TTDS CW2 Report

B204511

November 2024

1 Introduction

In this report I have performed: Analysis of IR systems and built a module to evaluate the output, Analysed contents of the 3 famous religious texts to understand what the most influential and distinguishing tokens for each text are, employed ML topic modelling technique - LDA - to perform topic modelling and cluster the texts based on their contents, and finally performed text classification to analyse tweet sentiment, for that I have trained a linear SVC from scratch and fine-tuned a transformer-based model to perform the same task and built an evaluation pipeline to collect performance reports and export them into a csv file.

In doing so, I have acquired new skills for performing independent data analysis and solidified my engineering skills to write reusable code of high quality with ample commentary and clear logic.

2 Information Retrieval

After aggregating and computing the mean performance of each system for each metric, the following results were obtained:

Table 1: System results

System	P@10	R@50	r-precision	AP	nDCG@10	nDCG@20
1	0.39	0.834	0.401	0.4	0.363	0.485
2	0.22	**0.867**	0.253	0.3	0.2	0.246
3	**0.41**	0.767	**0.449**	**0.451**	**0.42**	**0.511**
4	0.08	0.189	0.049	0.075	0.069	0.076
5	**0.41**	0.767	0.358	0.364	0.333	0.424
6	**0.41**	0.767	**0.449**	0.445	0.4	0.49

For each metric, I have highlighted the best performing system (in some cases, several systems had the same best score), to determine if the better performing system(s) are statistically better than the second scoring one, I have conducted 2-tailed t-tests with a critical value of 0.05. I set the null hypothesis and alternative hypothesis as:

$$H_0 : \mu_{\text{best}} = \mu_{\text{second best}}$$

$$H_1 : \mu_{\text{best}} > \mu_{\text{second best}}$$

P@10	R@50	r-precision	AP	nDCG@10	nDCG@20
3 vs 1 (0.751) 5 vs 1 (0.751) 6 vs 1 (0.751)	2 vs 1 (0.343)	3 vs 1 (p-value)	6 vs 1 (0.591)	3 vs 6 (0.272)	3 vs 6 (0.244)

Table 2: t-tests between best and second best systems for each metric

The results were as follows:

The p-value in all cases is > 0.05 and < 0.95, which means that we cannot reject the null hypothesis and there is not enough evidence to argue that any of the best performing systems (including the cases where there were several candidates) is statistically better than the second-best performing ones.

3 Token Analysis

OT		NT		Quran	
MI	χ^2	MI	χ^2	MI	χ^2
jesu 0.038	jesu 1328.566	jesu 0.056	jesu 2893.457	god 0.031	muhammad 1667.179
israel 0.036	lord 1206.064	christ 0.034	christ 1683.011	muhammad 0.03	god 1521.41
king 0.031	israel 1174.071	lord 0.024	lord 853.76	torment 0.021	torment 1209.627
lord 0.031	king 1042.85	israel 0.015	discipl 778.895	believ 0.02	believ 1197.831
ot 0.023	christ 703.673	discipl 0.015	nt 539.668	messeng 0.016	messeng 944.798
christ 0.02	god 695.067	peopl 0.012	peter 507.351	king 0.016	revel 846.744
believ 0.019	believ 682.372	king 0.011	paul 500.172	israel 0.016	quran 814.919
god 0.016	ot 631.652	nt 0.011	thing 457.814	quran 0.015	unbeliev 763.422
son 0.016	son 612.796	ot 0.011	israel 456.939	revel 0.014	guidanc 730.74
muhammad 0.016	muhammad 553.875	peter 0.01	spirit 406.494	unbeliev 0.013	disbeliev 708.902

Table 3: Word rankings for MI and χ^2 for the 3 corpuses.

Mutual information is a measure of how much a certain word tells us about the category of text. χ^2 tells us a level of association between a token and a corpus, essentially highlighting frequently appearing terms.

