DMT HW3: Natural Language Processing

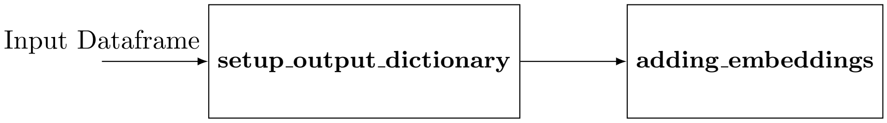
Alessandro Quattrociochi - 1609286

Tansel Simsek - 1942297

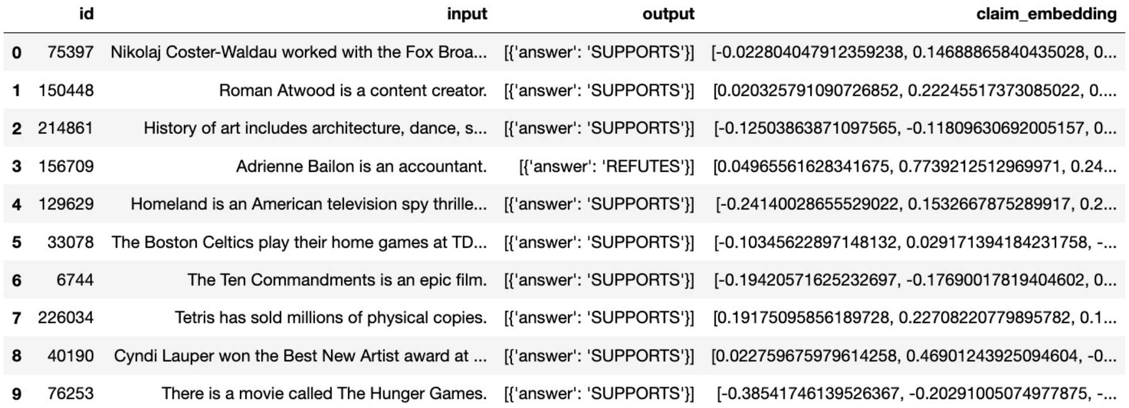
Part 1.1

Pre-processing and embeddings

In this part and in 2.1 we focus on text processing and some operations that led us to define the final form of the dataset on which, subsequently we applied the classification algorithms. We have three different datasets: train, test and dev, presented as list of dictionaries. For each dictionary, we considered only 3 fields: id, input, output while for the output we considered only the first dictionary (in case of multiples), particularly the key "answer" which had only two possible outcomes: SUPPORT and REFUTES. Although the initial form is, as mentioned, a list of dictionaries, we converted it to dataframe to have faster execution speed and to take advantage of the that structure. After the conversion to dataframe, we created two different functions: **setup_output_dictionary**, **adding_embeddings**. These functions, passed dataframe as input return as output a new one with the modification. The second function is the one of interest, as it exploits the SentenceTransformer library and the *paraphrase-distilroberta-base-v1* model to return an array of 768 elements, indicating the logarithm of similarity probability. To increase effectiveness, we made extensive use of the map() function to avoid loops on every line, especially when fitting models.



At the end of step 1.2, the final table was saved in two formats *.pickle* to have a save checkpoint and as *.jsonl* as required by the assignment.
Below is a printout of the train dataset after applying the two functions.



