Text Technology for Data Science

Coursework 2

s2845408

November 27 2025

Introduction:

This project covers three main components such as Information Retrieval Evaluation, Text Analysis and Text Classification. Across these tasks, various computational techniques were applied to evaluate system outputs and analyse model behaviour. The work included metric-based evaluation, statistical text processing, topic modelling and supervised machine learning to provide practical insight into modern text-analysis and retrieval systems.

Implementation:

1. Information Retrieval Evaluation:

The `ir_eval.csv` file reports results for six retrieval systems across ten queries using precision, recall, r-precision, AP and nDCG. These functions were implemented to calculate P@10, R@50, r-precision, Average Precision (AP) and nDCG@10/20 for each system query pair. Aggregated scores enabled system-level comparison, showing how retrieval effectiveness is measured and how small ranking changes influence metrics such as nDCG.

2. Text Analysis:

Text analysis combined feature-based and topic-based methods to study linguistic patterns in the Old Testament, New Testament and Quran. After tokenisation, stopword removal and stemming, the Mutual Information and Chi-square identified discriminative lexical features. An LDA model with twenty topics was trained to uncover thematic structure and average topic distributions were used to determine representative topics for each corpus.

3. Text Classification:

Sentiment classification followed Lab 7 using a bag-of-words representation of cleaned tweet text. A baseline SVC (C=1000) was compared against Logistic Regression, Random Forest, Multinomial Naive Bayes and LinearSVC. Models were evaluated using macro-F1 to determine the best-performing classifier.

Summary of Learn:

The coursework strengthened understanding of retrieval metrics and their sensitivity to ranking. Feature-selection techniques such as MI and Chi-square revealed distinctive vocabulary across corpora. LDA improved our understanding of topic modeling and thematic structure extraction. The classification task reinforced supervised learning fundamentals including preprocessing, model training and evaluation, demonstrating how algorithms behave under consistent feature settings.

Challenges:

The IR task required careful computation of rank-sensitive metrics such as AP and nDCG. Text analysis presented challenges in preprocessing large corpora, handling vocabulary imbalance and selecting discriminative features. LDA training required tuning to achieve coherent topics. The classification task involved working with sparse features and preventing model overfitting across multiple models.

IR Evaluation:

In this section, the system compares the six IR system using six retrieval metrics such as; P@10, R@50, r-precision, AP, nDCG@10 and nDCG@20, by using the per query values computed in `ir_eval.csv`. For each metric, the best system was found using the mean score across the 10 queries. Then, a paired two tailed t-test (p-value of 0.05) was applied between the systems to determine statistical significance.

Metric	Best System	Second System	Significantly Better
P@10	3 (tie with 5, 6)	1	No (p = 0.7509)
R@50	2	1	No (p = 0.3434)
r-precision	3 (tie with 6)	1	No (p = 0.5911)
AP	3	6	No (p = 0.6757)
nDCG@10	3	6	No (p = 0.2723)
nDCG@20	3	6	No (p = 0.2442)

Table 1: Best IR systems per metric and statistical significance of differences.

Across all six evaluation metrics, system 3 most frequently achieves the highest mean score. In some cases such as P@10 and r-precision, system 3 ties with systems 5 and 6, the next best distinct system is used for comparison. For each metric, a pair of two-tailed t-test was applied using the ten per query scores for the best and second best system. In this case, the resulting p-value is higher than the significance threshold level of 0.05. This indicates that the performance difference between the system are not significant enough in relation to their query-by-query variability. Hence, System 3 reaches the highest average performance, so no system can be declared statistically superior to others.

Token Analysis:

This analysis compares the feature ranking output produced Mutual Information (MI) and chi-square for the three corpora such as New Testament, Old Testament and Quran. The analysis is completely based on the ranking results obtained from the pre processing and scoring pipeline.

Difference in Rankings:

The ranking of Mutual Information (MI) and chi-square because MI rates the importance of words in terms of their distinctive ability in the lexicon by high scores to low frequency words occurring nearly entirely in one corpus, such as jesu, christ and lord in New Testament, israel, jesu and lord in Old Testament and god, muhammad and torment in Quran. These terms are strongly characteristic but not always frequent. In contrast, chi-square rewards high frequency tokens that vary strongly across corpora, upon surfacing dominant terms like jesu, christ, nt and lord in New Testament corpus, lord, ot, king, jesu in Old Testament corpus and god, Muhammad, believ, torment in Quran are very prominent. Overall, the two algorithms are very different because MI algorithm is very concerned with uniqueness in corpora, while the chi-square algorithm is concerned with commonality of words.

