Clustering sans réduction sur les données labélisées

Dataset Classic4

1. BERT

	k-means	Spectral clustering	HDBSCAN	CAH Ward	CAH Single	CAH Complete	CAH Average
Accuracy	0.76	0.77	0.38	0.77	0.45	0.43	0.44
NMI	0.64	0.73	0.19	0.72	0	0.24	0.01
ARI	0.47	0.52	-0.03	0.52	-0	0.02	-0.01

2. RoBERTa

	k-means	Spectral clustering	HDBSCAN	CAH Ward	CAH Single	CAH Complete	CAH Average
Accuracy	0.45	0.76	0.4	0.58	0.45	0.42	0.43
NMI	0.23	0.71	0.21	0.42	0	0.23	0.24
ARI	0.17	0.51	-0.02	0.32	-0	0.01	0.02

3. Interprétation :

les performances des diverses méthodes de clustering sont meilleures de manière général en utilisant la représentation textuelle BERT car elle opère l'algorithme de NSP contrairement à RoBERTa, par conséquent étant la nature de notre dataset qui est composé d'articles scientifiques NSP permet de comprendre mieux les dépendances à plus long terme entre les phrases.

Pour les deux représentations, c'est le clustering spectral qui donne les meilleurs résultats cela est dû au fait qu'il fait un partitionnement sur un graphe de ce fait la proximité des classes n'est pas un obstacle dans la classification car il n'a aucune hypothèse de départ sur la taille ou forme des classes les distinguer contrairement à k-means qui donne de moins bon résultat étant donnée la forme arbitraire des classes et le déséquilibre en matière de taille des clusters.

Dataset BBC

1. BERT

	k-means	Spectral clustering	HDBSCAN	CAH Ward	CAH Single	CAH Complete	CAH Average
Accuracy	0.75	0.56	0.27	0.7	0.23	0.29	0.24
NMI	0.57	0.6	0.02	0.49	0.01	0.11	0.04
ARI	0.51	0.37	0.27	0.38	-0	0.01	-0

2. RoBERTa

	k-means	Spectral clustering	HDBSCAN	CAH Ward	CAH Single	CAH Complete	CAH Average
Accuracy	0.54	0.72	0.23	0.5	0.23	0.31	0.26
NMI	0.38	0.66	0.01	0.35	0	0.1	0.05
ARI	0.31	0.56	-0	0.26	-0	0.04	-0

3. Interprétation :

Nous pouvons remarquer que pour le dataset BBC les résultats de Bert sont meilleurs et plus précisément k-means, mais on atteint les meilleurs résultats avec RoBERTa et le clustering spectral. A partir des graphes, il est évident que les classes ont une forme arbitraire pour RoBERTa ce qui explique les très bons résultats du clustering spectral comparé au reste des méthodes. HDBSCAN étant une approche basée sur la densité, ses mauvais résultats mauvais sont dû à la proximité entre les classes ce qui a rendu difficile la détection des différentes zones denses.

Les résultats du CAH sont aussi mauvais voir très mauvais car il est basé sur les distances et en vue de la grande dimension du dataset et de sa densité, la construction hiérarchique aboutit à de mauvaises classifications

Les approches Tandem sur les données labélisées

Dataset BBC

I. PCA

1. BERT

	k-means	Spectral clustering	HDBSCAN	CAH Ward	CAH Single	CAH Complete	CAH Average
Accuracy	0.49	0.53	0.17	0.47	0.23	0.37	0.47
NMI	0.42	0.38	0.24	0.37	0.01	0.32	0.37
ARI	0.28	0.29	0.01	0.24	-0	0.17	0.22

2. RoBERTa

	k-means	Spectral clustering	HDBSCAN	CAH Ward	CAH Single	CAH Complete	CAH Average
Accuracy	0.39	0.49	0.19	0.41	0.23	0.35	0.36
NMI	0.29	0.28	0.23	0.26	0	0.23	0.29
ARI	0.19	0.2	0.01	0.15	-0	0.1	0.15

II. t-SNE

1. BERT

	k-means	Spectral clustering	HDBSCAN	CAH Ward	CAH Single	CAH Complete	CAH Average
Accuracy	0.94	0.54	0.47	0.94	0.4	0.87	0.88
NMI	0.83	0.58	0.56	0.82	0.35	0.75	0.75
ARI	0.86	0.36	0.37	0.85	0.16	0.72	0.72

2. RoBERTa

	k-means	Spectral clustering	HDBSCAN	CAH Ward	CAH Single	CAH Complete	CAH Average
Accuracy	0.87	0.74	0.27	0.86	0.44	0.8	0.92
NMI	0.73	0.64	0.44	0.72	0.48	0.63	0.78
ARI	0.74	0.57	0.1	0.72	0.22	0.6	0.81

