Evaluation of Information Retrieval Models and Query Performance Predictors for Amharic *Adhoc* Task

Tilahun Yeshambel[1], Josiane Mothe[2] and Yaregal Assabie[3]
[1]IT Doctorial Program, Addis Ababa University, Addis Ababa, Ethiopia
[2]INSPE, Univ. de Toulouse, IRIT, UMR5505 CNRS, Toulouse, France
[3]Department of Computer Science, Addis Ababa University, Addis Ababa, Ethiopia

Keywords: Query Performance Prediction, Feature Analysis, Amharic, *Adhoc* Search, IR Models, Correlation.

Abstract: Query performance prediction (QPP) is the task of evaluating the quality of retrieval results of a query within the context of a retrieval model. Although several research activities have been carried out on QPP for many languages, query performance predictors are not studied yet for Amharic *adhoc* information retrieval (IR) task. In this paper, we present the effect of various IR models on Amharic queries, and make some analysis on the computed features for QPP methods from both Indri and Terrier indexes based on the Amharic *Adhoc* Information Retrieval Test Collection (2AIRTC). We conducted various experiments to attest the quality of Amharic queries and performance of IR models on 2AIRTC which is TREC-like test collection. The correlation degree between predictors is used to measure the dependence between various query performance predictors, or between a predictor and a retrieval score. Our finding shows that Jelinek-Mercer model outperformed the BM25 and Dirichlet models. The finding also indicates the correlation matrices between the query-IDF predictors and the evaluation measures show very low Pearson correlation coefficient values.

1 INTRODUCTION

QPP is an important task in IR and is the concern of many researchers in the IR community. The retrieval performances vary widely across different retrieval systems for a single query. The retrieval performance of an IR system is even inconsistent over different IR test collections. For example, *adhoc* retrieval effectiveness can vary across queries for different retrieval methods. *Adhoc* retrieval is the standard information retrieval task that aims to provide relevant documents for a given query within a given document corpus in ranked order based on relevance to a query (Teufel, 2007). As a result of ambiguities in user query or errors in a query, the performances of many IR systems vary while responding to different users' queries (Shtok *et al.*, 2012). A user query performance prediction deals with the retrieval performance of an IR system for a given query. The well-known QPP approaches are pre-retrieval and post-retrieval strategies (Carmel and Yom-Tov, 2010). The post-retrieval method evaluates the difficulty of a query by analyzing the most ranked retrieval results in a response to the query whereas the pre-retrieval approach analyzes a query and corpus prior to retrieval time and is based on linguistic and statistical features of a query and documents (Shtok *et al.*, 2012). Linguistic information (Mothe and Tanguy, 2005) and statistical features such as term frequency-inverse document frequency (TF-IDF), IDF and standard deviation values (Cronen-Townsend *et al.*, 2002) are examples of pre-retrieval prediction methods whereas analysis clarity (Cronen-Townsend *et al.*, 2002; Datta *et al.*, 2022), robustness (Zhou and Croft, 2006), normalized query commitment and weighted information gain (Datta *et al.*, 2022), and retrieval scores (Tao and Wu, 2014) are well-known post-retrieval prediction techniques. Unlike pre-retrieval, post-retrieval predictors are computationally expensive but outperform pre-retrieval predictors. Given a query and an IR system, a QPP method computes a real-valued score that indicates the effectiveness of a system for a given query. The quality of a QPP method is usually determined by evaluating the correlation between its predicted effectiveness scores and the values of some standard evaluation metric for a set of queries

Yeshambel, T., Mothe, J. and Assabie, Y.
Evaluation of Information Retrieval Models and Query Performance Predictors for Amharic Adhoc Task.
DOI: 10.5220/0012224000003598
In *Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2023) - Volume 1: KDIR*, pages 98-108
ISBN: 978-989-758-671-2; ISSN: 2184-3228

(Ganguly *et al*., 2022). The performance of a predictor is usually measured by computing the correlation coefficient between the output of a predictor and the actual performance of queries on a retrieval system. Pearson and Spearman correlation coefficients are the two widely used statistical measures.

Effective QPP enables a retrieval system to decide an appropriate action to be taken at the next turn. Estimating the performance of queries accurately is useful for various applications such as for managing resources, optimizing query, managing user experience, detecting queries with no relevant content and automatic spell correction, performing selective query expansion, selecting effective ranking algorithm for a query, estimating the retrieval quality for the top-ranked results and merging results in a distributed IR system (Akdere *et al*., 2012). For example, if the retrieval effectiveness of a query is good, then the query's performance can be further improved by an affirmative action such as automatic query expansion; otherwise, the system may ask the user for a refinement of the query.

