t-SNE and Autoencoder

Dejvis Toptani

2023-06-06

Datasets: bbc, Ecommerce, 20news

Number of classes: bbc 5, Ecommerce 4, 20news 20 (used only 5)

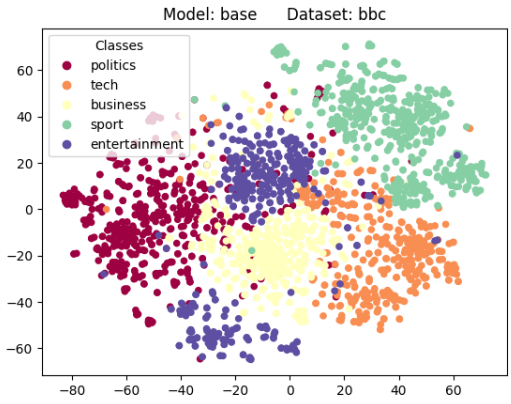
Embedding methods: BERT, tfidf, Word2Vec, fasttext

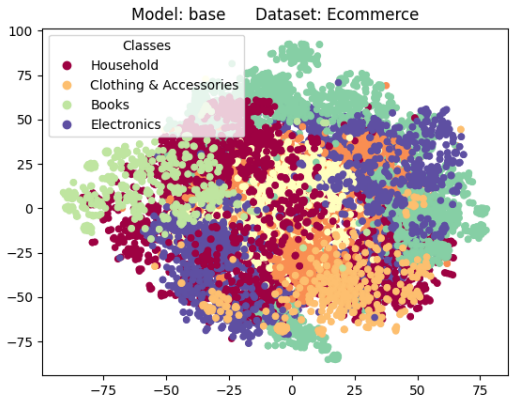
Dimensionality reduction: UMAP, t-SNE, Autoencoder

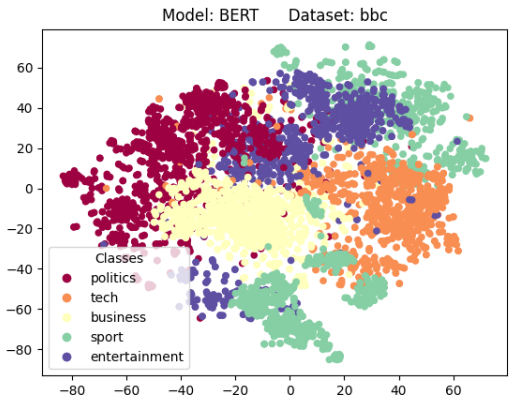
Algorithms: KMeans, DBSCAN, HDBSCAN, Meanshift, DBADV, DBHD, DPC, OPTICS, SNNDPC, SpectralClustering, SpectralACL