Learn about Rankings:

The ranking matrices clearly illustrate linguistic in the corpora. For the Quran, the top terms brought out by MI terms such as messeng, king, forgiv while chi-square mainly selects structural words like ‘muhammad’, ‘believ’, ‘revel’ because the subject resolves extensively around the command of god. Moving on to Old Testament, the terms brought out by MI are geographical terms such as israel, jesu, lord while the structural words like ‘lord’, ‘king’, ‘jesu’, ‘israel’ dominated by chi-square. Finally for New Testament, the terms ‘jesus’, ‘christ’, ‘lord’ dominates the MI because vocabulary revolves faith in jesus. Meanwhile, the term ‘jesu’, ‘christ’, ‘nt’ stand out in chi-square because the language revolves around jesus.

Table 2: Top 10 MI Tokens for New Testament, Old Testament and Quran

Rank	**NT Token**	**MI Score**	**OT Token**	**MI Score**	**Quran Token**	**MI Score**
1	jesu	0.0394	jesu	0.0268	god	0.0264
2	christ	0.0239	israel	0.0254	muhammad	0.0209
3	lord	0.0174	lord	0.0227	torment	0.0143
4	israel	0.0108	king	0.0220	believ	0.0141
5	discipl	0.0106	ot	0.0157	messeng	0.0112
6	peopl	0.0083	christ	0.0143	king	0.0111
7	king	0.0082	god	0.0131	forgiv	0.0110
8	nt	0.0076	believ	0.0129	quran	0.0102
9	ot	0.0076	son	0.0121	revel	0.0101
10	land	0.0073	muhammad	0.0112	unbeliev	0.0091

Table 3: Top 10 Chi-Square Tokens for New Testament, Old Testament, and Quran

Rank	**NT Token**	**Score**	**OT Token**	**Score**	**Quran Token**	**Score**
1	jesu	2919.95	ot	1755.38	god	2543.08
2	christ	1802.99	lord	1453.51	muhammad	1696.21
3	nt	1589.62	king	1387.20	believ	1342.33
4	lord	969.64	jesu	1354.53	torment	1236.95
5	discipl	817.45	israel	1299.07	messeng	1119.37
6	ot	733.93	god	975.11	quran	1038.31
7	thing	596.87	son	856.89	revel	898.34
8	peter	526.32	christ	753.84	guidanc	807.10
9	paul	522.76	believ	746.97	unbeliev	787.20
10	israel	513.96	land	669.73	disbeliev	785.72

Rank	NT Token	Weight	OT Token	Weight	Quran Token	Weight
1	jesu	881.0184	lord	1799.4951	god	3084.6533
2	thing	811.7532	god	880.4480	lord	915.4582
3	god	605.4264	word	807.6088	peopl	523.8709
4	christ	501.3980	fear	604.6053	fear	494.8636
5	spirit	299.6900	command	563.7937	merci	453.4935
6	faith	265.0224	nt	408.1103	messeng	444.2564
7	work	236.4003	heart	400.7994	worship	398.6101
8	man	187.6697	law	394.7200	believ	397.0213
9	discipl	180.6544	israel	388.7234	forgiv	248.3855
10	receiv	171.1061	peopl	372.0000	seek	243.0494

Table 4: Top 10 Dominant LDA Topic Tokens for New Testament, Old Testament and Quran

Topic Analysis:

This analysis examine the results of LDA model built on all the verses from Old Testament, New Testament and Quran. To build the model, it has been configured with k=20 topics for each document in the model where topic probability distribution was generated. To determine the average topic score for each corpus, the topic probability scores for all documents in corpus would be averaged. The topic with highest score was treated as most associated with that corpus.

Topic Labels:

- New Testament: Teaching and Dialogue (jesu, thing, god)
- Old Testament: Law and Covenant (lord, god, word)
- Quran: Judgement and Belief (god, lord, peopl)

LDA Model:

The results of LDA model points out the prominent themes in corpora through the topic having largest average probavility in each corpus. The New Testament corpus is mostly composed of teaching given in the form of conversation. Hence, the theme of New Testament corpus is centred on direct speech marked by words as jesu, thing, god. The theme of Old Testament corpus is related to commandments given in terms of land. This is marked by words as lord, god, word. Finally, the theme of Quran is mostly related to god's judgement on individual deeds in terms of belief. Then this is theme is marked by words as god, lord, peopl.

Common in 2 but not other:

The LDA results indicate beyond their individual dominant theme on certain topics are shared between pairs of corpora. The common topics amongst the Old and New Testament include the topics concern the authority of god and community identity that reflects in words such as lord, god, faith. Meanwhile there is an overlap between the New Testament and Quran includes topics that involves in their interpersonal and ethical prespective such as man, peopl and fear. Also the Old Testament and Quran share a instructive themes in words such as god, peopl, fear that point to string legalistic and authoritative tone.

