

# Workshop

# Text Classification for the Arabic Language

Dr. Fatemah Husain

[f.husain@ku.edu.kw](mailto:f.husain@ku.edu.kw)

21<sup>st</sup> September 2022



# Fatemah Husain

- Assistant Professor at the department of Information Science, College of Life Sciences, Kuwait University.
- Research: Arabic Natural Language Processing
- [The Information Science Lab - https://infoscilab.ku.edu.kw](https://infoscilab.ku.edu.kw)

# Agenda

- Introduction to text classification
- Selecting the dataset
- Exploratory data analysis
- Text preprocessing
- Preparing the dataset
- Feature extraction
- Classification model
- Performance evaluation
- Error Analysis

# What is Text Classification?

The image illustrates text classification through a magnifying glass over a blue background. The magnifying glass is positioned over a collection of text samples and icons representing different NLP tasks.

**Sentiment Analysis**

Icons: 😊 😐 😞 😡

**Topic Modeling**

Icons: 🍀 📖 🏃 🖥️

**Language Detection**

Icons: ES AR EN

**Text Samples:**

- Facebook Post:** A post from Abu Dhabi, United Arab Emirates, featuring a video of children playing and the text "الأطفال يلعبون مجاناً".
- Twitter Post:** A tweet from @elitaAlghamdi, dated August 10, 2022, at 4PM Kuwait Time, with the text "Unfold your world Join us as we unfold this #SamsungUnpacked".
- News Article:** A headline in Arabic: "الضبيب لـ 'الأنباء': بوتونكول شامل لتشخيص وعلاج امراض السمنة في الكويت".
- Amazon Customer Review:** A review of a dictionary, dated January 22, 2013, with the text "Oh dear! The print is so small that I have to use a magnifier as well as my reading glasses! A great pity as it is a brilliant dictionary with beautiful illustrations and covers everything. I would gladly have paid more for it in a larger sized copy. Not a dictionary one could use instantly in a necessary situation. Both my Arabic tutors shared my opinion and expressed same to their students; neither wore glasses, as I do."

# Is this a Spam?

[From: Mrs. Suzanne Mubarak ...]



Mrs. Suzanne Mubarak <mrssuzamubarak@gmail.com>



To: Recipients <mrssuzamubarak@gmail.com>

Sun 7/10/2022 7:06 AM

Dear Friend,

I am Mrs. Suzanne Mubarak, the wife of deposed and now late Egyptian President Hosni Mubarak who was jailed by the government of Egypt. You must have heard over the media reports and the Internet on the discovery of some fund in my husband secret bank account and companies and the allegations of some huge sums of money deposited by my late husband in my name of which I have refused to disclose or give up to the corrupt Egyptian Government.

In fact the total sum allegedly discovered by the Government so far is in the amount of about \$6.5 Billion Dollars. And they are not relenting on their effort to make me and my sons(Gamal & Alaa Mubarak) poor for life.

As you know, the Muslim community has no regards for women, more importantly when the woman is from a Christian background, hence my desire for a foreign assistance.

This arrangement will be known to you and I alone and all our correspondence should be strictly on email alone because our government

# Is this an Offensive Language Tweet?



Tweet



سرمدار معین  
@M0in\_Mushtaq



This is women empowerment. Hijab is old fashion which we don't want in our society anymore.

Pigs.