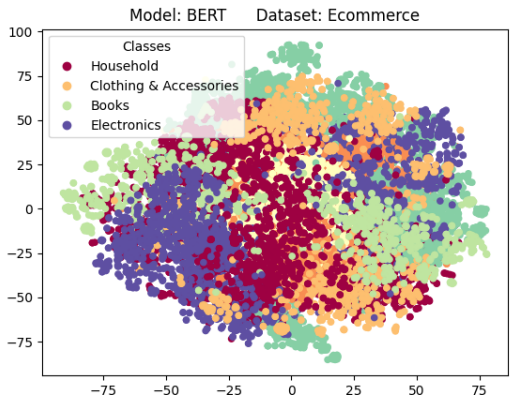
SVM Classifier for Evaluation

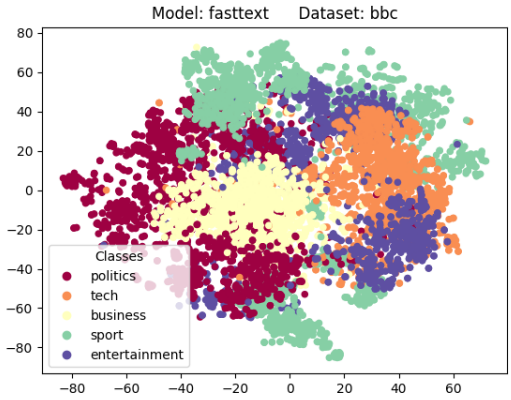
t-SNE Plots

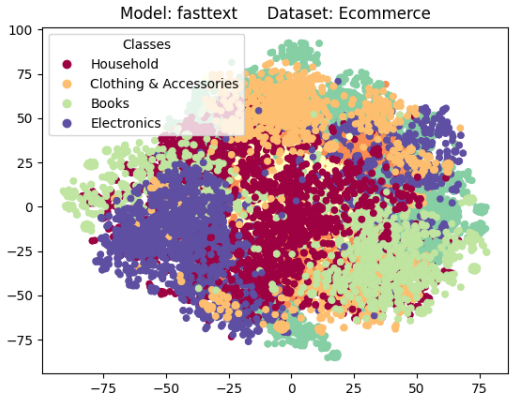


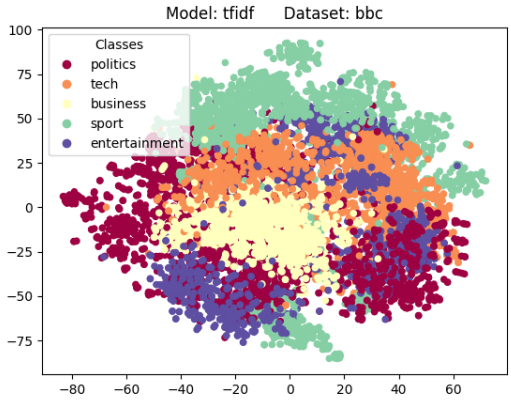


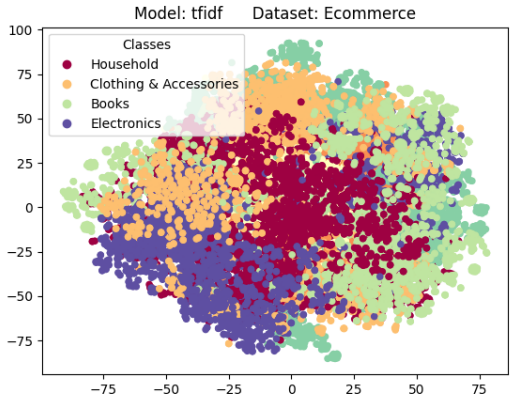


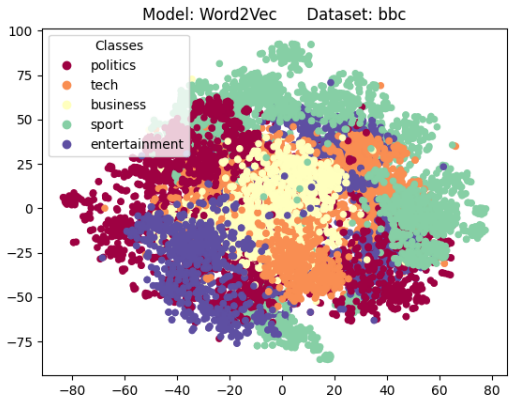


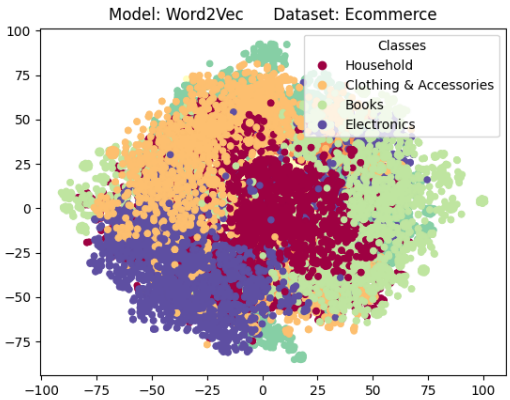

Model: fasttext Dataset: bbc
Classes
politics
tech
business
sport
entertainment


Model: fasttext Dataset: Ecommerce
Classes
Household
Clothing & Accessories
Books
Electronics

