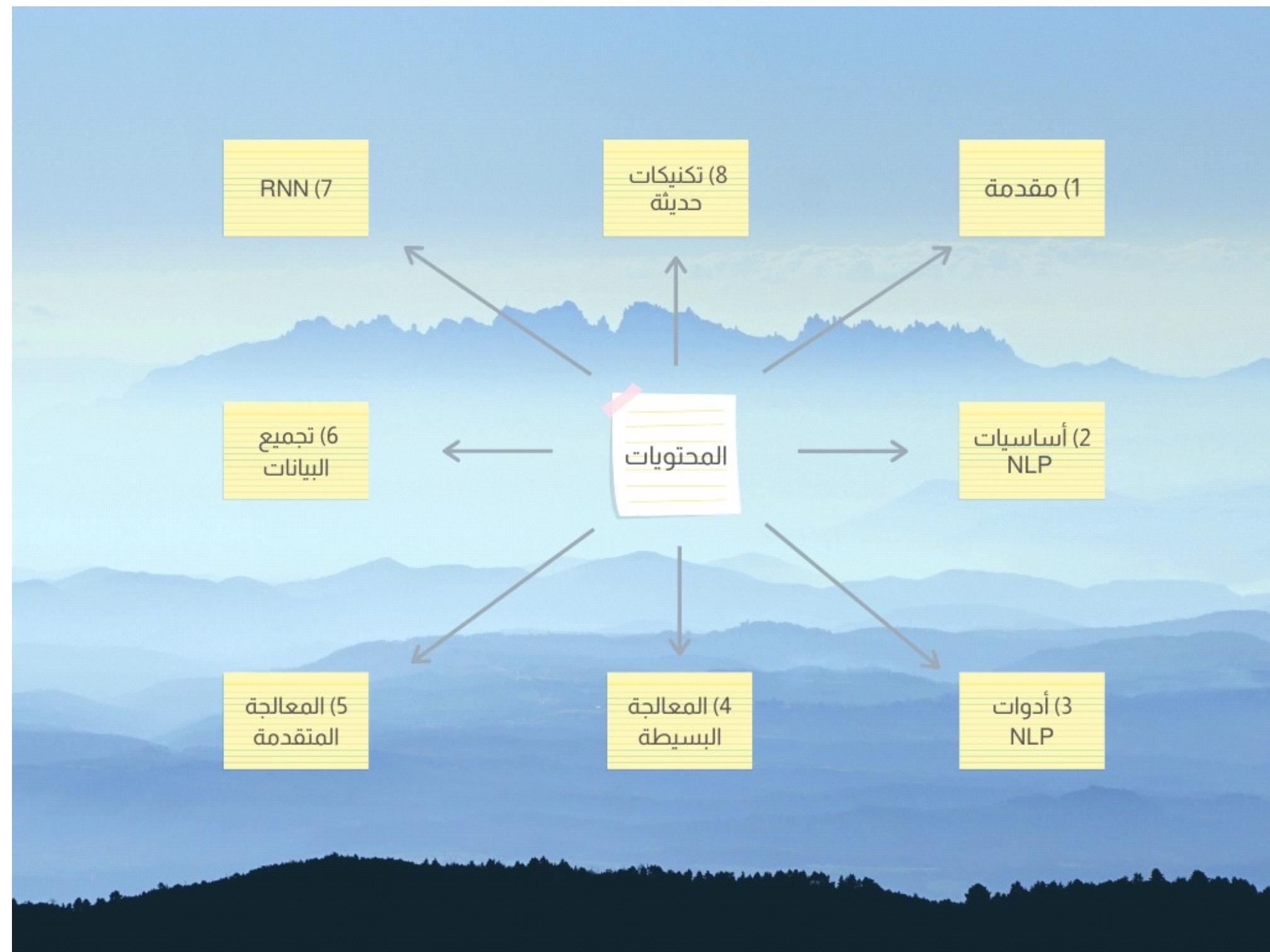


NATURAL LANGUAGE PROCESSING

المعالجة اللغوية الطبيعية



المحتويات

				التطبيقات	العقبات و التحديات	NLP تاريخ ملفات pdf	ما هو NLP الملفات النصية	المحتويات المكتبات	1) مقدمة
T.Visualization	Syntactic Struc.	Matchers	Stopwords	NER	Stem & Lemm	POS	Sent. Segm.	Tokenization	2) أساسيات NLP
	Dist. Similarity	Text Similarity	TF-IDF	BOW	Word2Vec	T. Vectors	Word embed	Word Meaning	3) أدوات NLP
T. Generation	NGrams	Lexicons	GloVe	L. Modeling	NMF	LDA	T. Clustering	T. Classification	4) المعالجة البسيطة
	Summarization & Snippets		Ans. Questions		Auto Correct	Vader	Naïve Bayes	Sent. Analysis	5) المعاجلة المتقدمة
Search Engine	Relative Extraction		Information Retrieval		Information Extraction		Data Scraping	Tweet Collecting	6) تجميع البيانات
					Rec NN\TNN	GRU	LSTM	Seq to Seq	7) RNN (
Chat Bot	Gensim	FastText	Bert	Transformer	Attention Model	T. Forcing	CNN	Word Cloud	8) تكنيكات حديثة