QPP techniques have been explored and analyzed on standard test collections for many information retrieval systems (Pérez-Iglesias and Araujo, 2010; Zendel *et al*., 2023). However, in spite of recent progresses made on the development of Amharic IR (Yeshambel *et al*., 2022; Yeshambel *et al*., 2023), there are no studies carried out to indicate the retrieval effectiveness of various IR techniques for the language. Therefore, the aim of this study is to evaluate:

(i) the quality of some retrieval models on Amharic topic set;

(ii) the correlation between various QPP features; and

(iii) the performance of some QPPs for estimating the quality of retrieval results.

We take the first initiation to examine the impact of some QPP methods on the retrieval effectiveness of Amharic *adhoc* retrieval setting. Some prediction methods considered as state-of-the-art have been implemented in this study. We explore the generalizability of the results from QPP methods for Amharic *adhoc* search for estimating the retrieval quality of an IR model based on top-rank lists. We also present analysis about the computed features and performance measures on the 2AIRTC Amharic collection created by Yeshambel *et al*. (2020c). The correlation between various query performance predictors and each predictor with the retrieval score is also investigated using mean average precision (MAP) and normalized discounted cumulative gain (NDCG) metrics. We performed QPP by considering query and document features. We made various experiments with two types of features: pre-retrieval query performance features, and Letor features.

The rest of this paper is organized as follows. Section 2 introduces the morphological features of Amharic and the progresses made on the construction of resources for Amharic IR. Section 3 discusses related works on QPP methods. In Section 4, we present the experimental setup. Experimental results along with detailed analysis of the features are discussed in Section 5. Finally, we make our conclusion in Section 6.

2 AMHARIC LANGUAGE

Amharic is the official language of the government of Ethiopia. It is used as a mother tongue by substantial segment of the population of the country and serves as the *lingua franca* of various communities speaking other languages. It is the second most spoken Semitic language in the world after Arabic (Gambäck, 2012). The language uses its own script for writing. The Amharic alphabet has 34 characters, each of which has 7 orders representing a combination of a consonant and 7 vowels (Argaw and Asker, 2006).

Owing to its Semitic characteristics, Amharic is morphologically complex language. Word formation involves affixation, reduplication, and Semitic stem interdigitation (Yimam, 2001). Words can be formed either from stems or roots by adding affixes on stems or by modifying the root itself internally. Stems take different types of prefixes and suffixes to form inflectional and derivational words. One of the unique features of Amharic morphology is root-pattern morphological phenomena. For example, Amharic verbs heavily rely on the arrangement of consonants and vowels in order to code different morph-syntactic properties. The root of an Amharic verb is represented by consonants which carry the semantic core of the verb. It is possible to create many words by attaching different affixes to a stem. For example, according to Abate and Assabie (2014), it is possible to generate thousands of words from a verbal root through a complex morphological process. The possibility of having multiple affixations on Amharic words may lead to ambiguity. Thus, contextual analysis is important in Amharic in order to understand the exact meaning of some words. The morphological complexity of

Amharic along with its implication in Amharic IR is exposed in detail by Yeshambel *et al*. (2022) and Yeshambel *et al*. (2023).

Although lots of Amharic documents are available in digital form, the language is still regarded as under-resourced as only few are usable. Among the Amharic resources worthy of mention are NLP corpus (Woldeyohannis and Meshesha, 2020; Yeshambel *et al*., 2021), Amharic IR test collection (Yeshambel *et al*., 2020c), pre-trained language models (Eshetu *et al*., 2020; Muhie *et al*., 2021; Yeshambel *et al*., 2023), stopwords (Alemayehu and Willett, 2002; Yeshambel *et al*., 2020b), and word analyzers (Alemayehu and Willett, 2002; Gasser, 2011; Abate and Assabie, 2014). Evaluations made on the aforementioned resources indicate that most of them are not in usable form for Amharic IR task (Yeshambel *et al*., 2020a). This shows that the status of research and development on Amharic IR is far behind in comparison to technologically advanced languages.

3 RELATED WORK

Zhao *et al*. (2008) proposed pre-retrieval predictors based on two sources of information: the similarity between a query and the underlying collection; and the variability with which query terms occur in documents. These predictors used collection and term distribution statistics. The proposed predictors were evaluated by investigating the correlation between the predictors and the actual performance of a retrieval system on each query. The finding indicates that the proposed predictors provide more consistent performance than existing pre-retrieval methods across informational and navigational search tasks on the Newswire and Web datasets.