Analysis:

The results from the LDA analysis are different from Mutual Information (MI) and chi-square analysis, which are supervised and discriminative methods that identify the words are strongly distinguish between each corpus. These methods highlight corpus specific words like jesu and christ for New Testament, israel and lord for Old Testament and allah, believ for Quran. On other hand LDA is unsupervised model grouping words into thematic clusters based on their pattern of co-occurrence, allowing it to reveal themes that apper across multiple corpora rather than just one. This approach identifies shared pattern, such as moral or social behaivour that crops up in New Testament and Quran or national identity common to both the Old Testament and New Testament. These cross corpus theme were not detected by MI or chi-score analysis, which focus only on vocabulary that sets one text apart from other. The LDA shows the both distinctive and shared conceptual structures across texts.

Classification:

The classification section describes the methodology and experimentation to build and improve the sentiment classification system required for the coursework. The aim is to construct a baseline model by following the structure of lab7, then design an improved version that performs better on both the development and test sets.

Baseline System:

The baseline system was followed from pipeline of Lab7. Here the dataset was shuffled and it was splitted into 90 percent training set and a 10 percent development set. The Tweets dataset was preprocessed using cleaning methods such as lowercasing and punctuation removal, while the stopwords removal, stemming and lemmatization are included to keep the sparse Bag of Words environment practically defined. A vocabulary was built only from the training data by assigning an integer index to each token. Then all the tweets were represented as count vector in a sparse dok matrix. The vocabulary which are out of terms were mapped to a OOV index, so that feature alignment across different splits would be maintained. A baseline classifier was implemented by a Linear Support Vector Machine using (sklearn.svm.svc) with c=1000. A Macro F1 score was generated on the both development and test set such as 0.560 and 0.558, where this indicates the performance are at moderate level. From the calculation, it exposed common error patterns such as misclassifying neutral tweets, difficulty with multi word sentiment expressions such as not, impressed and inability to read sarcastic content within the framework.

System Improvement:

The improvement is analyzed by exploration of different classifier and techniques with the same feature representation of Bag of Words. The following models were explored such as Logistic Regression, Random Forest Classifier, Multinomial Navie Bayes and Linear Support Vector Classifier. All models were trained with the same features of Bag of Words used by baseline system. The aim is to find the classifier which yields the highest Macro F1 score on development set. Some of these approaches did not result in meaningful improvements. In Random Forest model, the effect in structured data performed badly in this context because it does not handle very high dimensional data and sparse matrices. This resulted in unstable decision boundaries. In Multinomial Navie Bayes model, it gave a high recall for positive and negative classes but low precision and low performance on neutral tweets. These results in need to explore other models that suit best sparse and high dimensional text data.

Best Model and Improvement:

The different models are evaluated using the data but the overall performance of Logistic Regression is the best and it is selected as best because it handles high dimensional data and sparse input representations, which characteristics of Bag of Words (i.e) it allows the model to capture sentiment distinctions when words occur frequently. When compared to the baseline Support Vector Machine, Logistic Regression generalized better on unseen data and produce more stable decision boundaries when dealing with large vocabularies. This model obtained a Macro F1 score of 0.6147 on development set and 0.6094 on test set when compared to the baseline model. These consistent gains in both evaluation splits indicate that Logistic Regression generalize well and does not over fit. The improvement comes from its appropriateness for sparse textual data and ability to model class smoothly, hence it handles better in neutral class.

Development vs Test Performance:

The improved system achieved the score of both the data is given below:

- Development Macro F1: 0.6147
- Test Macro F1: 0.6094

The close correspondence between development and test scores suggests that the model is not overfitting to the development and test data. This consistency shows that the improvements are stable across differing evaluation condition. Hence, the performance is stable across splits indicates that the enhancement generalise reliably and reflect improvements in the model ability to classify sentiment.

Conclusion

This project provided a n integrated exploration of Information Retrieval, Text analysis and Text Classification. The IR evaluation showed how ranking metrics like AP and nDCG reflect retrieval effectiveness. In text analysis underlined distinctive vocabulary across three corpora, while LDA showed thematic patterns and cross-corpus connection. In text classification, preprocessing, model choice and subsequent evaluation demonstrated how sparse textual data can be effectively handled. Among the model tested, Logistic Regression yield the improvement and well suitable high dimensional features. Overall, this project enhanced both practical skills and conceptual understanding of NLP and IR techniques.

References

1. Sentiment Analysis and Opinion Mining, Bing Liu - Morgan & Claypool Publishers, May 2012.

2. NLP for Sentiment Analysis, Prof. Dr. Dileep Kumar M, S.R. Jena -Xoffencer International Book Publication, 2024.

3. Topic Modeling in Sentiment Analysis: A Systematic Design, Toqir A. Rana, Yu-N cheah and Sukumar Letchmunan -Journal of ICT Research and Application, Vol-10, pp. 76-93, 2016.