القسم الخامس : المعالجة المتقدمة للنصوص

الجزء الأول : **Text Classification**

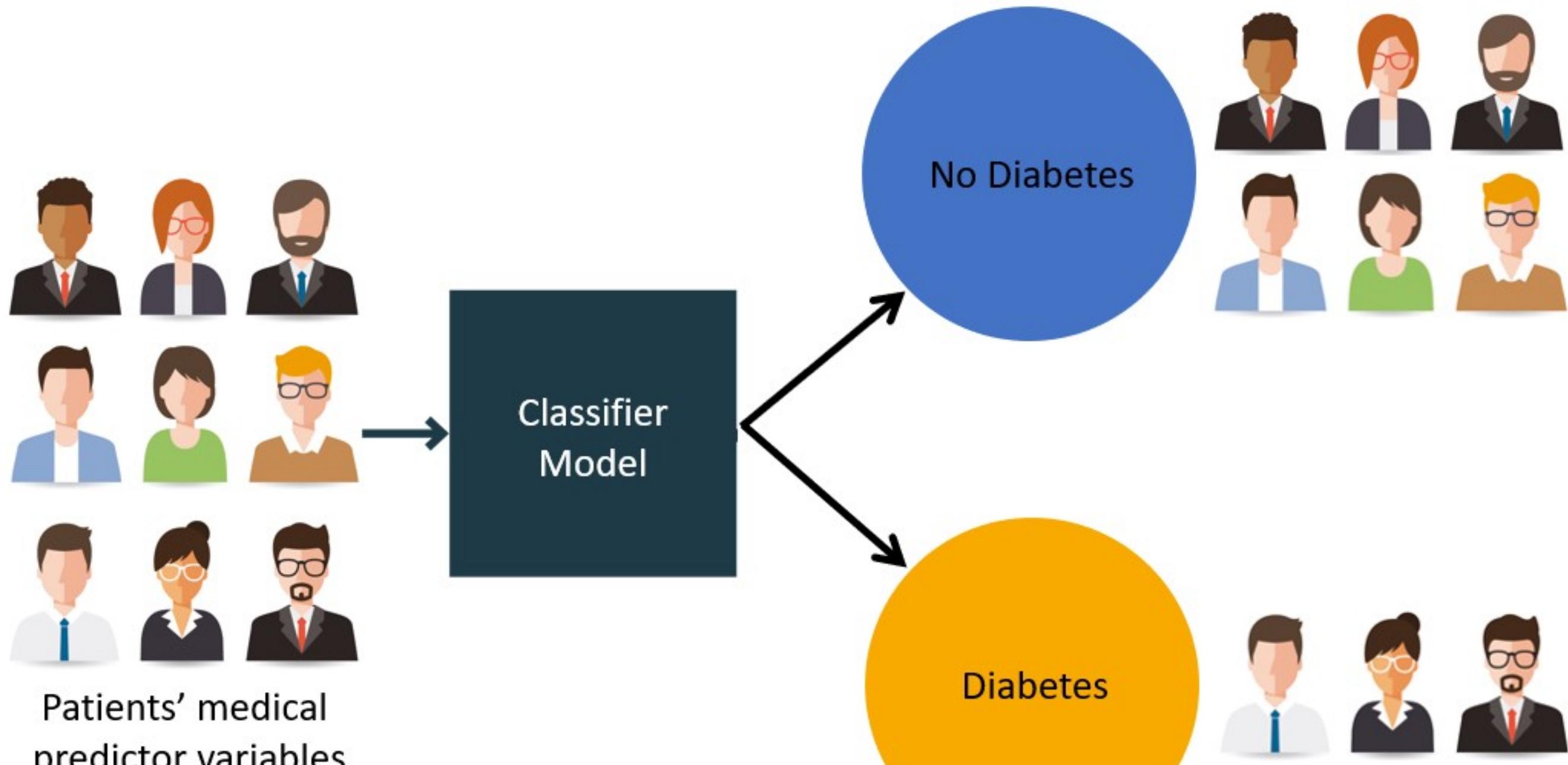
تصنيف النصوص **Text Classification** هي من التطبيقات المنتشرة و الهامة في مجال الـ NLP ، و يتزايد الطلب عليه بشكل دائم

و يقصد به استخدام تكنولوجيات NLP لتصنيف النصوص الى أكثر من فئة ، بشكل تلقائي اعتمادا على محتوياتها النصية