III. UMAP

1. BERT

	k-means	Spectral clustering	HDBSCAN	CAH Ward	CAH Single	CAH Complete	CAH Average
Accuracy	0.93	0.56	0.21	0.92	0.76	0.93	0.93
NMI	0.82	0.55	0.21	0.8	0.73	0.81	0.81
ARI	0.85	0.36	0.1	0.82	0.65	0.83	0.83

2. RoBERTa

	k-means	Spectral clustering	HDBSCAN	CAH Ward	CAH Single	CAH Complete	CAH Average
Accuracy	0.9	0.58	0.28	0.89	0.47	0.87	0.92
NMI	0.78	0.59	0.46	0.77	0.49	0.75	0.81
ARI	0.76	0.37	0.19	0.74	0.25	0.73	0.81

IV. Autoencoder

1. BERT

	k-means	Spectral clustering	HDBSCAN	CAH Ward	CAH Single	CAH Complete	CAH Average
Accuracy	0.69	0.73	0.3	0.55	0.23	0.5	0.51
NMI	0.48	0.52	0.32	0.42	0.01	0.5	0.36
ARI	0.41	0.48	0.07	0.31	0	0.19	0.22

2. RoBERTa

	k-means	Spectral clustering	HDBSCAN	CAH Ward	CAH Single	CAH Complete	CAH Average
Accuracy	0.68	0.67	0.24	0.69	0.23	0.55	0.71
NMI	0.63	0.61	0.03	0.66	0.01	0.54	0.66
ARI	0.5	0.49	-0	0.55	-0	0.37	0.52

V. Interprétation :

pour le dataset BBC, nous constatons la même chose que pour le dataset classic4, cependant les meilleurs résultats sont obtenus cette fois avec la représentation BERT et la méthode tandem t-SNE/ CAH ward. Nous remarquons aussi que nous atteignons de très bon taux (0.9 pour ARI et 0.94 pour accuracy) ceci est dû à la nature du dataset. UMAP a donné aussi de trés bon résultats avec K-means. les méthodes UMAP et t-SNE donnent de bons résultats avec les méthodes k-means et CAH car elles permettent d'obtenir des clusters qui préserve la topologie de voisinage de l'espace d'origine.

HDBSCAN a des performances assez mauvaises à cause de la proximité des classes entre eux ce qui rend la classification avec l'approche de densité mauvaise.

Dataset Classic4

I. PCA

1. BERT

	k-means	Spectral clustering	HDBSCAN	CAH Ward	CAH Single	CAH Complete	CAH Average
Accuracy	0.51	0.5	0.43	0.62	0.45	0.58	0.43
NMI	0.46	0.46	0.24	0.45	0	0.44	0.24
ARI	0.28	0.27	0.02	0.32	-0	0.32	0.02

2. RoBERTa

	k-means	Spectral clustering	HDBSCAN	CAH Ward	CAH Single	CAH Complete	CAH Average
Accuracy	0.47	0.47	0.31	0.47	0.45	0.48	0.49
NMI	0.24	0.24	0.18	0.24	0	0.25	0.25
ARI	0.19	0.18	-0.08	0.19	-0	0.13	0.14

II. t-SNE

1. BERT

	k-means	Spectral clustering	HDBSCAN	CAH Ward	CAH Single	CAH Complete	CAH Average
Accuracy	0.75	0.59	0.63	0.77	0.79	0.77	0.77
NMI	0.69	0.52	0.57	0.74	0.78	0.73	0.74
ARI	0.51	0.29	0.4	0.53	0.62	0.52	0.53

2. RoBERTa

	k-means	Spectral clustering	HDBSCAN	CAH Ward	CAH Single	CAH Complete	CAH Average
Accuracy	0.56	0.6	0.77	0.55	0.77	0.56	0.77
NMI	0.55	0.54	0.73	0.57	0.73	0.55	0.73
ARI	0.39	0.37	0.52	0.41	0.52	0.39	0.52

III. UMAP

1. BERT

	k-means	Spectral clustering	HDBSCAN	CAH Ward	CAH Single	CAH Complete	CAH Average
Accuracy	0.77	0.59	0.61	0.77	0.77	0.77	0.77
NMI	0.74	0.47	0.57	0.74	0.74	0.74	0.74
ARI	0.53	0.26	0.39	0.53	0.53	0.53	0.53

2. RoBERTa

	k-means	Spectral clustering	HDBSCAN	CAH Ward	CAH Single	CAH Complete	CAH Average
Accuracy	0.77	0.68	0.61	0.77	0.77	0.77	0.77
NMI	0.73	0.53	0.59	0.73	0.73	0.73	0.73
ARI	0.52	0.34	0.4	0.52	0.52	0.52	0.52