	id	input	output	claim_embedding
0	75397	Nikolaj Coster-Waldau worked with the Fox Broa...	[{'answer': 'SUPPORTS'}]	[-0.022804047912359238, 0.14688865840435028, 0...
1	150448	Roman Atwood is a content creator.	[{'answer': 'SUPPORTS'}]	[0.020325791090726852, 0.22245517373085022, 0....
2	214861	History of art includes architecture, dance, s...	[{'answer': 'SUPPORTS'}]	[-0.12503863871097565, -0.11809630692005157, 0...
3	156709	Adrienne Bailon is an accountant.	[{'answer': 'REFUTES'}]	[0.04965561628341675, 0.7739212512969971, 0.24...
4	129629	Homeland is an American television spy thrille...	[{'answer': 'SUPPORTS'}]	[-0.24140028655529022, 0.1532667875289917, 0.2...
5	33078	The Boston Celtics play their home games at TD...	[{'answer': 'SUPPORTS'}]	[-0.10345622897148132, 0.029171394184231758, -...
6	6744	The Ten Commandments is an epic film.	[{'answer': 'SUPPORTS'}]	[-0.19420571625232697, -0.17690017819404602, 0...
7	226034	Tetris has sold millions of physical copies.	[{'answer': 'SUPPORTS'}]	[0.19175095856189728, 0.22708220779895782, 0.1...
8	40190	Cyndi Lauper won the Best New Artist award at ...	[{'answer': 'SUPPORTS'}]	[0.022759675979614258, 0.46901243925094604, -0...
9	76253	There is a movie called The Hunger Games.	[{'answer': 'SUPPORTS'}]	[-0.38541746139526367, -0.20291005074977875, -...

Figure 1. Final dataset for 1.1

Part 1.2

We have picked 2 algorithms, 1st one as Logistic Regression and 2nd one as knn. The details about selected parameters and best parameters found after hyper-parameter tuning will be given below.

Grid Of Parameters

- **Logistic Regression**
  solver : ['newton-cg', 'lbfgs', 'liblinear'],
  penalty : ['l1', 'l2', 'elasticnet'],
  C : loguniform(1e-5, 100)

- **knn**
  n_neighbors : [13,15,17]
  weights : ['uniform', 'distance']
  metric : ['euclidean', 'manhattan']

Best Configurations for Each Classifier

- **Logistic Regression**
  'C': 1.1700993456412832, 'penalty': 'l1', 'solver': 'liblinear'

- **knn**
  'weights': 'distance', 'n_neighbors': 17, 'metric': 'manhattan'

Confusion Matrices

Classifier 1	**Predicted Refutes**	**Predicted Supports**	**Recall**
Actual Refutes	3735	1517	0.711
Actual Supports	1339	3853	0.742
Precision	0.736	0.718	0.727

Classifier 2	**Predicted Refutes**	**Predicted Supports**	**Recall**
Actual Refutes	3332	1920	0.634
Actual Supports	1510	3682	0.709
Precision	0.688	0.657	0.672

From the results above, We can state that Logistic Regression performs much better than knn for all recall, precision and accuracy metrices.

Hyper-parameter optimisation

First of all, it is fundamental to mention about data preprocessing that lead to find the above results. Because of the unbalanced train dataset (73% as Supports, 27% as Refutes), We have decided to get only 27% of the 0 labeled train data to acquire fair precision. Randomized Search algorithm with 5-cross validation is applied in both Logistic Regression and knn algorithm. Additionally, number of iterations taken as 500 for the Logistic Regression. Because of the cost of computational time, different k values are tried in different small datasets then the most frequent values are picked to construct our parameter space for k in knn algorithm.

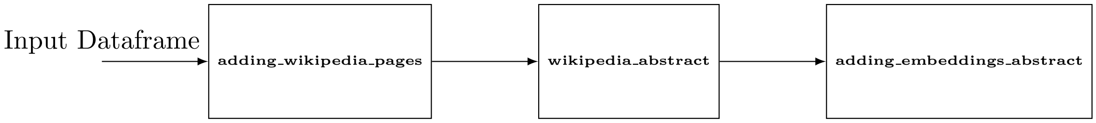
if the goal was to be sure to correctly catch all REFUTES claims (i.e. label them as REFUTES), which metric should you look at? And what trivial solution could be adopted?