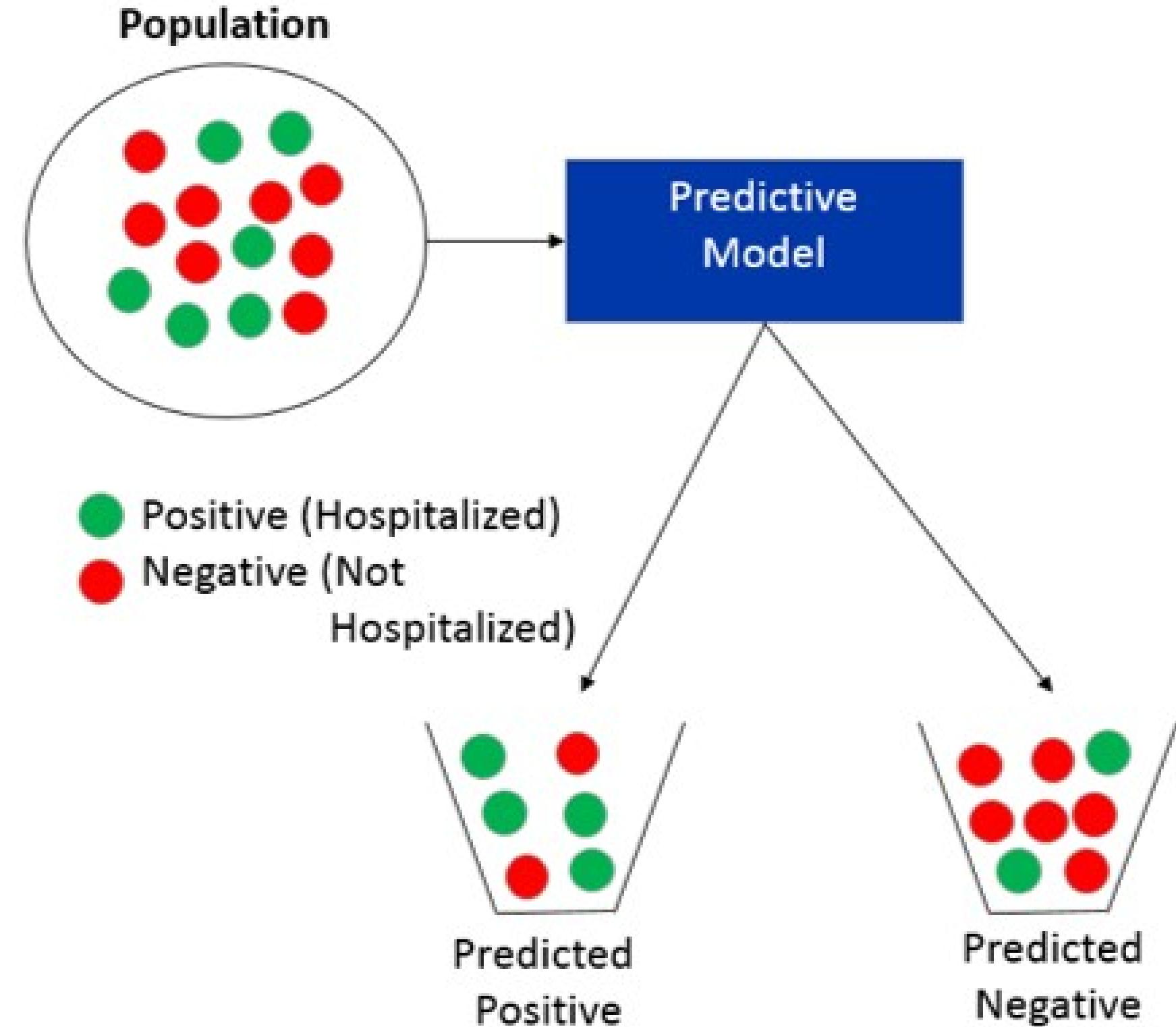
و علينا أولا ان نتذكرة ، ما هو التصنيف ..

--*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*

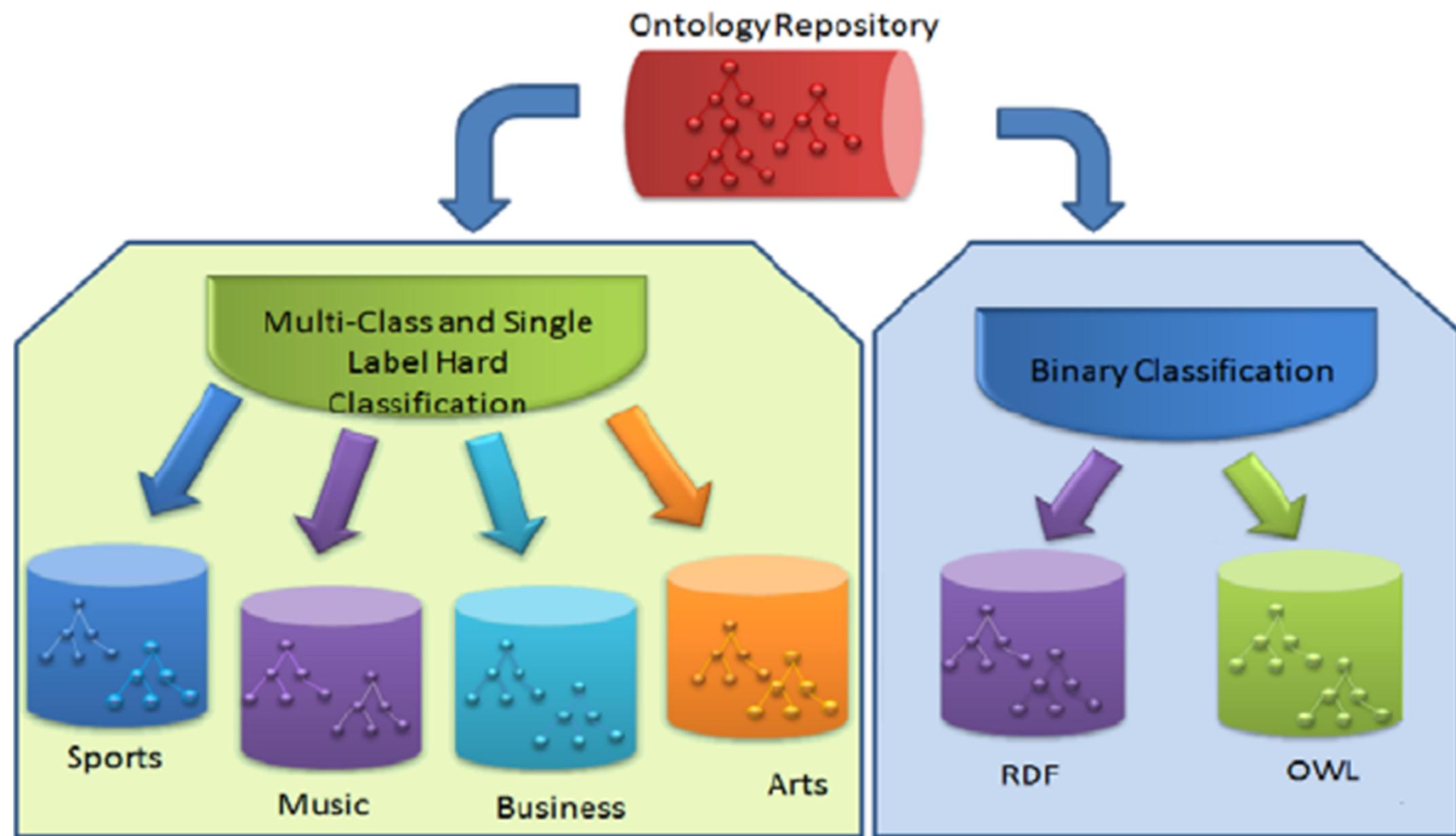
هو استخدام خوارزميات الذكاء الاصطناعي ، في تحديد الفئة التي تتنمي لها العينة