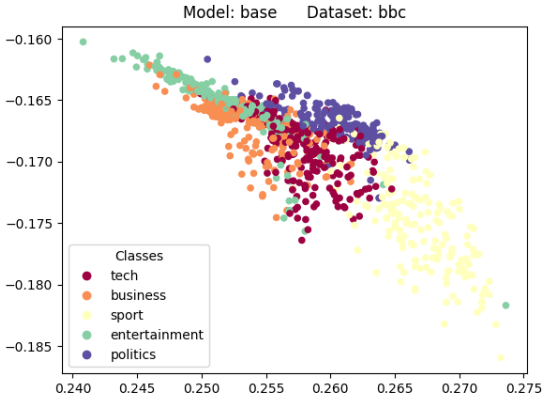

Model: tfidf Dataset: bbc
Classes
politics
tech
business
sport
entertainment

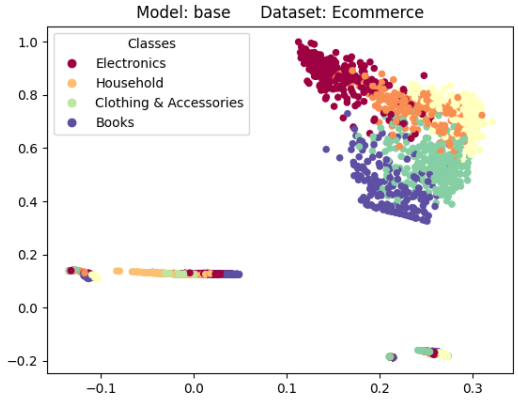

Model: tfidf Dataset: Ecommerce
Classes
Household
Clothing & Accessories
Books
Electronics

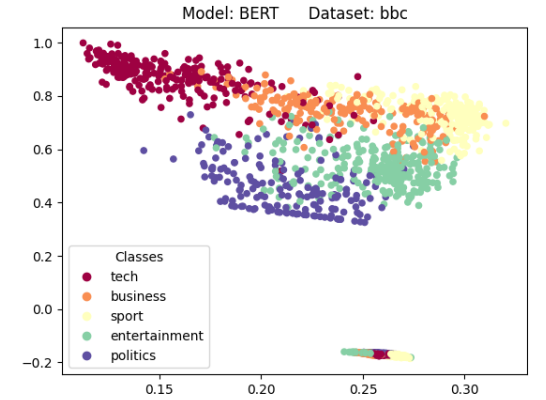

Model: Word2Vec Dataset: bbc
Classes
politics
tech
business
sport
entertainment

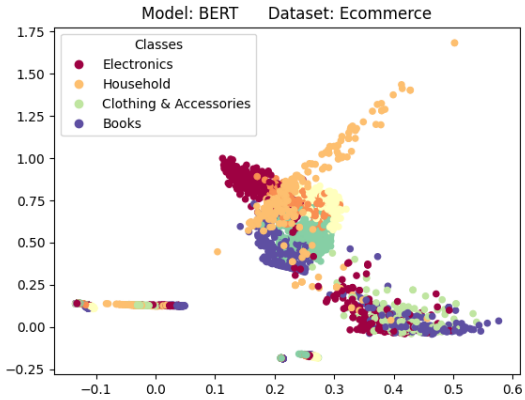

Model: Word2Vec Dataset: Ecommerce
Classes
Household
Clothing & Accessories
Books
Electronics