Paul Williams @freemonotheist · Jul 23

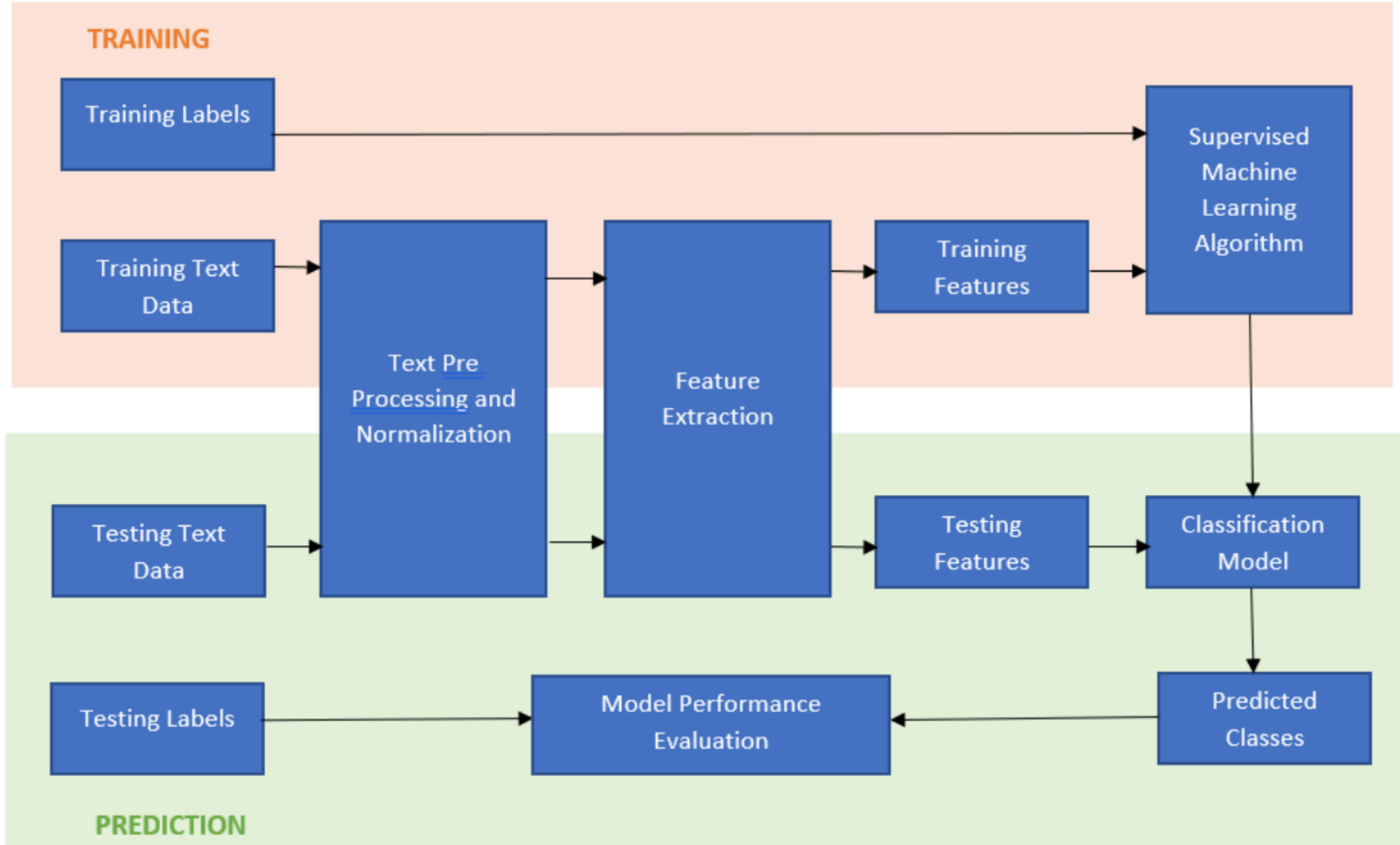


**HIJAB, JILBAB, LONG DRESSES BANNED**

**OFFICIALS CUTTING DRESSES OF UYGHUR  
MUSLIM WOMEN BECAUSE THEY ARE 'LONG'**