IV. Autoencoder

1. BERT

	k-means	Spectral clustering	HDBSCAN	CAH Ward	CAH Single	CAH Complete	CAH Average
Accuracy	0.53	0.67	0.5	0.52	0.45	0.47	0.53
NMI	0.48	0.56	0.45	0.57	0	0.36	0.56
ARI	0.28	0.35	0.29	0.35	-0	0.12	0.34

2. RoBERTa

	k-means	Spectral clustering	HDBSCAN	CAH Ward	CAH Single	CAH Complete	CAH Average
Accuracy	0.75	0.63	0.5	0.77	0.45	0.58	0.52
NMI	0.6	0.59	0.46	0.72	0	0.48	0.56
ARI	0.43	0.46	0.29	0.52	-0	0.31	0.35

V. Interprétation :

Les diverses métriques de performances nous montrent que l'utilisation de la représentation textuelle bert et la méthode de clusetring CAH ward donne de bons résultats avec PCA, t-SNE, cependant l'autocondeur donne les meilleurs résultats avec RoBERTa et CAH. La meilleure approche pour ce dataset est UMAP / kmeans ou CAH. Ceci est expliqué par le fait que UMAP essaie de préserver la structure de voisinage ce qui permet de garder les instances similaires proches mais en plus de s'assurer que la distance entre classes soit significative. Par conséquent l'utilisation de bert et UMAP a permis d'avoir des clusters assez espaces de forme similaire et par conséquent kmeans et CAH ont pu donner les meilleures performances.

En conclusion, en comparant les résultats des approches tandem obtenues avec word2vec, glove, BERT et RoBERTa nous constatons que c'est glove qui nous permet d'avoir les meilleurs résultats pour les deux datasets

Les approches combinées sur les données labélisées

Dataset Classic4

1. BERT

	Reduced kmeans	Factorial Kmeans	DCN	DKM
Accuracy	0.52	0.7	0.76	
NMI	0.52	0.45	0.63	
ARI	0.32	0.35	0.45	

2. RoBERTa

	Reduced kmeans	Factorial Kmeans	DCN	DKM
Accuracy	0.6	0.53	0.39	
NMI	0.53	0.37	0.21	
ARI	0.39	0.23	0.08	

3. Interprétation :

Le DCN retourne de meilleurs résultats que Reduced kmeans Factorielle kmeans pour la représentation BERT. Alors que le Reduced kmeans performe mieux avec la représentation RoBERTa que Factorielle kmeans et le DCN.

Dataset BBC

1. BERT

	Reduced kmeans	Factorial Kmeans	DCN	DKM
Accuracy	0.49	0.58	0.36	0.59
NMI	0.41	0.41	0.2	0.47
ARI	0.28	0.32	0.13	0.36

2. RoBERTa

	Reduced kmeans	Factorial Kmeans	DCN	DKM
Accuracy	0.61	0.47	0.35	0.4
NMI	0.6	0.31	0.16	0.24
ARI	0.48	0.23	0.12	0.17

Interprétation :

DKM est meilleure que Reduced kmeans, Factorielle kmeans et DCN pour la représentation BERT. Le Reduced kmeans encore une fois performe mieux que les autres algorithmes avec la représentation RoBERTa.

En conclusion, on remarque que les méthodes tandem donnent de meilleurs résultats que les méthodes combinées pour les deux dataset Classic4 et BBC.

Les données non labélisées

1. Nous nous attendions à ce que BERT soit moins performant que RoBERTa sur les données Article1, étant donné que Roberta a été entraîné sur des articles de presse (ainsi que d'autres) alors que BERT n'a été entraîné que sur BOOK CORPUS et English Wikipedia, mais il semble que dans le cas des données originales, les performances sont similaires.

2. L'un des objectifs d'UMAP est que la distance entre les clusters de points soit significative. Cela signifie que les clusters peuvent se retrouver étalés avec une bonne quantité d'espace entre eux. En conséquence, les clusters eux-mêmes peuvent être visuellement plus compacts qu'avec t-SNE ou PCA, par exemple. C'est ce qu'on observe en comparant les d'UMAP des deux datasets avec les autres méthodes de réduction de dimensionalité.

3. En conséquence : PCA, t-SNE et les auto-encodeurs combinées avec Kmeans, Spectral clustering, CAH Ward et CAH complete étaient les mieux à souligner les différences entre les classes pour l'analyse Tandem.

4. Pour les méthodes de clustering combinées, le Reduced Kmeans a obtenu la meilleure séparation des classes avec le DCN et le DKM, mais ces deux derniers avaient une variance plus faible et les distances entre les points du clustering ont étaient considérablement réduites.