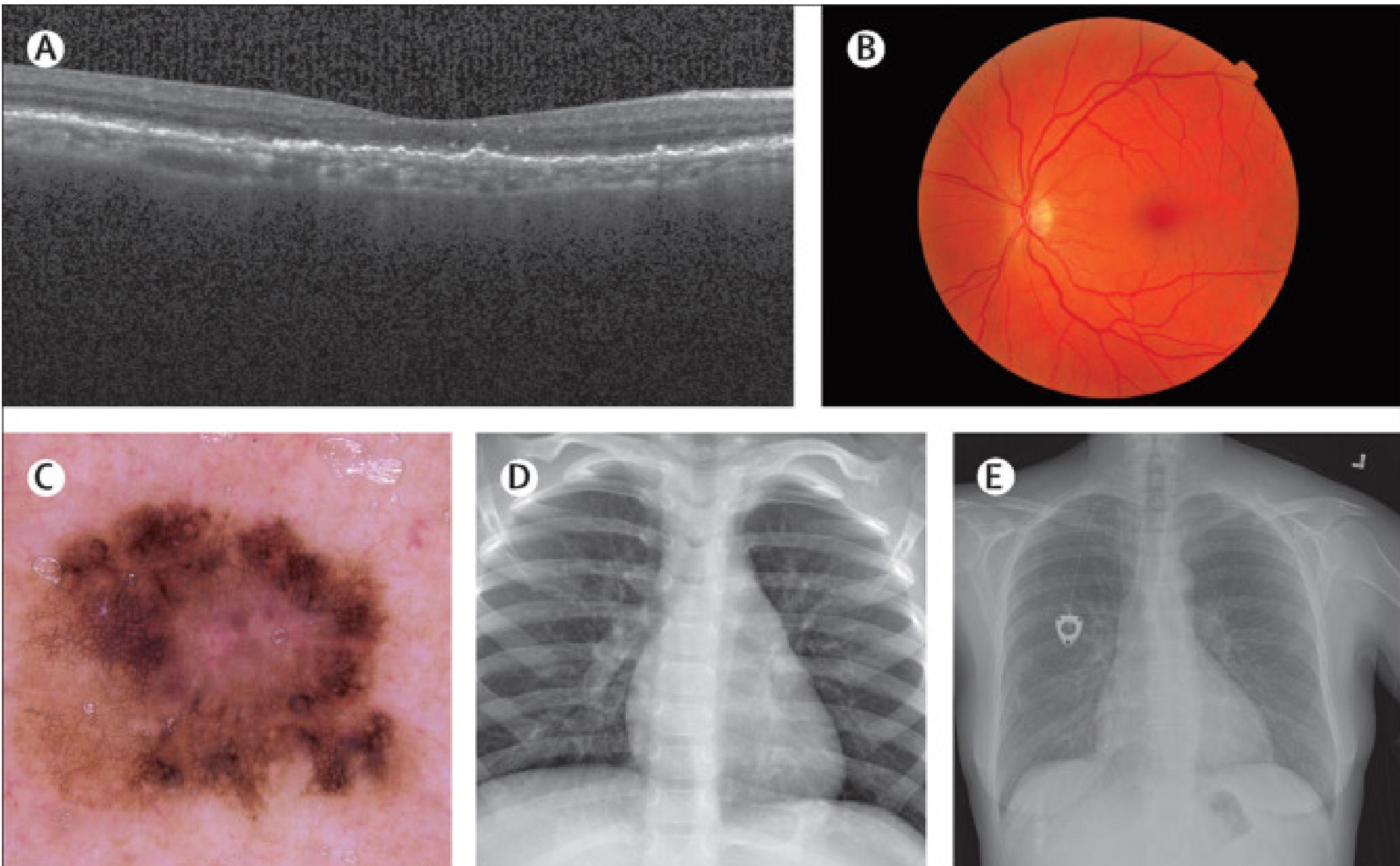
لا تكون بالضرورة جميع الخوارزميات ذات كفاءة عالية



و لها نوعين التصنيف الثنائي ، والمتعدد



و لها استخدامات طبية عديدة .