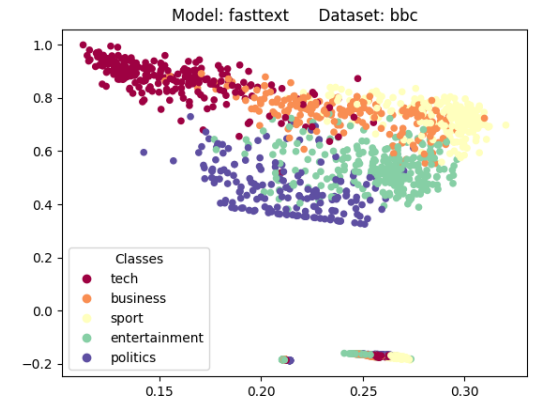
Autoencoder Plots

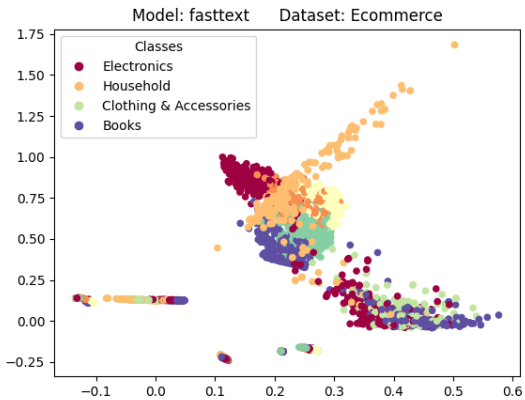


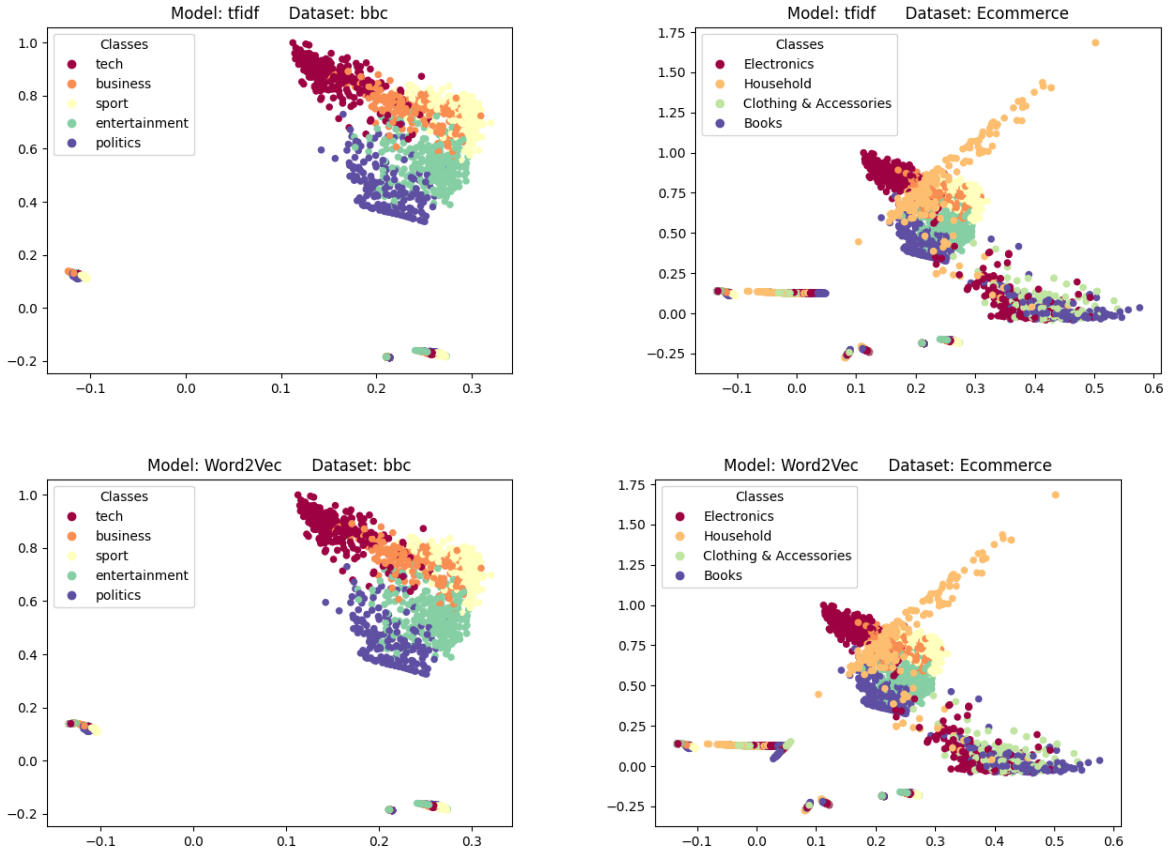












Clustering of bbc dataset with t-SNE

DBSCAN

min_samples	eps	tsne_n_components	NMI	embedding
5	1.000	2	0.204	BERT
5	1.000	2	0.204	BERT
5	1.000	2	0.204	BERT
5	1.000	2	0.212	fasttext
5	1.000	2	0.212	fasttext
5	1.000	2	0.212	fasttext
5	0.989	2	0.046	tfidf
5	0.985	2	0.046	tfidf
5	0.987	2	0.046	tfidf
5	1.000	2	0.226	word2vec
5	1.000	2	0.226	word2vec
5	0.999	2	0.226	word2vec

KMeans

n_clusters	tsne_n_components	NMI	embedding
5	2	0.840	BERT
5	2	0.840	BERT
5	2	0.840	BERT
5	2	0.811	fasttext
5	2	0.811	fasttext
5	2	0.810	fasttext
5	2	0.765	tfidf
5	2	0.762	tfidf
5	2	0.762	tfidf
5	2	0.800	word2vec
5	2	0.800	word2vec
5	2	0.800	word2vec

OPTICS

min_samples	xi	tsne_n_components	NMI	embedding
103	0.030	2	0.764	BERT
87	0.014	2	0.759	BERT
87	0.013	2	0.759	BERT
128	0.023	2	0.785	fasttext
180	0.010	2	0.784	fasttext
180	0.011	2	0.784	fasttext
156	0.012	2	0.680	tfidf
141	0.011	2	0.660	tfidf
181	0.010	2	0.640	tfidf
116	0.011	2	0.770	word2vec
116	0.011	2	0.770	word2vec
116	0.012	2	0.770	word2vec

