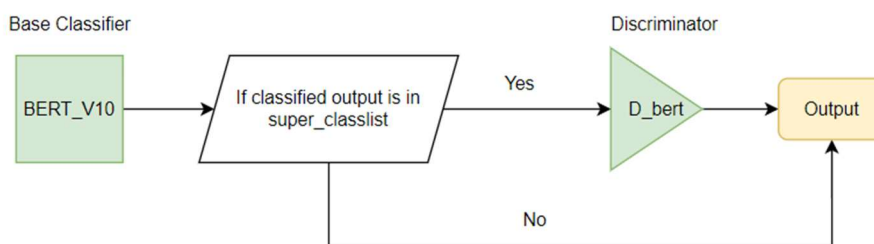


Fibe Hack the Vibe Season 2

Team - ml_dtypes

The final submission for the text classification problem consisted of the following approach:

1. **Initial Pretrained Embeddings & MLP Classifier:** Started with pretrained BERT embeddings and a custom MLP classification head, achieving a 40% F1 score.
2. **KNN Voting Method:** Moved to a KNN-based voting approach using cosine similarity on vector embeddings, improving the F1 score to 65%.
3. **DistilBERT Model:** Due to resource constraints, switched to DistilBERT (an encoder-only model), reaching an F1 score of 84%.
4. **BERT Base Classifier & DistilBERT Discriminator:** Used BERT as the base classifier and DistilBERT as a discriminator to differentiate between semantically similar class pairs (e.g., business & finance, personal finance, real estate). This approach crossed the 85% F1 score mark.



5. **RoBERTa Model:** Switched to RoBERTa for successive training. The first 2 epochs were on a train subset, followed by 1.5 epochs on the full training set with a slow learning rate. This boosted performance to 86%.
6. **Text Cleaning & Augmentation:** Applied text cleaning and augmentation on the test set before inference, achieving the final F1 score of 87.64%. This approach delivered the best result for the competition.

Confusion matrix to see mismatch and between what labels are the models switching during training. This is done to actually understand which labels is the model finding hard to classify.