و ايضا في تصنیف الصور



airplane



automobile



bird



cat



deer



dog



frog



horse



ship



truck



و لها معادلة خطأ cost function و التي تمثل الفارق بين القيمة المتوقعة و الحقيقة

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m Cost(h_\theta(x^{(i)}), y^{(i)})$$

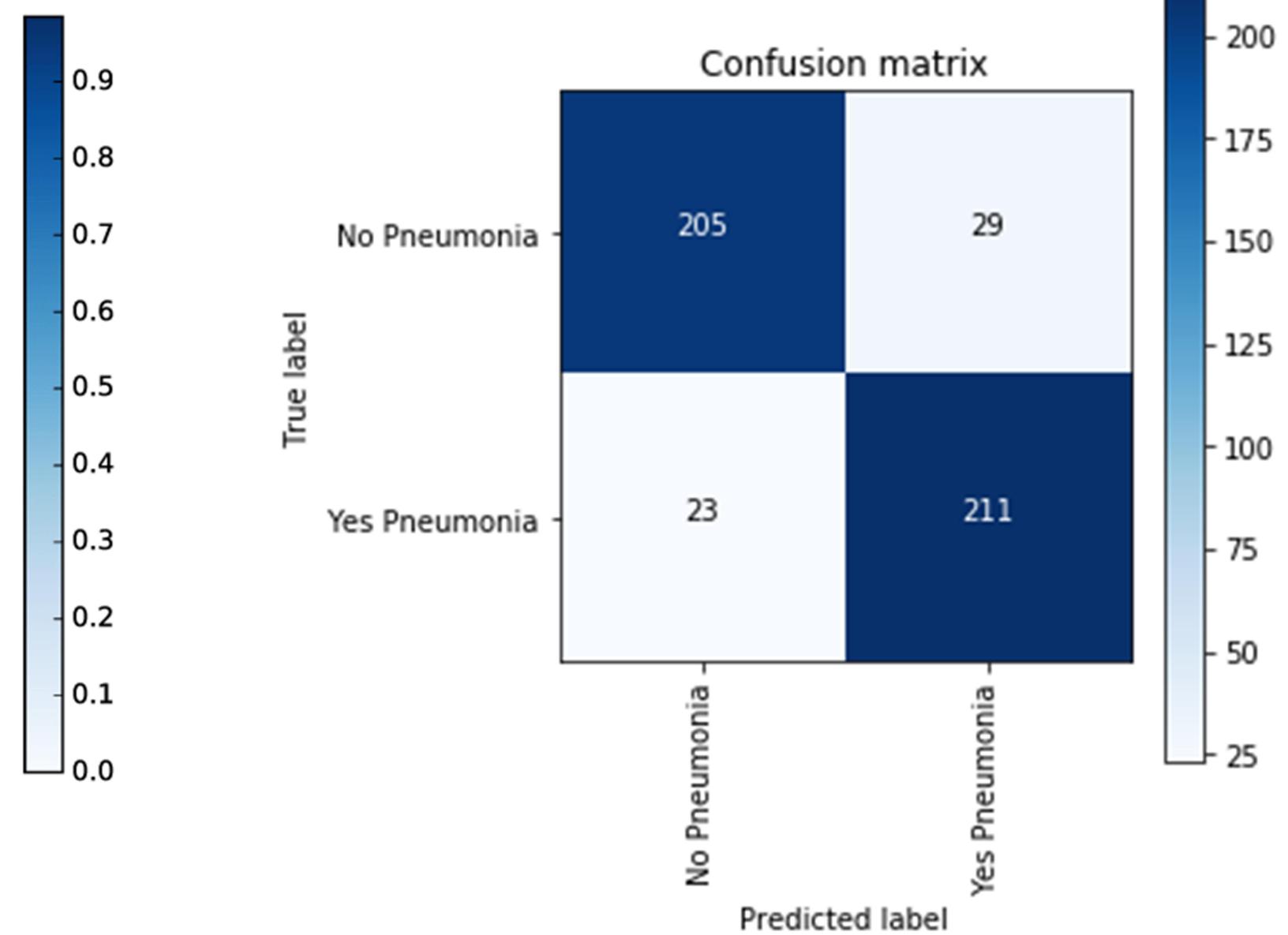
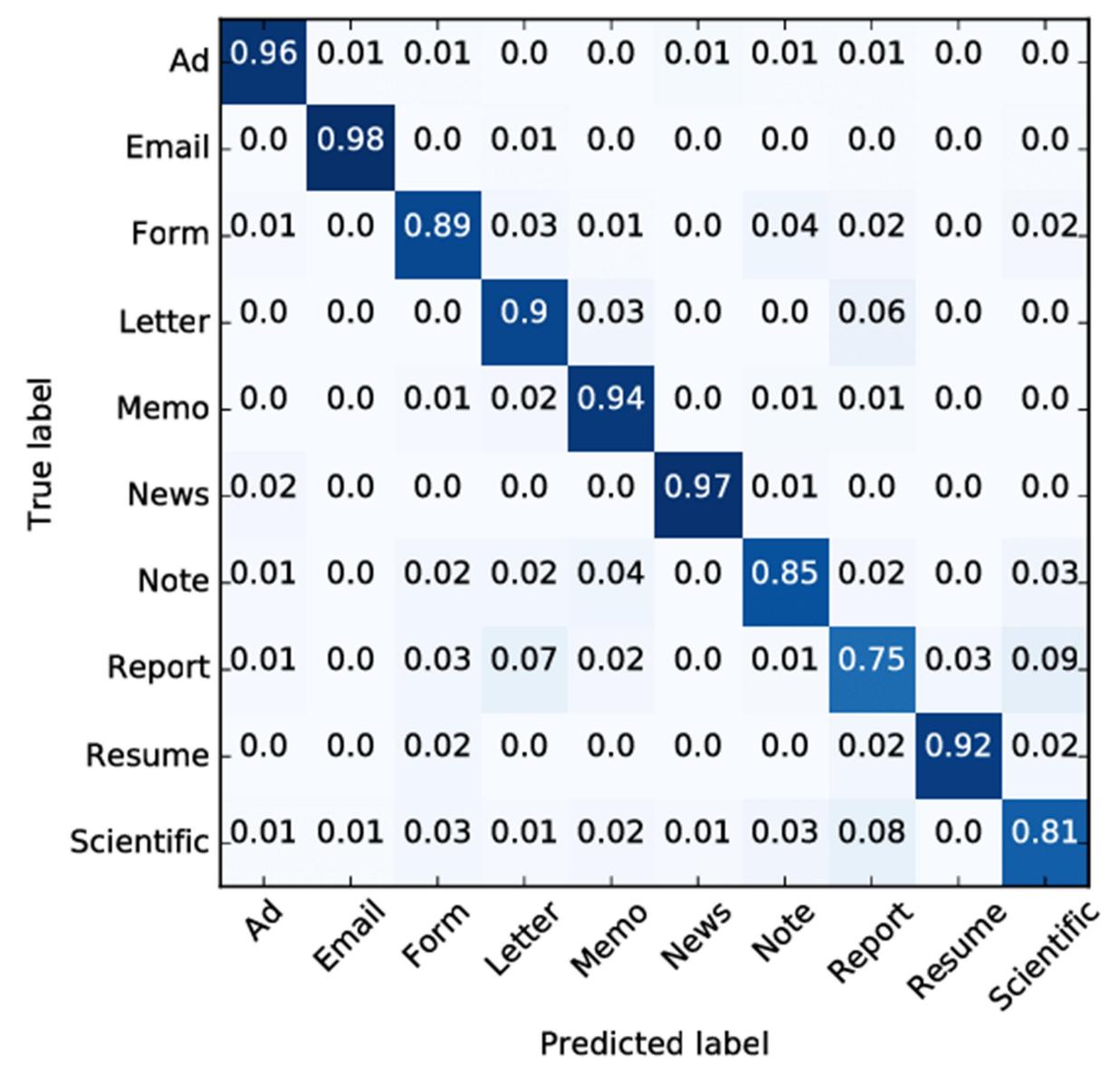
$$J(\theta) = \frac{1}{m} \left[\sum_{i=1}^m -y^{(i)} \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)})) \right]$$