According to both methods, for all corpuses the top ranking words are religiously related with words like jesus, lord, christ, god, muhammad leading. Since the words are similar, it might be difficult to judge based on top 1/2 tokens, however if we look deeper into the rankings, we can find more specific words that distinguish the corpuses. For the old testament, words like believ and son are telling us more, as are discipl, peter for the new testament and torment, messeng for Quran. Certain words like muhammad only appear for 2 out of 3 corpora. Overall, the words are indicative of the topic - religion - and tell us more detail about each of the texts, where Old Testament and New Testament are focused on Israel and Jesus, whereas Quran is more focused on God and Muhammad.

4 Topic Analysis

OT (Topic 18: 0.095)	NT (Topic 18: 0.1122)	Quran (Topic 16: 0.1291)
god 0.1782	god 0.1782	god 0.1157
lord 0.0942	lord 0.0942	life 0.049
believ 0.0564	believ 0.0564	deed 0.0457
truth 0.037	truth 0.037	lord 0.0425
word 0.0341	word 0.0341	live 0.0357
peopl 0.0332	peopl 0.0332	evil 0.0354
reward 0.0264	reward 0.0264	good 0.033
angel 0.0229	angel 0.0229	love 0.0284
hear 0.0157	hear 0.0157	peopl 0.0283
righteous 0.0139	righteous 0.0139	soul 0.0246

Table 4: Top 10 tokens for each top ranking Topic for each Corpus

To model and identify the most prevailing topics across the texts, I have run the LDA on the entire set of verses from

