TD : Comprendre les mesures d'évaluation en catégorisation de texte

1. Comparer des systèmes de classification binaire

Deux modèles de classification supervisée ont été choisis et appliqués sur une population test. Les résultats sont donnés dans le tableau suivant :

Indiv. X'	Ref.	Hyp. 1	Hyp. 2
1	+	+	+
2	+	+	+
3	+	+	+
4	-	+	-
5	+	+	+
6	+	+	-
7	+	+	+
8	-	+	-
9	-	-	-
10	+	+	+
11	+	+	-
12	+	+	-
13	-	+	-
14	-	-	-
15	-	-	+
16	-	-	-
17	-	-	-
18	-	-	-
19	-	-	-
20	-	-	-

a. Quel type d'erreur peut-on évaluer ?
b. Calculer les tables de contingence.
c. Quels sont le taux de mauvaise classification et l'accuracy pour chacun des systèmes ?
d. Donnez les valeurs de rappels et précisions pour chacun des systèmes.
e. Quel est selon vous le meilleur système ? Justifiez.
f. Avez-vous considéré la significativité de vos résultats ?

Aide : $$CER \pm 1.96\sqrt{\frac{CER(1-CER)}{N}}$$

2. Calcul de mesures d'évaluation dans un cadre multi-classe

On souhaite associer chaque individu d'une population donnée à l'une des classes suivantes : A, B, C ou D. Soit X un ensemble de 30 individus utilisés pour construire le modèle de classification et X' un ensemble de 10 autres. Les performances sur chacune de ces populations sont indiquées dans les tableaux ci-dessous :

Indiv. X	Réf.	Hyp. 1
1	A	A
2	A	A
3	A	A
4	A	A
5	A	A
6	C	A
7	D	A
8	C	C
9	B	B
10	A	C
11	D	B
12	B	B
13	B	B
14	C	A
15	B	B
16	C	A
17	D	A
18	A	A
19	A	A
20	A	A
21	A	A
22	A	D
23	D	B
24	C	A
25	C	C
26	A	C
27	A	A
28	A	A
29	B	B
30	A	A

Indiv. X'	Réf.	Hyp. 1
31	C	A
32	A	A
33	C	C
34	A	A
35	B	C
36	C	A
37	A	A
38	C	C
39	A	A
40	B	C

a. Ecrire la table de contingence qui permet d'évaluer les performances de prédiction du modèle.
b. À partir de cette table, calculer les indicateurs suivants :
    i. Taux d'erreur et accuracy
    ii. Macro-mesures (précision et rappel)
    iii. Micro-mesures (précision et rappel)
c. Quelles remarques faites-vous sur le pouvoir de prédiction du modèle 1 ?

3. Analyser et comparer plusieurs systèmes de DEFT2017

Ces tableaux sont extraits des actes de DEFT 2017, disponibles à l'adresse :
https://deft.limsi.fr/2017/actes_DEFT_2017.pdf

Détails des corpus :

Tâche 1	Entrainement	Test	TOTAL
Objectif	1 642 (42,1%)	411 (42%)	**2 053 (42,05%)**
Positif	494 (12,6%)	123 (12,65%)	**617 (12,65%)**
Négatif	1 268 (32,5%)	317 (32,5%)	**1 585 (32,45%)**
Mixte	502 (12,8%)	125 (12,85%)	**627 (12,85%)**
TOTAL	**3 906 (80%)**	**976 (20%)**	**4 882**

TABLE 2 – Répartition des tweets pour la tâche 1

Tâche 2	Entrainement	Test	TOTAL
Non figuratif	3 906 (66,7 %)	976 (66,7 %)	**4 882 (66,7 %)**
Figuratif	1 947 (33,3 %)	488 (33,3 %)	**2 435 (33,3 %)**
TOTAL	**5 853 (80 %)**	**1 464 (20 %)**	**7 317**

TABLE 3 – Répartition des tweets pour la tâche 2

Résultats de la campagne DEFT'2017, tâches 1 et 2

Tâche 1	Micro-précision				Macro f-mesure			
	1	2	3	Rang	1	2	3	**Rang**
LIA (équipe 8)	0,682	0,704	**0,710**	1	0,602	0,634	**0,650**	**1**
Advanse, LIRMM (équipe 4)	0,640	0,628	**0,653**	4	0,555	0,539	**0,557**	**2**
MELODI, IRIT (équipe 14)	**0,697**	0,695	0,695	2	**0,547**	0,545	0,546	**3**
LIUM-OCTO (équipe 12)	0,621	**0,625**	0,589	7	0,534	0,539	**0,545**	**4**
LS2N (équipe 17)	0,632	**0,635**	0,634	5	0,493	**0,534**	0,487	**5**
Tweetaneuse (équipe 2)	**0,622**	0,573	0,571	6	**0,531**	0,488	0,498	**6**
IRISA (équipe 5)	0,644	0,659	**0,676**	3	0,512	0,512	**0,514**	**7**
OrangeLabs (équipe 16)	**0,620**			8	0,467	**0,502**		**8**
Tw-StAR (Université Libre de Bruxelles, Selcuk University, Ibn Zohr University ; équipe 1)	0,537	0,230	0,579	9	0,406	0,217	**0,492**	**9**
LIRMM (équipe 18)	0,341	**0,541**		10	0,219	**0,353**		**10**
Amrita University (équipe 15)	**0,387**	0,364	0,351	11	**0,276**	0,228	0,210	**11**
Équipe 13	**0,308**	0,308		12	**0,239**	0,239		**12**