Cronen-Townsend *et al*. (2002) developed QPP method by calculating the relative entropy between a query language model and the corresponding collection language model. Various experiments were conducted to evaluate the correlation between clarity score and average precision score for various TREC *Adhoc* Track test collections and the title section of TREC topics using the language modelling. The obtained clarity score was used to measure the coherence of language usage in documents and its models to generate a query. The finding indicates that clarity score can measure the ambiguity of a query regarding to a corpus and they correlate positively with average precision in a variety of TREC datasets.

He and Ounis (2006) explored the effect of query length, IDF, query scope, query clarity, standard deviation of the IDF of the terms in a query, average inverse collection TF (AvICTF) and the ratio of maximum IDF value (IDF_max) to minimum IDF value (IDF_min) of a query. The correlations of the proposed predictors with query performance are evaluated on various TREC collections using title, description and narrative fields of the topic. Among the six proposed predictors, a simplified clarity score (SCS) and the AvICTF have the strongest correlation with average precision for short queries. The standard deviation of IDF, SCS and AvICTF are the most correlated predictors with average precision for normal and long queries. On the contrary, the use of two statistically diverse document weighting models does not have an impact on the overall effectiveness of the proposed predictors.

Zendel *et al*. (2023) proposed entropy to estimate the retrieval effectiveness of neural re-ranking models. The idea is to measure the entropy of the retrieval scores for a re-ranking model if there is no training data or corpus related statistics. The proposed QPP method was tested using Query Likelihood and NeuralRanker on Robust04 *adhoc* retrieval TREC collection. The NeuralRanker slightly outperformed the query likelihood model using the entropy predictor. For query likelihood, the performance of entropy is similar to the standard deviation. However, entropy outperformed in the case of NeuralRanker retrieval approach.

Pérez-Iglesias and Araujo (2010) proposed QPP evaluation framework to avoid some limitations of correlation coefficients to evaluate QPP methods. The two proposed evaluation frameworks are: (i) to evaluate the performance of the method to detect 'easy' or 'hard' queries; and (ii) to make explicit the accuracy of the method for different types of topics. The proposed evaluation frameworks were tested against set of different prediction methods on a standard TREC collection using the query likelihood language modelling with a Dirichlet prior smoothing parameter. The proposed predictor is used to classify queries as hard, average and easy. The most accurate prediction methods in terms of grouping predictions are SCS and ScoreDesv. The Pearson and Kendall correlation coefficients suggested an equivalent performance for SCS and AvICTF methods.

While we see that QPP has been a subject of research and development for various IR systems, to our best knowledge, there is no such attempt for Amharic *adhoc* IR task. Thus, our study is aimed at addressing this issue using the 2AIRTC test collection.

4 EXPERIMENTAL SETUP

In this study, we conducted various experiments to investigate the effect of QPP on Amharic IR, the performance of some widely used QPP methods and the correlation between them. The details of the experimental setup are presented as follows.

4.1 The Amharic IR System

We made our experiment based on the Amharic IR system developed by Yeshambel *et al*. (2020d). In the system, both documents and queries pass through similar pre-processing, morphological analysis and stopword removal tasks. The pre-processing task involves tokenization, character normalization and removal of punctuation marks. Morphological analysis was used to extract stems and roots from surface forms of words. Stopwords were removed from queries and documents using the morpheme-based stopword list constructed by Yeshambel *et al*. (2020b). Document indexing was performed based on morphologically analyzed words. We created three indexes using the retrieval models BM25, language modelling with Dirichlet smoothing (LMDir) and language modelling with Jelinek-Mercer smoothing (LMJM).

The system performed matching between index terms and morphologically analyzed query terms. Finally, ranking was made based on the results obtained from the matching task. Figure 1 shows the architecture of the Amharic IR system.

4.2 Test Collection

The 2AIRTC (Yeshambel *et al*., 2020c) is the only publicly accessible and usable Amharic IR test collection. We used 6,069 documents and 240 queries from this test collection. The relevance has binary values with 1 for relevant and 0 for irrelevant. This collection has been used in some Amharic IR studies. We used the titles of the topic set.

4.3 Features and Predictors

Various query-document, query-IDF and document features have been used in many IR studies for QPP. In this study, we have tested some pre-retrieval and post-retrieval predictors which have performed well in past evaluations. We used IDF with TF since it has been used to rank retrieved documents based on relevance to a given query and enable the retrieval system to select relevant documents to the query based on TF-IDF score. Moreover, IDF statistics have been used in many studies. Ranking is one of the main tasks of *adhoc* retrieval and can be done in various ways. In our case, we did it using Letor features which are associated with query-document pairs. The details of query-IDF, query-document Letor features and their variants are presented in Table 1.

Table 1: Indri query-IDF and Terrier query-document Letor features.