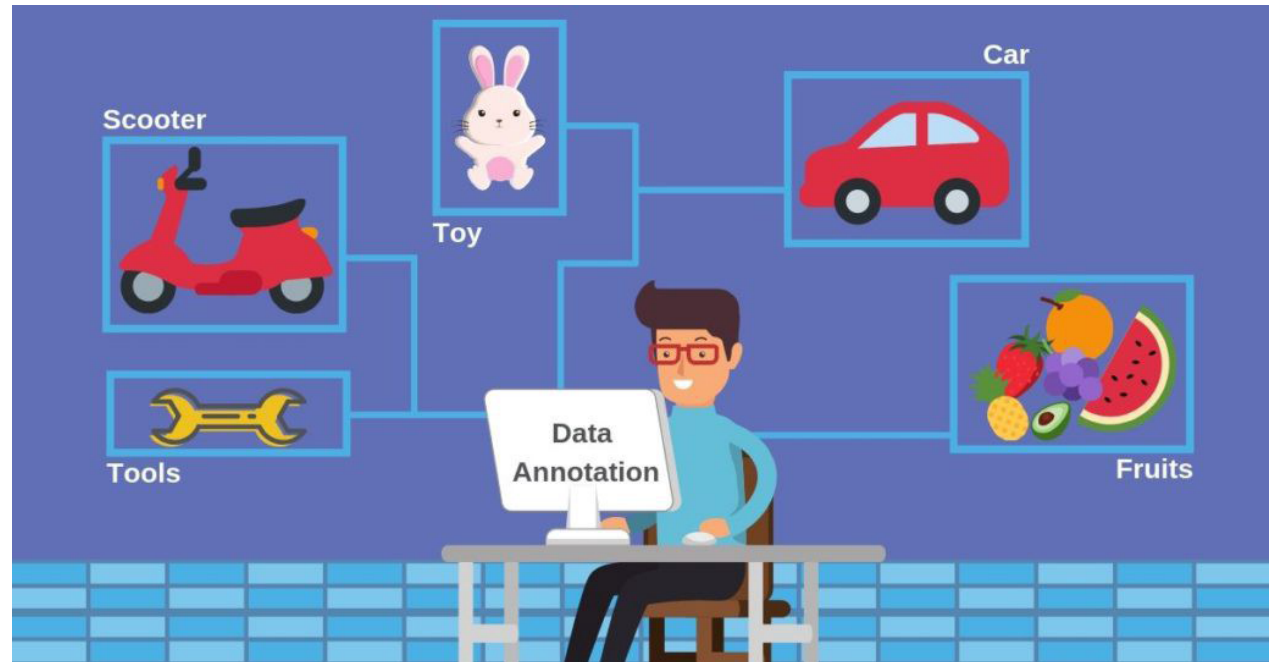


# Text Classification Pipeline



# Dataset Selection

- Dataset sizes
- Distribution of samples (classes)
- Annotation of the datasets
- Source of the dataset