DBHD

min_cluster_size	rho	beta	tsne_n_components	NMI	embedding
84	1.189	0.397	2	0.812	BERT
84	1.170	0.395	2	0.810	BERT
84	1.187	0.394	2	0.810	BERT
192	1.948	0.108	2	0.831	fasttext
190	1.970	0.101	2	0.831	fasttext
191	1.892	0.109	2	0.831	fasttext
25	0.649	0.120	2	0.638	tfidf
36	1.205	0.112	2	0.611	tfidf
32	0.117	0.131	2	0.603	tfidf
77	1.292	0.364	2	0.763	word2vec
108	1.434	0.253	2	0.760	word2vec
113	1.366	0.284	2	0.753	word2vec

Meanshift

bandwidth	tsne_n_components	NMI	embedding
0.907	2	0.373	BERT
0.910	2	0.373	BERT
0.909	2	0.373	BERT
0.910	2	0.374	fasttext
0.909	2	0.374	fasttext
0.909	2	0.374	fasttext
0.910	2	0.362	tfidf
0.909	2	0.362	tfidf
0.909	2	0.362	tfidf
0.907	2	0.374	word2vec
0.907	2	0.374	word2vec
0.909	2	0.374	word2vec

SNN-DPC

k	nc	tsne_n_components	NMI	embedding
29	6	2	0.816	BERT
29	6	2	0.816	BERT
29	6	2	0.816	BERT
8	6	2	0.765	fasttext
8	6	2	0.765	fasttext
8	6	2	0.765	fasttext
45	6	2	0.679	tfidf
45	6	2	0.679	tfidf
45	6	2	0.679	tfidf
24	6	2	0.749	word2vec
24	6	2	0.749	word2vec
24	6	2	0.749	word2vec

HDBSCAN

min_cluster_size	min_samples	tsne_n_components	NMI	embedding
125	29	2	0.794	BERT
146	29	2	0.794	BERT
141	29	2	0.794	BERT
131	35	2	0.736	fasttext
126	35	2	0.736	fasttext
86	35	2	0.736	fasttext
117	18	2	0.650	tfidf
112	18	2	0.650	tfidf
102	18	2	0.650	tfidf
103	16	2	0.709	word2vec
107	16	2	0.709	word2vec
134	16	2	0.709	word2vec

SpectralACL

n_clusters	epsilon	tsne_n_components	NMI	embedding
4	0.239	8	0	BERT
4	0.526	8	0	BERT
4	0.594	16	0	BERT
4	0.863	16	0	fasttext
4	0.383	4	0	fasttext
4	0.406	4	0	fasttext
4	0.897	2	0	tfidf
4	0.151	2	0	tfidf
4	0.872	16	0	tfidf
4	0.703	2	0	word2vec
4	0.273	16	0	word2vec
4	0.785	2	0	word2vec

DBADV

perplexity	MinPts	probability	tsne_n_components	NMI	embedding
12	30	0.997	2	0.666	BERT
12	30	0.997	2	0.666	BERT
12	30	0.997	2	0.666	BERT
9	22	0.997	2	0.651	fasttext
9	22	0.997	2	0.651	fasttext
9	22	0.997	2	0.651	fasttext
2	1	0.997	2	0.427	tfidf
2	1	0.997	2	0.427	tfidf
2	1	0.997	2	0.427	tfidf
8	20	0.997	2	0.498	word2vec
3	7	0.997	2	0.479	word2vec
3	7	0.997	2	0.479	word2vec

DPC

density_threshold	distance_threshold	tsne_n_components	NMI	embedding
0.352	0.998	2	0.382	BERT
0.029	0.998	2	0.382	BERT
0.011	0.998	2	0.382	BERT
0.620	1.000	2	0.384	fasttext
0.636	1.000	2	0.384	fasttext
0.588	1.000	2	0.384	fasttext
0.286	0.999	2	0.365	tfidf
0.177	0.999	2	0.365	tfidf
0.151	0.999	2	0.365	tfidf
0.860	1.000	2	0.386	word2vec
0.915	1.000	2	0.386	word2vec
0.863	1.000	2	0.386	word2vec