m = number of samples

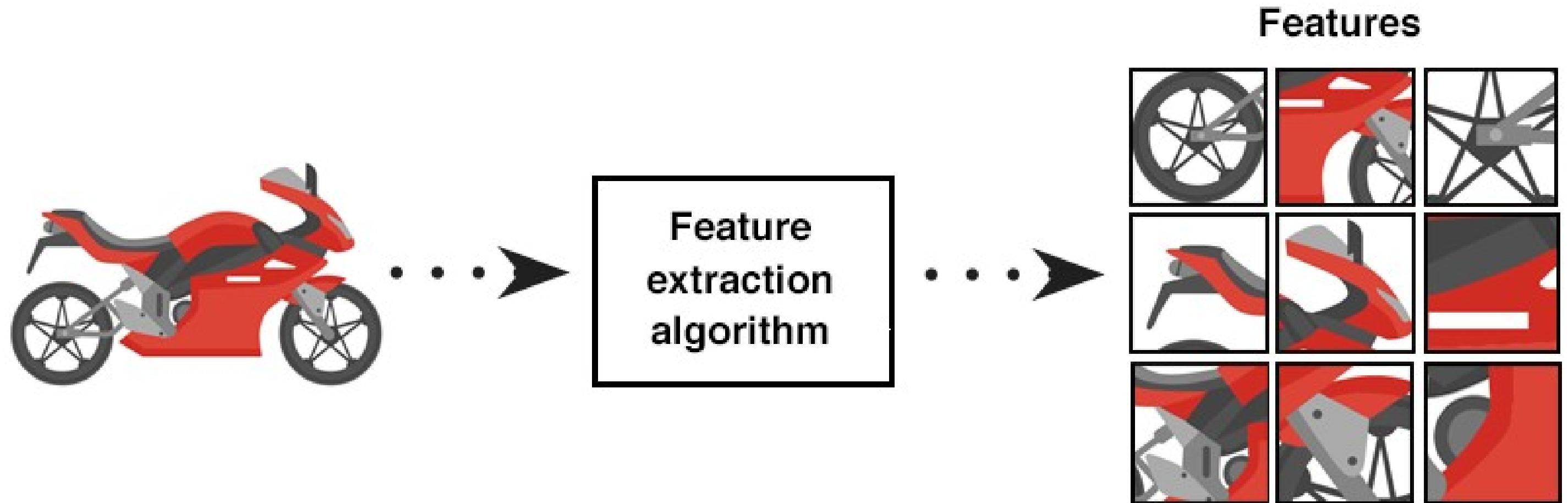
كما اننا نستخدم مصفوفة التشتت لاظهار مدى كفاءتها

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)	



و الأهم هو خطوة feature extraction لاستخلاص المعلومات من المحتوى

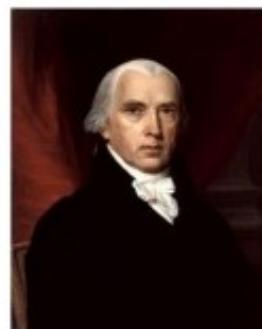


--*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*

الآن ، ما هو تصنيف النصوص . .