For BERT MODEL –

		Confusion Matrix																											
y_test \ y_pred	technology and computing	0	342	34	88	109	123	6	40	481	206	20	115	53	20	81	42	33	43	74	19	30	29	10	44	14	13		
	academic interests	47	0	7	14	9	7	2	1	29	82	7	21	14	18	52	40	65	13	15	15	7	13	0	9	0	12		
	hobbies and interests	204	94	0	174	131	185	0	69	64	174	253	79	297	250	102	132	205	97	252	72	232	207	2	259	18	48		
	family and relationships	30	73	21	0	13	16	1	1	18	54	39	11	15	29	47	27	22	10	42	17	219	27	1	97	21	11		
	sports	45	42	14	16	0	19	3	4	182	350	15	63	10	50	31	43	19	7	18	56	20	27	12	22	58	172		
	personal finance	44	14	46	25	20	0	0	55	30	30	404	30	203	69	38	29	21	23	82	5	110	58	0	62	2	22		
	health	5	3	0	0	1	0	0	5	1	0	0	0	0	0	0	1	2	1	0	0	0	0	0	0	0	0		
	books and literature	35	9	24	7	12	97	5	0	23	47	52	4	23	87	31	24	54	17	23	5	44	24	0	12	1	3		
	television	71	28	14	2	38	26	1	3	0	33	13	19	27	6	18	4	10	10	25	21	25	15	1	13	20	14		
	news and politics	27	117	13	21	101	35	0	9	51	0	33	41	59	97	62	24	30	32	53	9	60	25	2	10	3	81		
	food and drinks	1	24	16	14	8	25	0	5	18	22	0	5	136	67	35	13	11	17	121	5	35	10	0	19	1	0		
	pets	17	51	4	3	106	7	0	0	22	22	5	0	3	7	21	9	5	5	7	10	1	5	0	11	11	119		
	style and fashion	11	9	16	7	1	35	0	4	34	16	54	7	0	19	17	8	9	7	54	6	44	9	0	6	3	3		
	business and finance	22	62	79	14	89	44	0	27	11	60	111	35	55	0	41	41	306	27	118	10	51	48	0	57	2	40		
	travel	31	184	24	31	57	40	1	8	32	104	78	23	71	42	0	48	72	77	79	29	30	89	1	11	14	29		
	arts and culture	9	89	16	14	21	22	2	7	11	15	14	13	13	31	45	0	96	25	25	52	22	30	0	20	2	9		
	movies	46	232	107	26	53	18	2	27	38	74	17	12	36	123	103	103	0	42	59	10	43	39	0	48	1	8		
	healthy living	17	34	12	7	5	8	1	4	10	11	25	3	9	15	27	28	14	0	13	11	16	30	0	13	1	3		
	video gaming	9	15	13	4	6	4	0	1	12	9	70	8	50	41	19	15	9	5	0	4	26	12	0	15	0	2		
	shopping	17	64	19	18	61	20	0	1	44	39	15	35	9	11	27	51	8	13	77	9	14	35	0	108	29	33		
	music and audio	7	10	20	45	7	20	0	2	77	11	33	5	36	20	8	6	9	7	20	7	0	15	0	16	2	8		
	home and garden	24	26	35	35	17	82	0	10	22	64	12	22	26	58	55	82	43	44	26	33	80	0	0	70	7	30		
	automotives	24	1	0	0	17	2	0	3	8	0	0	1	0	0	1	0	0	0	0	0	1	0	1	0	1	0		
	careers	3	9	21	19	9	54	0	2	15	2	49	2	31	26	30	14	22	12	49	62	24	18	0	0	1	4		
	real estate	4	3	4	12	66	0	0	0	52	1	1	29	19	5	4	3	2	0	17	23	8	12	0	3	0	12		
	pharmaceuticals, conditions, and symptoms	4	1	2	2	47	17	0	0	8	13	0	0	2	21	11	6	3	2	14	0	24	1	1	0	0	0		
y_test	technology and computing	0	342	34	88	109	123	6	40	481	206	20	115	53	20	81	42	33	43	74	19	30	29	10	44	14	13		
	academic interests	47	0	7	14	9	7	2	1	29	82	7	21	14	18	52	40	65	13	15	15	7	13	0	9	0	12		
	hobbies and interests	204	94	0	174	131	185	0	69	64	174	253	79	297	250	102	132	205	97	252	72	232	207	2	259	18	48		
	family and relationships	30	73	21	0	13	16	1	1	18	54	39	11	15	29	47	27	22	10	42	17	219	27	1	97	21	11		
	sports	45	42	14	16	0	19	3	4	182	350	15	63	10	50	31	43	19	7	18	56	20	27	12	22	58	172		
	personal finance	44	14	46	25	20	0	0	55	30	30	404	30	203	69	38	29	21	23	82	5	110	58	0	62	2	22		
	health	5	3	0	0	1	0	0	5	1	0	0	0	0	0	0	1	2	1	0	0	0	0	0	0	0	0		