TABLE 5 – Résultats et classement des systèmes selon la micro-précision et la macro f-mesure pour la tâche 1. Le classement est fait sur la base de la meilleure soumission de chaque participant

Tâche 2	Micro-précision				Macro f-mesure			
	1	2	3	Rang	1	2	3	**Rang**
LIA (équipe 8)	**0,807**	0,801	0,802	2	**0,783**	0,774	0,744	**1**
LIUM-OCTO (équipe 12)	0,699	**0,810**	0,721	1	0,659	**0,774**	0,677	**2**
Advanse, LIRMM (équipe 4)	0,788	0,788	**0,790**	3	**0,750**	0,749	0,750	**3**
Tweetaneuse (équipe 2)	**0,779**	0,761	0,761	5	**0,746**	0,716	0,716	**4**
IRISA (équipe 5)	0,779	0,742	**0,782**	4	0,741	0,642	**0,745**	**5**
LS2N (équipe 17)	0,745	**0,758**	0,751	7	0,718	**0,720**	0,704	**6**
LIRMM (équipe 18)	0,587	**0,755**	0,712	8	0,457	**0,715**	0,630	**7**
MELODI, IRIT (équipe 14)	**0,768**	0,755	0,766	6	**0,709**	0,692	0,706	**8**
OrangeLabs (équipe 16)	**0,699**			9	**0,663**			**9**
Équipe 11	0,613	**0,692**	0,688	10	**0,577**	0,550	0,542	**10**
Amrita University (équipe 15)	0,488	0,490	**0,497**	11	**0,557**	0,550	0,542	**11**

TABLE 7 – Résultats et classement des systèmes selon la micro-précision et la macro f-mesure pour la tâche 2. Le classement est fait sur la base de la meilleure soumission de chaque participant

a. Est-ce que le système gagnant est significativement meilleur que les autres?

4. Courbe de précision-rappel

Ci-dessous sont notés les scores d'un classifieur binaire pour la classe + sur un ensemble test de 20 documents. L'association d'un document à la classe + se fait en fonction de ce score selon la règle suivante : si score > a alors le document est classé +.
Calculer les valeurs de precision et rappel pour les seuils suivants : 0,3 0,4 0,5 0,6 0,8
Esquisser une courbe précision/rappel.

id	classe ref	score (+)
18	-	0,25
6	-	0,28
14	+	0,28
5	-	0,33
13	-	0,37
17	-	0,37
16	-	0,47
2	+	0,48
3	-	0,52
12	+	0,54
15	-	0,57
4	+	0,64
8	-	0,64
7	-	0,74
19	+	0,79
1	+	0,85
20	-	0,87
9	-	0,89
11	+	0,95
10	+	0,98

5. Comprendre et commenter des tableaux de résultats

Voici les tableaux résultats issus d'un rapport de recherche. Selon vous :

- Quels sont les commentaires à associer obligatoirement à ces tableaux ?
- Y a-t-il des résultats qui sont manquants pour mieux comprendre le phénomène ?
- Quelles autres expériences auriez-vous faites ?

a. Sur une classification binaire dont la répartition est la suivante :

	A	B
Nb documents	6909	158

Classifieur	F-mesure (macro et weighted)	Accuracy	Durée
KNeighbors	57.04	50.78	00 :00 :01 :09
SVC (Linear)	**75.52**	**75.78**	00 :00 :02 :56
DecisionTree	64.93	66.4	00 :00 :03 :07
RandomForest	58.51	63.28	00 :00 :02 :12
XGBoost (cv = 10)	79.62	80.46	00 :10 :38 :02
MultinomialNB	68.31	73.44	00 :00 :02 :14

TABLE 13 – Baseline - Sans Traitement

Classifieur	F-mesure (macro et weighted)	Accuracy	Durée
KNeighbors	52.17	48.02	00 :00 :00 :59
SVC (Linear)	**78.73**	**78.72**	00 :00 :02 :40
DecisionTree	64.26	65.19	00 :00 :02 :11
RandomForest	58.26	60.85	00 :00 :01 :39
XGBoost (cv = 10)	77.44	78.9	00 :05 :17 :02
MultinomialNB	68.19	71.5	00 :00 :01 :33

TABLE 14 – RAKE

b. Sur une classification multi-classe dont la répartition est la suivante :

	A	B	C	D	E	F	G	H	I
Nb documents	547	291	1096	91	2785	958	1141	40	47

Classifieur	F-mesure (macro et weighted)	Accuracy	Durée
KNeighbors (3)	92.14	93.62	00 :04 :52
SVC (Linear)	93.28	93.64	00 :52 :45
DecisionTree	**95.17**	**95.21**	00 :05 :32
RandomForest	92.01	91.47	00 :04 :43
XGBoost (cv = 10)	97.05	97.02	06 :50 :30
MultinomialNB	89.60	89.74	**00 :03 :51**

TABLE 16 – Baseline - Sans Traitement

Classifieur	F-mesure (macro et weighted)	Accuracy	Durée
KNeighbors (3)	91.01	91.60	**00 :01 :17**
SVC (Linear)	94.79	96.83	00 :15 :43
DecisionTree	93.83	**94.09**	00 :03 :09
RandomForest	92.19	92.68	00 :01 :46
XGBoost (cv = 10)	96.41	96.72	03 :16 :29
MultinomialNB	89.15	89.21	00 :01 :21

TABLE 17 – RAKE