في الستينيات تم النجاح في تحديد من كتب عدة خطابات رسمية متعلقة بالدستور الأمريكي ، عبر استخدام طرق احصائية مثل Bayesian

- 1787-8: anonymous essays try to convince New York to ratify U.S Constitution: Jay, Madison, Hamilton.
- Authorship of 12 of the letters in dispute
- 1963: solved by Mosteller and Wallace using Bayesian methods



James Madison

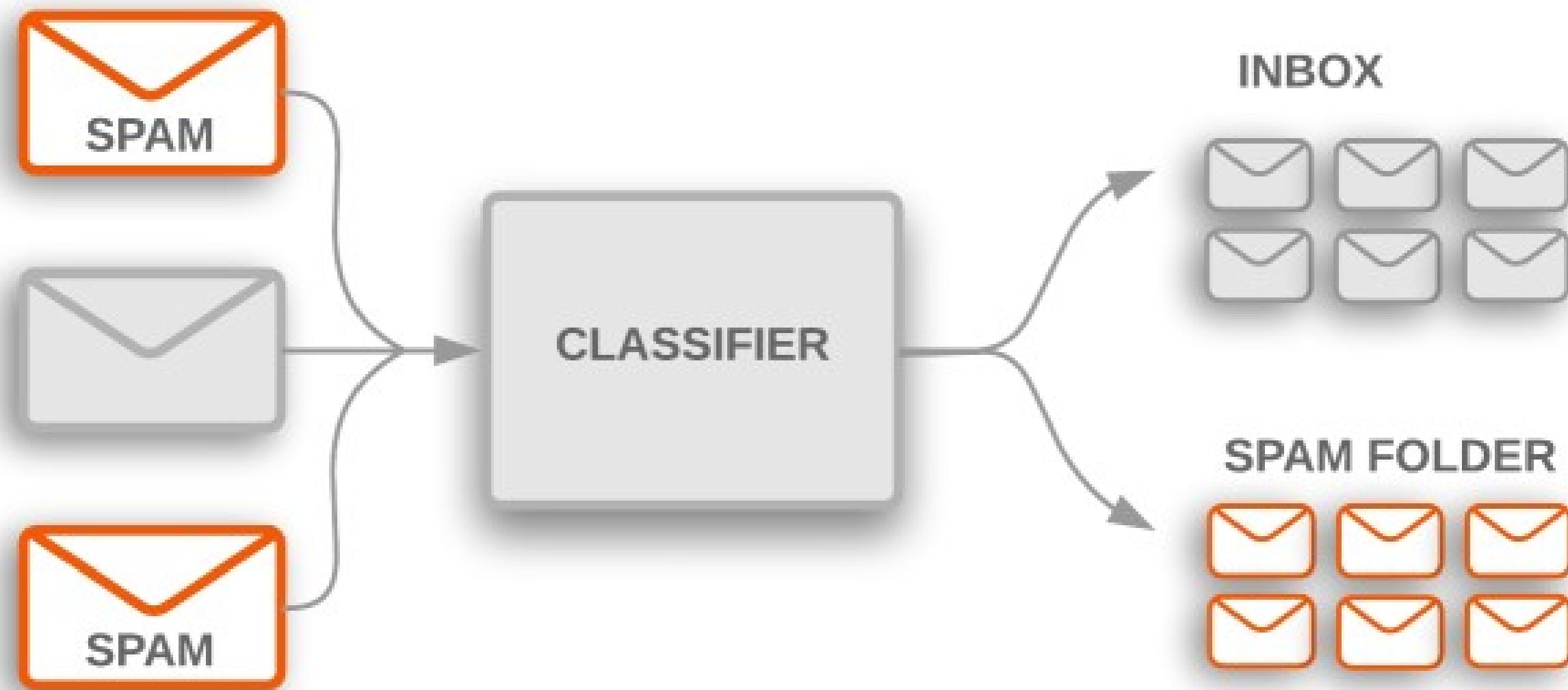


Alexander Hamilton

كما انا يمكن أن نقوم بتصنيف النصوص حسب نوع الكاتب او سنه , بناء علي معايير معينة في النص

1. By 1925 present-day Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony of Cochinchina; the central area with its imperial capital at Hue was the protectorate of Annam...
2. Clara never failed to be astonished by the extraordinary felicity of her own name. She found it hard to trust herself to the mercy of fate, which had managed over the years to convert her greatest shame into one of her greatest assets...

و من أشهر أمثلتها : تصنیف الإيميلات الى عادية و spam



ايضا تحليل المشاعر في النصوص و استنتاج هل هو كلام سلبي ام ايجابي sentimental analysis



