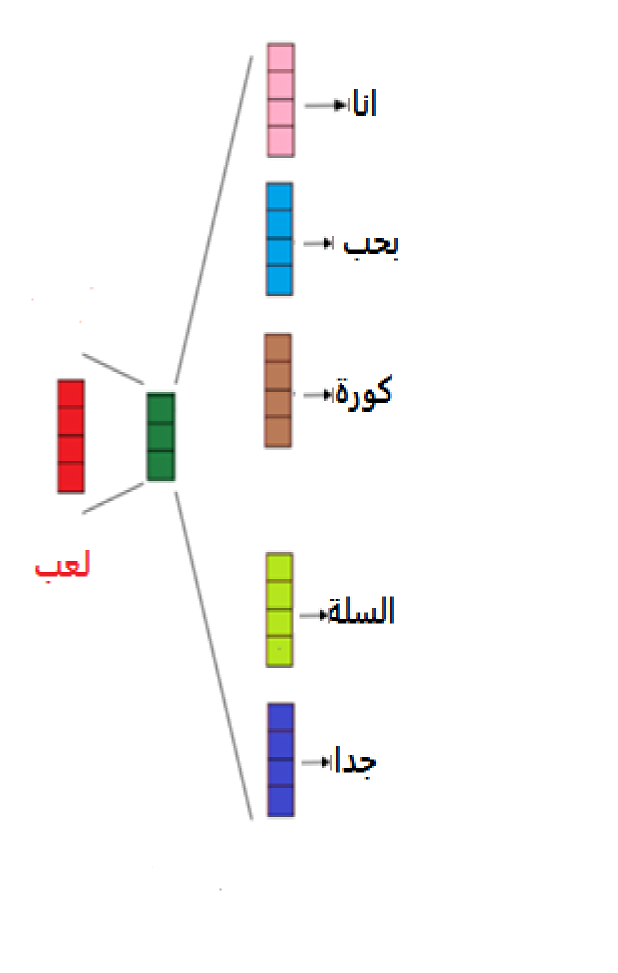
- Continuous Bag-of-Words model (CBOW)

predicting the word given its context.



- Skip-gram model

predicting the context given a word.



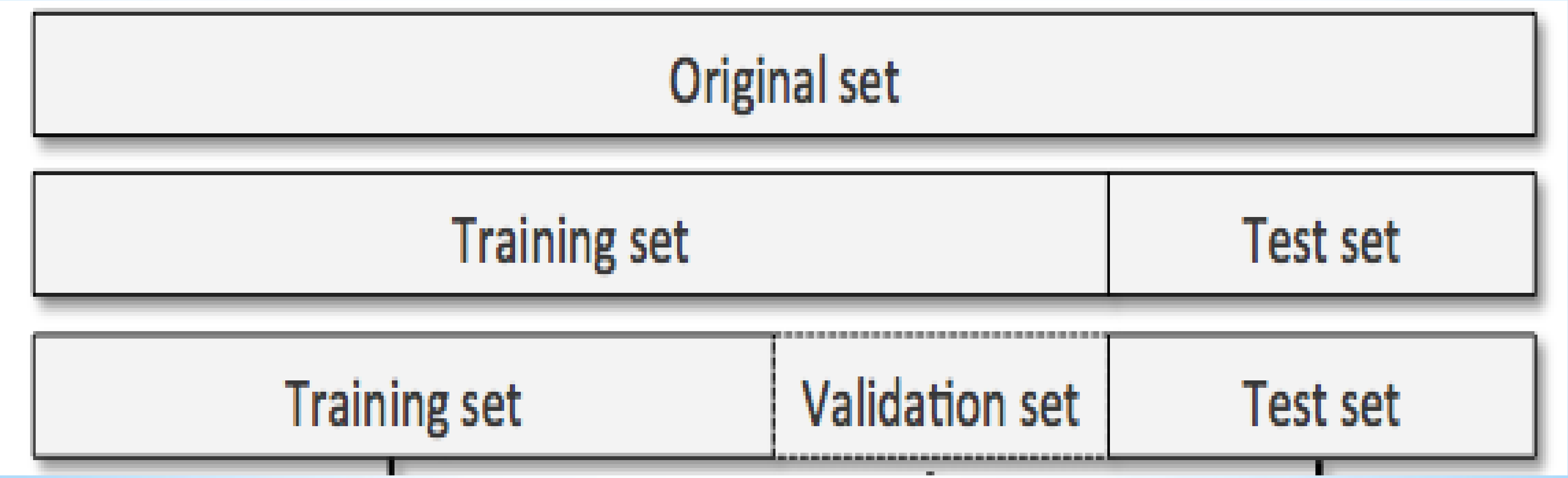
4- Data Splitting

It’s split into training, validation, and test sets.