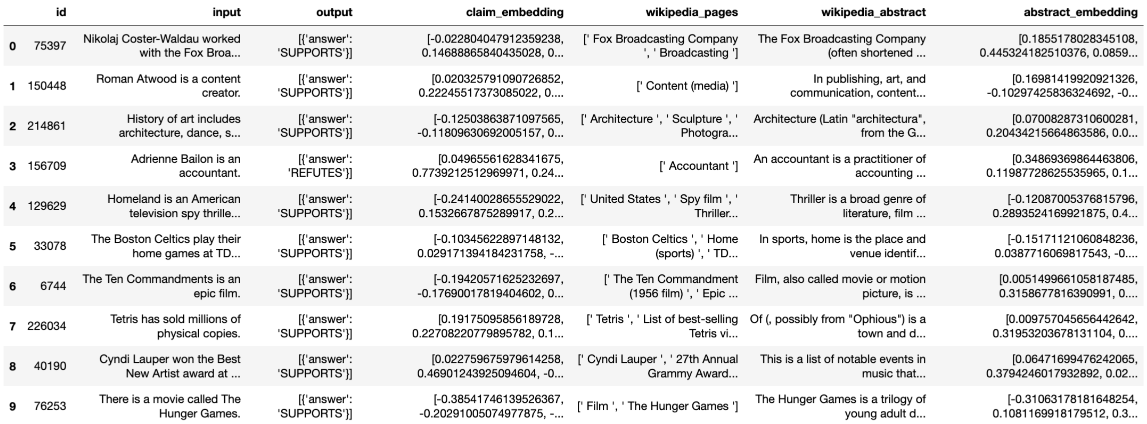
In that case, we should look at the True Negative Recall rate and pick the classifier with the highest TNR rate which is Classifier 1 (Logistic Regression) for our setting while trial solution can be taken as to collect many REFUTES labeled classes with providing few labeled SUPPORTS classes to increase TNR rate for REFUTES.

Part 2.1

Pre-processing and embeddings

This second part follows the same pattern as part 1.1 In fact, also in this case we have maintained the approach of writing functions that would take the dataset from part_1 output data folder and from those we have added the last three columns related to wikipedia pages, abtracts and finally embedding for the abstracts from wikipedia. In more detail, we used the pre-trained model ”*hf_e2e_entity_linking_wiki_abs*” and from the outputs, after fitting GENRE library model, we selected only the predicted pages in square brackets. For each claim we created a list that contained all the resulting pages. After creating the column related to wikipedia pages, we used the file *abstract_kilt_knowledgesource.json* to query the abstract related to the previously found wikipedia pages, stored and finally ranked in descending order of length. At the end, we reapplied the function defined in Section 1.1 to compute the embeddings for each claim. For this last case, the maximum length of the input sequence was set equal to 256. As before, we appended to each row an array of size 768 containing the logarithm of the similarity probability. For this computationally intensive part of the homework, we used GPU accelerators and divided the datasets into batches of 1000 samples each. Nevertheless, in the final version of the code we report the driver code version without the split. Also in this case we report a screen of the final version of the dataset, subsequently converted in jsonl.





