Part 1: Text Classification

1. Task: 3 Boolean tasks (InfoTheory, CompVis and Math labels)
2. Algorithm: BERT and Logistic Regression.
3. Tokenizers: Tokenizers used are Bert Tokenizer and LemmaTokenizer.
4. Data size: training on the first 1000 records in the training set and training on all the records in the training set.

BERT
Tokenizer : BERT Tokenizer

Labels ///// Metrics	InfoTheory (1000 records)	InfoTheory (All records)	CompVis (1000)	CompVis (All records)	Math (1000)	Math (all records)
Accuracy	81.60%	92.69%	89.06%	89.06%	69.86%	69.86%
Precision (Macro)	40.8%	88.5%	44.53%	44.53%	34.93%	34.90%
Recall	50%	86.64%	50%	50%	50%	50%
F1 score (Macro)	44.90%	87.53%	47.10%	47%	41.12%	41%

Tokenizer : LemmaTokenizer

Labels ///// Metrics	InfoTheory (1000 records)	InfoTheory (All records)	CompVis (1000)	CompVis (All records)	Math (1000)	Math (all records)
Accuracy	81.62%	92.69%	89.06%	89.06%	69.86%	69.86%
Precision (Macro)	40.81%	88.5%	44.53%	44.53%	34.93%	34.90%
Recall	50%	86.64%	50%	50%	50%	50%
F1 score (Macro)	44.94%	87.53%	47.10%	47%	41.12%	41%

- Using BERT, output predictions don't change with different preprocessing because BERT uses BPE (Byte pair encoding to shrink its vocab size).
- Because of the long training time, I choose only 16 as a max length for input. Max tokens which I received were 745 after including [CLS] and [SEP] to the encoded tokens. BERT can work up to a maximum of 512 tokens. Because of the system crashing and making too long with 512 tokens, I chose 16 to make it simple. I may miss so much essential information with

the input length as 16, but it performed well on InfoTheory all labels and performed ok on Math and CompVis labels.

- It worked almost the same with different preprocessing when the training was performed on the first 1000 records, maybe because all the training labels belong to the same class.

Statistical Model: Logistic Regression
Tokenizer: LemmaTokenizer

Labels ///// Metrics	**InfoTheory (1000 records)**	**InfoTheory (All records)**	**CompVis (1000)**	**CompVis (All records)**	**Math (1000)**	**Math (all records)**
Accuracy	**81.62%**	**94.77%**	**89.06%**	**95.14%**	**69.86%**	**87.51%**
Precision (Macro)	**40.81%**	**94.71%**	**44.53%**	**95.41%**	**34.93%**	**86.84%**
Recall	**50%**	**87.4%**	**50%**	**78.9%**	**50%**	**82.6%**
F1 score (Macro)	**44.94%**	**90.54%**	**47.10%**	**84.87%**	**41.12%**	**84.34%**

Statistical Model: Logistic Regression
Tokenizer: BERT Tokenizer

Labels ///// Metrics	**InfoTheory (1000 records)**	**InfoTheory (All records)**	**CompVis (1000)**	**CompVis (All records)**	**Math (1000)**	**Math (all records)**
Accuracy	**81.51%**	**81.60%**	**88.86%**	**89.06%**	**69.86%**	**69.86%**
Precision (Macro)	**52.06%**	**40.8%**	**48.69%**	**44.53%**	**34.93%**	**34.93%**
Recall	**50%**	**50%**	**49.96%**	**50%**	**50%**	**50%**
F1 score (Macro)	**45.15%**	**44.94%**	**47.23%**	**47.10%**	**41.12%**	**41.11%**

- F1 score is slightly improved when we use preprocessing which includes BERT Tokenizer compared to other preprocessing when we train on the first 1000 records.
- The model accuracy is pretty good with the preprocessing which includes lemma tokenizer in it when the training is performed on all the records.
- Even though the accuracy is good when trained on 1000 records, we should consider precision because the accuracy is bad in terms of imbalance.
- Recall value is higher in the first table when trained on all the records, which clearly says that the number of correct positive predictions made out of all positive predictions that could have been made.

Part 2: Topic Modelling

These are the one of the topic screenshots with different training data i.e one with 1000 training and other with 20000 training records.