- create sets for the features and the labels.

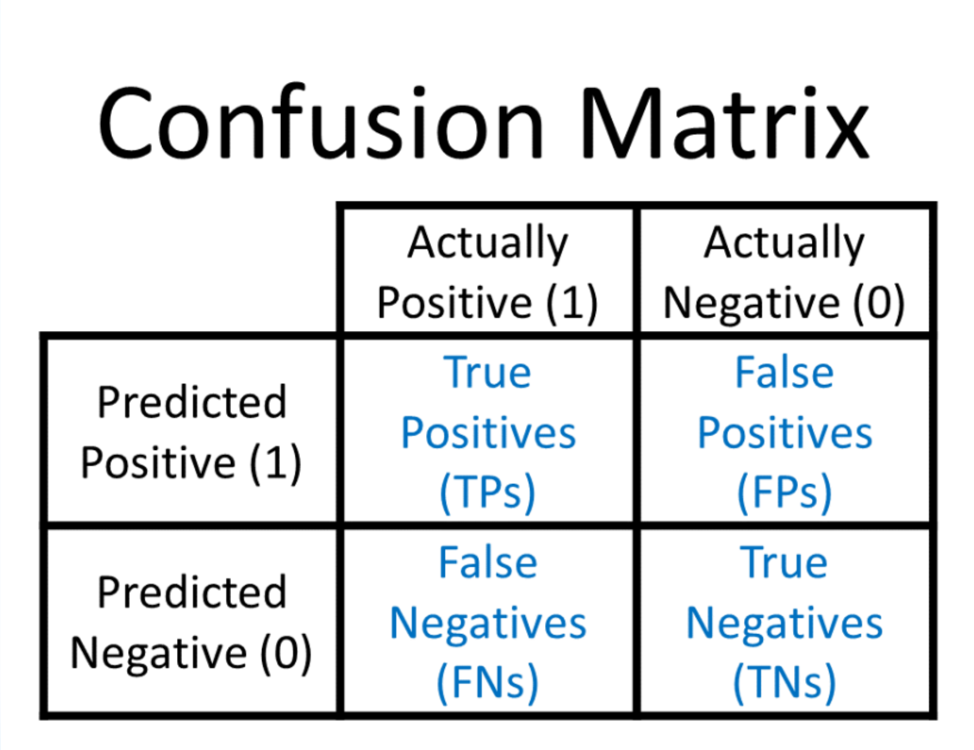
train_x and train_y

test_x and test_y



Validation Acc	Testing Acc	Result
LOW	LOW	Under fit
HIGH	LOW	Over fit
HIGH	HIGH	Fit

Evaluation



Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Accuracy = (TP+TN) / (TP+TN+FP+FN)	**(1)**
Precision=TP/ (TP+FP)	(2)
Recall= TP/ (TP+FN)	(3)
F1 score= (2* Precision*Recall)/ (Precision+ Recall)	(4)

Deep Learning

- ✓ Deep learning (also known as deep structured learning or hierarchical learning).
- ✓ Deep learning is a branch of machine learning.

Advantages:

- Reduces the need for feature engineering, one of the most time-consuming parts of machine learning practice.

Requirements:

- Requires a large amount of data.
- GPU.

Recurrent Neural Network(RNNs)

- RNNs have gained tremendous attention in the NLP field, and they have been employed to handle many tasks, including machine translation.
- Its objective, or the problem it solves is the problem of prediction.
- Long Short Term Memory (LSTM) is method that learn from a sequence of words and outperformed on several feature-engineering approaches.

LSTM

1. Input Gate.
2. Output Gate.
3. Forget Gate.
4. Memory Cell.

Let's go to the Jupyter

Set up Anaconda, Jupyter Notebook
Install TensorFlow and Keras
for studying Deep Learning

ANACONDA
jupyter
TensorFlow
K Keras

thank you