	id	input	output	claim_embedding	wikipedia_pages	wikipedia_abstract	abstract_embedding
0	75397	Nikolaj Coster-Waldau worked with the Fox Broa...	[{'answer': 'SUPPORTS'}]	[-0.022804047912359238, 0.14688865840435028, 0...	[' Fox Broadcasting Company ', ' Broadcasting ']	The Fox Broadcasting Company (often shortened ...	[0.1855178028345108, 0.445324182510376, 0.0859...
1	150448	Roman Atwood is a content creator.	[{'answer': 'SUPPORTS'}]	[0.020325791090726852, 0.22245517373085022, 0....	[' Content (media) ']	In publishing, art, and communication, content...	[0.16981419920921326, -0.10297425836324692, -0...
2	214861	History of art includes architecture, dance, s...	[{'answer': 'SUPPORTS'}]	[-0.12503863871097565, -0.11809630692005157, 0...	[' Architecture ', ' Sculpture ', ' Photogra...	Architecture (Latin "architectura", from the G...	[0.07008287310600281, 0.20434215664863586, 0.0...
3	156709	Adrienne Bailon is an accountant.	[{'answer': 'REFUTES'}]	[0.04965561628341675, 0.7739212512969971, 0.24...	[' Accountant ']	An accountant is a practitioner of accounting ...	[0.34869369864463806, 0.11987728625535965, 0.1...
4	129629	Homeland is an American television spy thrille...	[{'answer': 'SUPPORTS'}]	[-0.24140028655529022, 0.1532667875289917, 0.2...	[' United States ', ' Spy film ', ' Thriller...	Thriller is a broad genre of literature, film ...	[-0.12087005376815796, 0.2893524169921875, 0.4...
5	33078	The Boston Celtics play their home games at TD...	[{'answer': 'SUPPORTS'}]	[-0.10345622897148132, 0.029171394184231758, -...	[' Boston Celtics ', ' Home (sports) ', ' TD...	In sports, home is the place and venue identif...	[-0.15171121060848236, 0.0387716069817543, -0....
6	6744	The Ten Commandments is an epic film.	[{'answer': 'SUPPORTS'}]	[-0.19420571625232697, -0.17690017819404602, 0...	[' The Ten Commandment (1956 film) ', ' Epic ...	Film, also called movie or motion picture, is ...	[0.0051499661058187485, 0.3158677816390991, 0....
7	226034	Tetris has sold millions of physical copies.	[{'answer': 'SUPPORTS'}]	[0.19175095856189728, 0.22708220779895782, 0.1...	[' Tetris ', ' List of best-selling Tetris vi...	Of (, possibly from "Ophious") is a town and d...	[0.009757045656442642, 0.31953203678131104, 0....
8	40190	Cyndi Lauper won the Best New Artist award at ...	[{'answer': 'SUPPORTS'}]	[0.022759675979614258, 0.46901243925094604, -0...	[' Cyndi Lauper ', ' 27th Annual Grammy Award...	This is a list of notable events in music that...	[0.06471699476242065, 0.3794246017932892, 0.02...
9	76253	There is a movie called The Hunger Games.	[{'answer': 'SUPPORTS'}]	[-0.38541746139526367, -0.20291005074977875, -...	[' Film ', ' The Hunger Games ']	The Hunger Games is a trilogy of young adult d...	[-0.31063178181648254, 0.1081169918179512, 0.3...

Figure 2. Final dataset for part 2.1

Part 2.2

We have picked 2 algorithms, 1st one as Logistic Regression and 2nd one as knn. The details about selected parameters and best parameters found after hyper-parameter tuning will be given below.

Grid Of Parameters

- **Logistic Regression**
  solver : ['newton-cg', 'lbfgs', 'liblinear'],
  penalty : ['l1', 'l2', 'elasticnet'],
  C : loguniform(1e-5, 100)

- **knn**
  n_neighbors : [13,15,17]
  weights : ['uniform', 'distance']
  metric : ['euclidean', 'manhattan']

Best Configurations for Each Classifier

- **Logistic Regression**
  'C': 0.4115283426435553, 'penalty': 'l1', 'solver': 'liblinear'

- **knn**
  'weights': 'distance', 'n_neighbors': 15, 'metric': 'manhattan'

Confusion Matrices

Classifier 1	**Predicted Refutes**	**Predicted Supports**	**Recall**
Actual Refutes	3683	1569	0.701
Actual Supports	1300	3892	0.750
Precision	0.739	0.713	0.725

Classifier 2	**Predicted Refutes**	**Predicted Supports**	**Recall**
Actual Refutes	3090	2162	0.588
Actual Supports	1731	3461	0.667
Precision	0.641	0.616	0.627

From the results above, We can state that Logistic Regression performs much better than knn for all recall, precision and accuracy metrices.