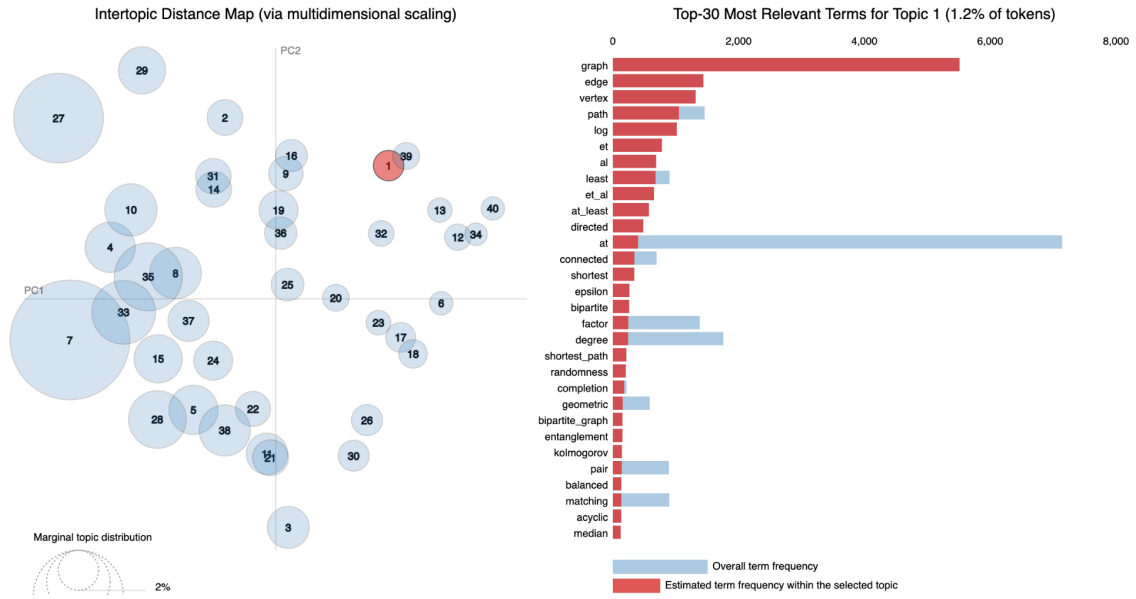

Figure 1: K=40 and 20000 Training data

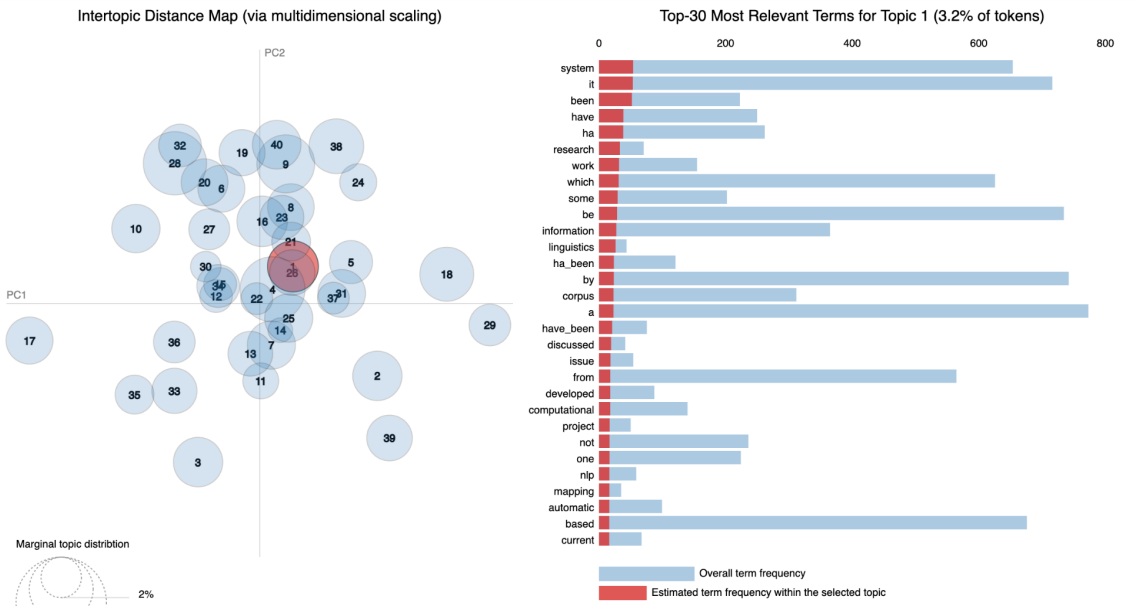

Figure 2: K=40 and 1000 Training Data

The majority topics are related to system information, configurations, wireless communication, grammar, semantics/syntactic and some technical terms which are included in the wireless technology i.e signals, networks, graphs, channel related information.

All the tops in the topic are very close enough when considering twenty thousand records. When there is smaller data from which we are considering 40 topics, in this case the top words are not close enough. We can see from Figure 2, [System, it,have, been] are the top words which don't have any connection at all.

Example: Part of the Article

"The technique developed is a variant of dependency-directed backtracking that uses only polynomial space while still providing useful control information and retaining the completeness guarantees provided by earlier approaches."

The article clearly talks about the graphs and something related to assurance of the information. There is a high probability to see this article in the various topics like it can be seen in graph topics and can also be seen in the information delivery/communication and can also be seen in the information transfer related topic and communication medium as well. As we can see from figure 1, it is purely related to graphs. So we can highly expect to see this article on that topic. In terms of 1000 training data and 20000 training data, it is more likely to see closely related words in a topic when we have more data. In this case, 20K data is more effective in generating meaning topics with context related words in it.