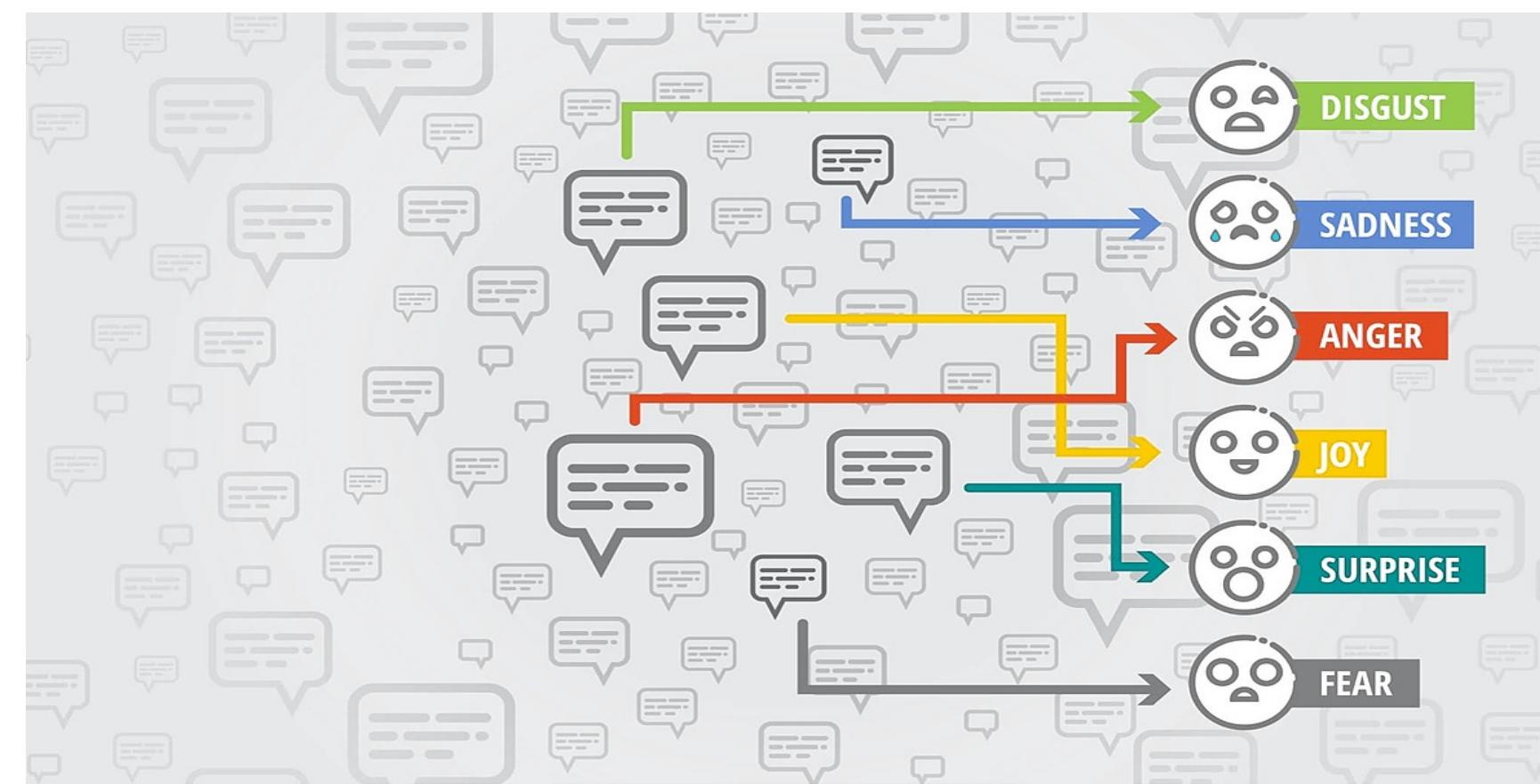
- unbelievably disappointing
- Full of zany characters and richly applied satire, and some great plot twists



- this is the greatest screwball comedy ever filmed



- It was pathetic. The worst part about it was the boxing scenes.

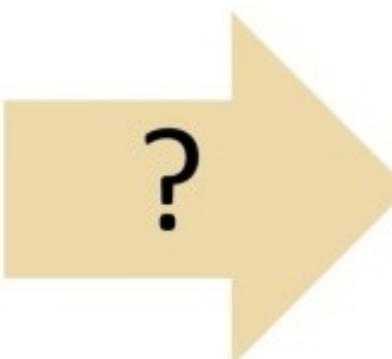


بالإضافة الي تحديد نوع هذا النص ، و الي اي صنف ينتمي

MEDLINE Article



MeSH Subject Category Hierarchy



- Antagonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
-

و هناك عدد من استخدامات أخرى مثل : text classification

- Assigning subject categories, topics, or genres
- Spam detection
- Authorship identification
- Age/gender identification
- Language Identification
- Sentiment analysis

--*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*

خطوات تصنیف النصوص :

و تتم هذه العملية عبر عدد من الخطوات هي :

- تجميع البيانات المعنونة labeled data
- عملية معالجة النصوص
- استخراج الـ features المستخدمة فيها
- اختيار الخوارزم المطلوب
- تقييم عملية التصنیف

--*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*

و نري في الكود العملي ، عدد من الخوارزميات الخاصة بتصنيف النصوص ، والتي نجحت بشكل كبير

و تعتمد الفكرة العامة لها ، على استخراج الـ **features** من النصوص ، ثم ادخالها في احد خوارزميات **classification**

و سنري في الامثلة العملية ان الـ **features** الاساسية هي قيم **tf-idf** ، بينما يمكن اضافة قيم اخرى ، مثل :

- طول النص
- عدد الكلمات
- عدد الحروف
- تواجد الارقام
- تواجد علامات الترقيم
- تواجد كلمات معينة (كلمة اقتصادية او كلمة مسيئة علي سبيل المثال)
- غيرها من الـ **features** التي يراها المبرمج مناسبة

مع العلم ان هذه الأدوات ستعمل بشكل جيد علي اللغة العربية ، ولكن المشكلة في البحث عن داتا مناسبة