Hyper-parameter optimisation

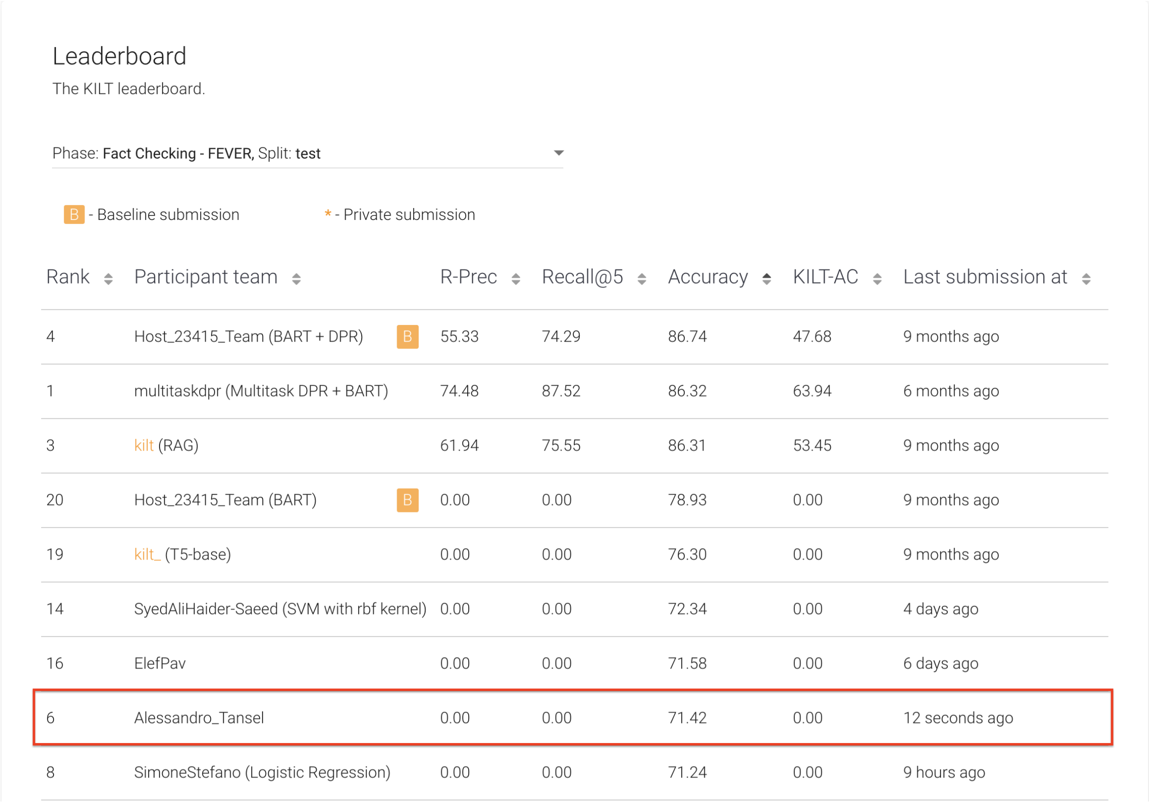
Randomized Search algorithm with 5-cross validation is applied both Logistic Regression and knn algorithm. Additionally, number of iterations taken as 500 for the Logistic Regression. Because of the cost of computational time, different k values are tried in different small datasets then the most frequent values are picked to construct our parameter space for k in knn algorithm.

Bonus Part

Prediction file used

new_set_pred_1.jsonl

Result



Leaderboard

The KILT leaderboard.

Phase: Fact Checking - FEVER, Split: test

B - Baseline submission   * - Private submission

Rank	Participant team		R-Prec	Recall@5	Accuracy	KILT-AC	Last submission at
4	Host_23415_Team (BART + DPR)	B	55.33	74.29	86.74	47.68	9 months ago
1	multitaskdpr (Multitask DPR + BART)		74.48	87.52	86.32	63.94	6 months ago
3	kilt (RAG)		61.94	75.55	86.31	53.45	9 months ago
20	Host_23415_Team (BART)	B	0.00	0.00	78.93	0.00	9 months ago
19	kilt_ (T5-base)		0.00	0.00	76.30	0.00	9 months ago
14	SyedAliHaider-Saeed (SVM with rbf kernel)		0.00	0.00	72.34	0.00	4 days ago
16	ElefPav		0.00	0.00	71.58	0.00	6 days ago
6	Alessandro_Tansel		0.00	0.00	71.42	0.00	12 seconds ago
8	SimoneStefano (Logistic Regression)		0.00	0.00	71.24	0.00	9 hours ago

Figure 3. Ranking and score in the leader board