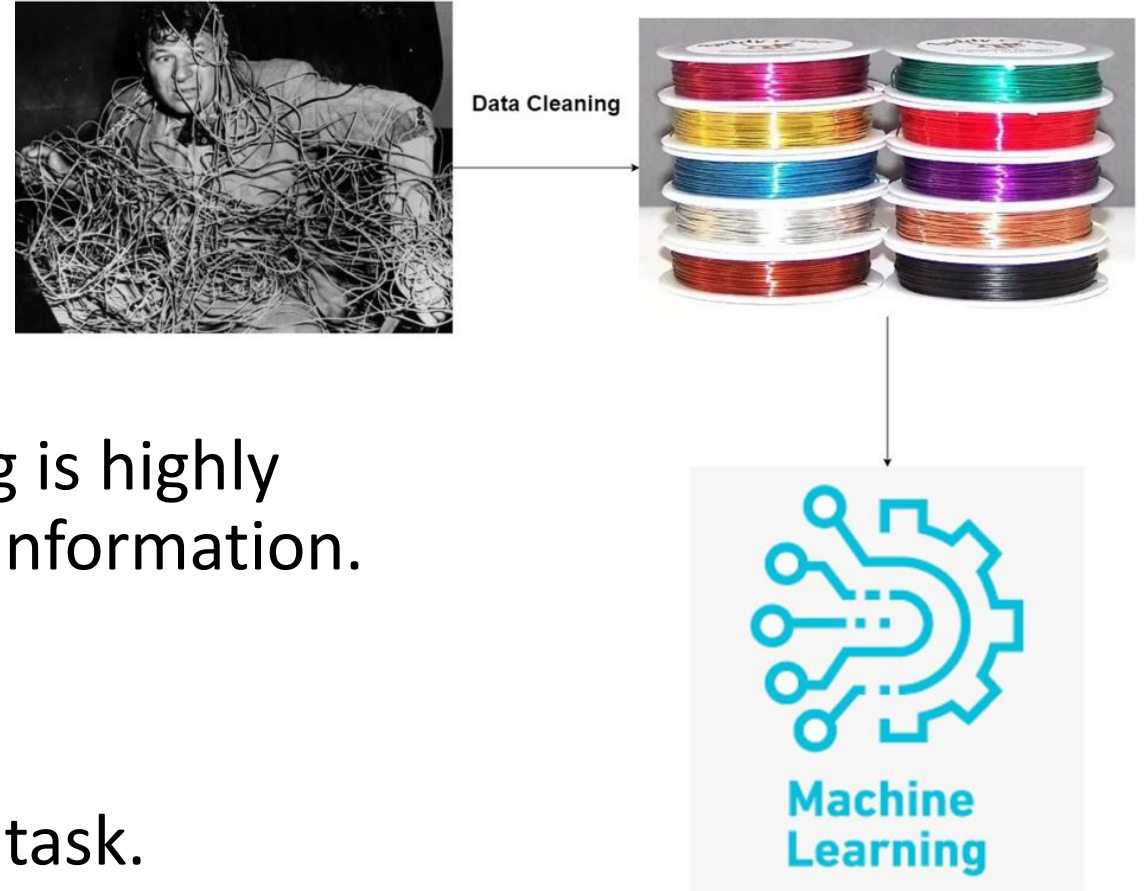


# Exploratory Data Analysis

- Initial dataset investigation
- Help to discover patterns
- Help to understand the dataset
- Learn the main characteristics of the dataset

# Text Preprocessing

- Raw documents without preprocessing is highly unstructured and contains redundant information.
- dimensionality reduction.
- Bring your text into a form that is **predictable** and **analyzable** for your task.
- Not directly transferable from task to task



# Text Preprocessing

## Emoji Conversion

يا اتربيتوا عليها يا لاً 🙄

*you were raised on it  
or no*

يا اتربيتوا عليها يا لاً **وجه**  
**مبتسم مع هالة**

## Dialects Normalization

لا تعتب عليه هيدا اسمه ابو  
**صرماي**

*Don't worry his name  
is shoe's owner*

لا تعتب عليه هيدا اسمه ابو  
**حذاء**

# Text Preprocessing

## Word Categorization

بدل ما تخسر تبدیل یا **حمار**

*better than losing,  
donkey*

بدل ما تخسر تبدیل یا **حيوان**

## Letters Normalization

يا أنت يا حبيبي يا أنيق

*Lovely, you are elegant*

يا انت يا حبيبى يا انق

# Text Preprocessing

## Hashtags Segmentation

نحبك يا تاج الرأس يا كل  
أيامنا أنت **#يوم\_الأم**

*We love you, you are  
our days mother's day*

نحبك يا تاج الرأس يا كل  
أيامنا أنت **يوم الأم**

## Miscellaneous

يا **RT : @USER @USER**  
نموت احرار يا نمشي  
ماليزيا

*We either die free or we  
will leave to Malaysia*

نموت احرار نمشي  
ماليزيا

# Dataset Preparation

- Partitioning the dataset
  - Train set - used in teaching the model
  - Validation/Development/Evaluation set - used for initial evaluation of the model's parameters
  - Test set - used for testing the mode
- Converting labels to numeric format

# Feature Extraction

- Numerical representation for individual words
- Word embeddings
- Assign particular weights to words that tell us how important they are in the document

# Feature Extraction

- **Bag-Of-Words:** represents text as multisets (bags) without preserving the order of the words but keeping their frequencies.

	about	bird	heard	is	the	word	you
About the bird, the bird, bird bird bird	1	5	0	0	2	0	0
You heard about the bird	1	1	1	0	1	0	1
The bird is the word	0	1	0	1	2	1	0