	books and literature	35	9	24	7	12	97	5	0	23	47	52	4	23	87	31	24	54	17	23	5	44	24	0	12	1	3		
	television	71	28	14	2	38	26	1	3	0	33	13	19	27	6	18	4	10	10	25	21	25	15	1	13	20	14		
	news and politics	27	117	13	21	101	35	0	9	51	0	33	41	59	97	62	24	30	32	53	9	60	25	2	10	3	81		
	food and drinks	1	24	16	14	8	25	0	5	18	22	0	5	136	67	35	13	11	17	121	5	35	10	0	19	1	0		
	pets	17	51	4	3	106	7	0	0	22	22	5	0	3	7	21	9	5	5	7	10	1	5	0	11	11	119		
	style and fashion	11	9	16	7	1	35	0	4	34	16	54	7	0	19	17	8	9	7	54	6	44	9	0	6	3	3		
	business and finance	22	62	79	14	89	44	0	27	11	60	111	35	55	0	41	41	306	27	118	10	51	48	0	57	2	40		
	travel	31	184	24	31	57	40	1	8	32	104	78	23	71	42	0	48	72	77	79	29	30	89	1	11	14	29		
	arts and culture	9	89	16	14	21	22	2	7	11	15	14	13	13	31	45	0	96	25	25	52	22	30	0	20	2	9		
	movies	46	232	107	26	53	18	2	27	38	74	17	12	36	123	103	103	0	42	59	10	43	39	0	48	1	8		
	healthy living	17	34	12	7	5	8	1	4	10	11	25	3	9	15	27	28	14	0	13	11	16	30	0	13	1	3		
	video gaming	9	15	13	4	6	4	0	1	12	9	70	8	50	41	19	15	9	5	0	4	26	12	0	15	0	2		
	shopping	17	64	19	18	61	20	0	1	44	39	15	35	9	11	27	51	8	13	77	9	14	35	0	108	29	33		
	music and audio	7	10	20	45	7	20	0	2	77	11	33	5	36	20	8	6	9	7	20	7	0	15	0	16	2	8		
	home and garden	24	26	35	35	17	82	0	10	22	64	12	22	26	58	55	82	43	44	26	33	80	0	0	70	7	30		
	automotives	24	1	0	0	17	2	0	3	8	0	0	1	0	0	1	0	0	0	0	0	1	0	1	0	1	0		
	careers	3	9	21	19	9	54	0	2	15	2	49	2	31	26	30	14	22	12	49	62	24	18	0	0	1	4		
	real estate	4	3	4	12	66	0	0	0	52	1	1	29	19	5	4	3	2	0	17	23	8	12	0	3	0	12		
	pharmaceuticals, conditions, and symptoms	4	1	2	2	47	17	0	0	8	13	0	0	2	21	11	6	3	2	14	0	24	1	1	0	0	0		
y_test \ y_pred	technology and computing	0	342	34	88	109	123	6	40	481	206	20	115	53	20	81	42	33	43	74	19	30	29	10	44	14	13		
	academic interests	47	0	7	14	9	7	2	1	29	82	7	21	14	18	52	40	65	13	15	15	7	13	0	9	0	12		
	hobbies and interests	204	94	0	174	131	185	0	69	64	174	253	79	297	250	102	132	205	97	252	72	232	207	2	259	18	48		
	family and relationships	30	73	21	0	13	16	1	1	18	54	39	11	15	29	47	27	22	10	42	17	219	27	1	97	21	11		
	sports	45	42	14	16	0	19	3	4	182	350	15	63	10	50	31	43	19	7	18	56	20	27	12	22	58	172		
	personal finance	44	14	46	25	20	0	0	55	30	30	404	30	203	69	38	29	21	23	82	5	110	58	0	62	2	22		
	health	5	3	0	0	1	0	0	5	1	0	0	0	0	0	0	1	2	1	0	0	0	0	0	0	0	0		
	books and literature	35	9	24	7	12	97	5	0	23	47	52	4	23	87	31	24	54	17	23	5	44	24	0	12	1	3		
	television	71	28	14	2	38	26	1	3	0	33	13	19	27	6	18	4	10	10	25	21	25	15	1	13	20	14		
	news and politics	27	117	13	21	101	35	0	9	51	0	33	41	59	97	62	24	30	32	53	9	60	25	2	10	3	81		
	food and drinks	1	24	16	14	8	25	0	5	18	22	0	5	136	67	35	13	11	17	121	5	35	10	0	19	1	0		
	pets	17	51	4	3	106	7	0	0	22	22	5	0	3	7	21	9	5	5	7	10	1	5	0	11	11	119		
	style and fashion	11	9	16	7	1	35	0	4	34	16	54	7	0	19	17	8	9	7	54	6	44	9	0	6	3	3		
	business and finance	22	62	79	14	89	44	0	27	11	60	111	35	55	0	41	41	306	27	118	10	51	48	0	57	2	40		
	travel	31	184	24	31	57	40	1	8																				

Final Model - [roberta-base-v1](#)