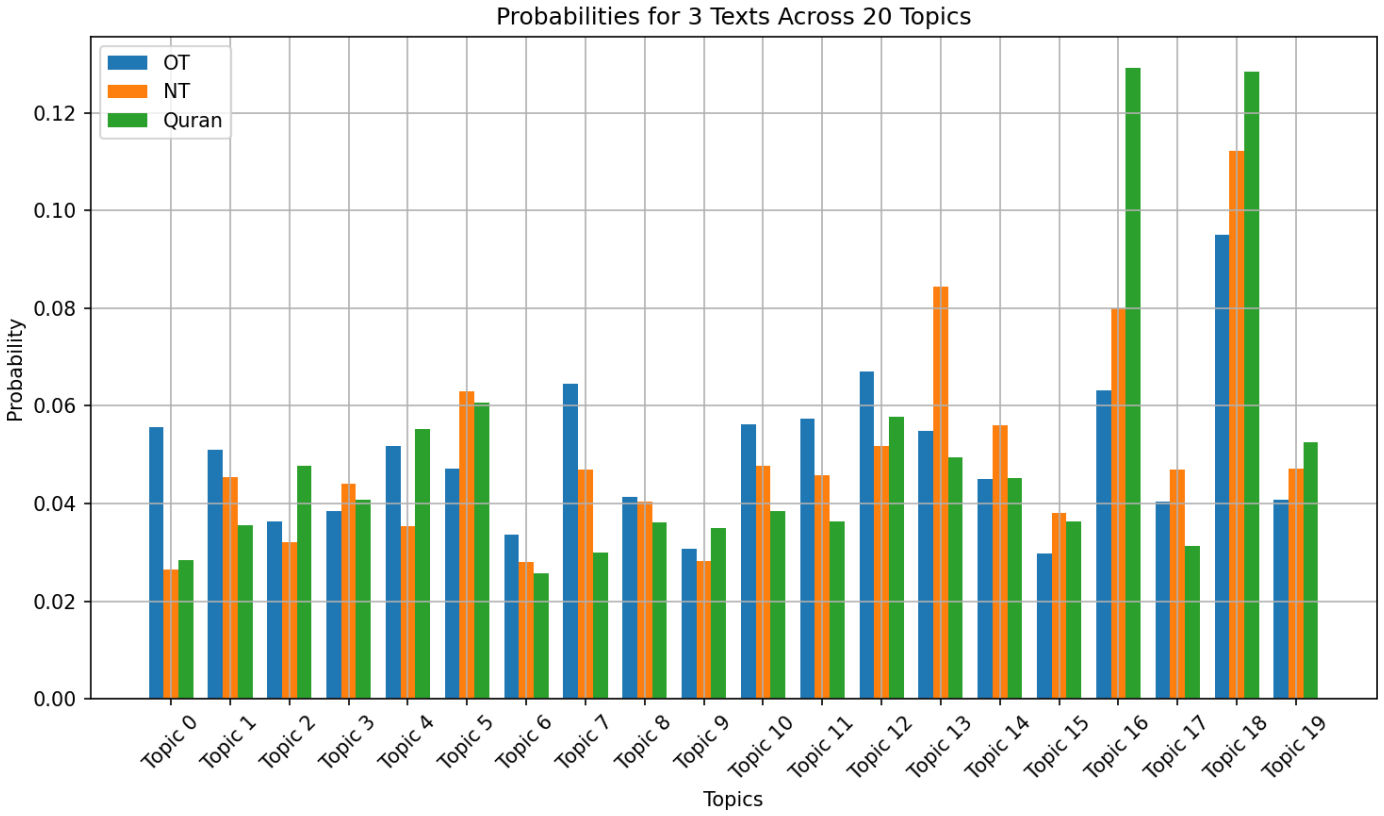

Figure 1: Topic probabilities

all 3 texts. Then, for each of the corpuses I have computed corpus average topic probabilities and found the ones scoring the highest. For each of the best-scoring topics, I have listed the top 10 tokens for that topic and listed them in Table 4.

Looking at the most common words, I would describe Topic 18 as **Faith outside** and Topic 16 as **Faith inside**. That is because 18 has words like God, Lord, People, Reward, Angel, which describe elements of faith outside relative to an individual, whereas 16 has words like Life, Deed, Live, Evil, Good, Love, Soul, which are closer to the faith elements that are inherent to an individual person.

To study the distribution of topics in detail, I have plotted the allocated probabilities on a bar chart for each of the texts and the results can be found in Figure 1.

We can see that even though Quran was classified as Topic 16, it was very close to be Topic 18, which means all texts have a a lot that would describe them as Faith outside, but for Quran there was a bit more of Faith inside. Regardless, 16 and 18 were the highest scoring for all, followed by 13 and 12. The word list is different to that identified by MI and χ^2 methods, which were quite generic and didn't tell us anything specific about the content of the texts, other than they were religious and had something to do with Israel.

5 Classification

To classify tweet text sentiment into Positive, Negative and Neutral, I have trained a SVC with C=1000 and a Linear Kernel. The training dataset was not preprocessed in any way other than tokenization, which lowercases the text and splits on non-alphanumeric characters. This gave me a decent baseline with the following results for the Test dataset:

Class	Precision	Recall	F1-score
Positive	0.55	0.61	0.58
Neutral	0.60	0.60	0.60
Negative	0.54	0.45	0.49
Macro Avg	0.56	0.55	0.55

Table 5: Baseline Test set results

To get a sense of what the results were like, I've looked into 3 mislabeled examples on the Dev set and they were:

- Gold: Positive Pred: Neutral. "rt benwerd trump supporters please read this article do you support this what possible response is there"
- Gold: Negative Pred: Neutral. "tomwarren i don t know she just recited jay-z to me this may have yours beat for best reply"
- Gold: Neutral Pred: Positive. "these gas stations need to have some blackfriday prices for uber drivers lbvs use my link for 200 extra"

In all cases, I thought BoW the approach lacks connection between words in a sequence that encode sentiment the way we interpret it. To improve upon the baseline result, I decided to fine-tune a Transformer-based Encoder of a distilled version of BERT - DistilBert, as it is a lightweight model that is easy to setup and fine-tune. Transformer-based model has attention mechanisms that allow to draw long range dependencies and allow every token in a sequence to attend to every other when. I converted the training set into a compatible format and further trained the model from publicly available weights for 3 epochs on the Training portion of the Train dataset. Using the data, in about 2 hours of training, I was able to achieve the following improved results:

Class	Precision	Recall	F1-score
Positive	0.76	0.68	0.72
Neutral	0.68	0.75	0.71
Negative	0.67	0.63	0.65
Macro Avg	0.7	0.69	0.69

Table 6: Transformer Test set results