# Feature Extraction

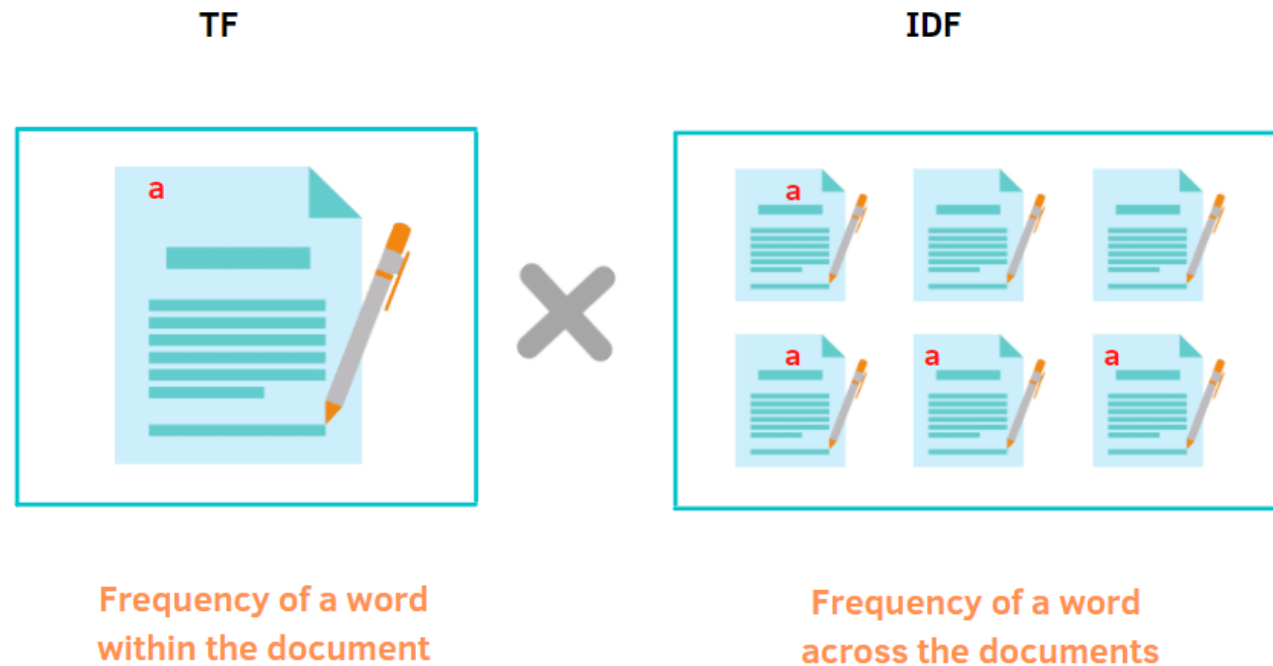
- **Term Frequency-Inverse Document Frequency (TF-IDF):** a form of sparse word embedding
- Provide semantic relationship between words in the vocabulary
- $tf_{t,d} = \log_{10}(\text{count}(t,d)+1)$
- $df_t$  is the number of documents  $t$  occurs in

$$idf_t = \log_{10} \left( \frac{N}{df_t} \right)$$

$N$  is the total number of documents in the collection

# Feature Extraction

- Final tf-idf weighted value for a word  $w_{t,d} = \text{tf}_{t,d} \times \text{idf}_t$



# Classification Model Development

- Computers can teach themselves to use data and learn from their experiences to make more accurate decisions by using machine learning models
- A classification function to estimate the class for each instance
- Examples:
  - Naive Bayes is a generative classifier
  - Logistic regression is a discriminative classifier

# Classification Model Development

## Generative and Discriminative Classifiers

Suppose we're distinguishing cat from dog images



imagenet



imagenet

# Classification Model Development

## Generative Classifier:

- Build a model of what's in a cat image
  - Knows about whiskers, ears, eyes
  - Assigns a probability to any image:
    - how cat-y is this image?



Also build a model for dog images

Now given a new image:

**Run both models and see which one fits better**

# Classification Model Development

## Discriminative Classifier

Just try to distinguish dogs from cats



Oh look, dogs have collars!  
Let's ignore everything else

# Performance Evaluation

- **Precision:** how many of the returned labels are correct
- **Recall:** how many of the labels that should have been returned are actually returned
- **Accuracy:** the classifier does what it is supposed to do

		<i>gold standard labels</i>		
		gold positive	gold negative	
<i>system output labels</i>	system positive	true positive	false positive	precision = $\frac{tp}{tp+fp}$
	system negative	false negative	true negative	
		recall = $\frac{tp}{tp+fn}$		accuracy = $\frac{tp+tn}{tp+fp+tn+fn}$

# Performance Evaluation

- **F1 measure** : a balance between the quantity and the quality of labels

$$F1 = \frac{2 * (Precision * Recall)}{(Precision + Recall)}$$



# Error Analysis

- Manually looking at the errors (the examples in the validation dataset that the simple algorithm doesn't work properly) to generate more insights.
- Try different ideas and cross-check whether they are improving your application or not

# Project: Offensive Language Detection

- Login to your Gmail
- Download the Jupyter Notebook (Text Classification Basics - Workshop) from this repository:
- <https://github.com/Fatemah-Husain>
- Start a new Colab project using the same file and follow its instruction.