Feature	Variant
Query-IDF	IDF_Q1, IDF_Q3, IDF_mean, IDF_median, IDF_std, IDF_std2, IDF_sum, IDF_min and IDF_max
Query-document and document Letor	IFB2, In_expB2, In_expC2, InB2, InL2, Js_KLs, LemurTF_IDF, LGD, MDL2, ML2, PL2, TF and TF-IDF

4.4 Retrieval Models

The analysis and performance measures were performed using both the Indri and Terrier retrieval systems. Indri uses BM25 and language model methods. Terrier is an efficient, flexible, extensible, effective and interactive IR platform. We used Terrier platform to compute Letor features.

IR models are one of the main components of QPP. We employed three retrieval models: Okapi BM25, language modelling with Dirichlet smoothing and language modelling with Jelinek-Mercer smoothing. The values of k1 and b in BM25 were set to 0.7 and 0.3, respectively. We used λ=0.6 for LMJM, and the value of the smoothing parameter μ for LMDir was set to 1000.

4.5 Evaluation Metrics

We measured the ground truth effectiveness of the system by the trec-eval toolkit to compute precision, MAP, recall, NDCG, relative precision, set-based measures, success and number-based metrics. The prediction quality of QPP features was measured by Pearson correlation between the ground truth effectiveness values and the prediction values for the queries. Pearson measures the strength of the linear relationship between two variables. The number of top-retrieved documents in all the pre-retrieval predictors was selected from the top {5, 10, 15, 20, 30, 100, 200, 500, 1000} retrieval results.

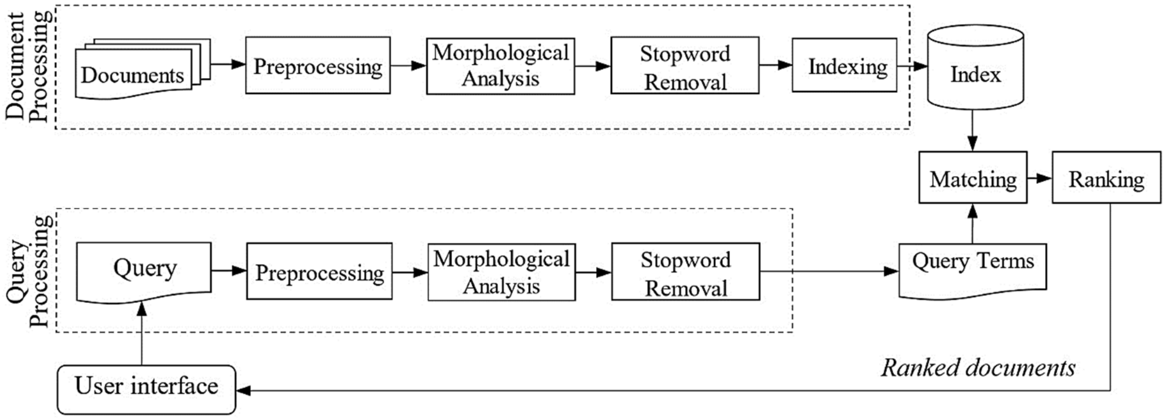

Figure 1: Architecture of Amharic IR system (Yeshambel *et al*., 2020d).

Table 2: Retrieval effectiveness of BM25, Dirichlet and Jelinek-Mercer models.

Model	**Retrieval Effectiveness**									
	P@5	**P@10**	**P@15**	**MAP**	**Rprec**	**bpref**	**recip_rank**	**NDCG**	**RNDCG**	**11pt_avg**
BM25	0.60	0.54	0.49	0.48	0.50	0.48	0.70	0.64	0.59	0.50
Dirichlet	0.65	0.59	0.53	0.56	0.54	0.54	0.78	0.72	0.65	0.58
Jelinek-Mercer	**0.70**	**0.61**	**0.55**	**0.60**	**0.58**	**0.58**	**0.83**	**0.75**	**0.69**	**0.61**

5 RESULT AND DISCUSSION

In this section, the retrieval performance of various IR models, analysis on the correlation between different query performance predictors and comparisons between some pre-retrieval predictors on Amharic IR test collections are presented.

5.1 Effectiveness of IR Models

The retrieval performances of BM25, Dirichlet and Jelinek-Mercer retrieval models are measured by issuing queries against the 2AIRTC. The overall retrieval performances of the models are presented in Table 2 where the best performances are depicted in bold font. As shown in Table 2, the three retrieval models are generally far to each other in terms of retrieval results. It can be seen that Jelinek-Mercer model outperformed the other two models. The correlation between the three retrieval models is computed and presented in Figure 2. The minimum and maximum extreme features of the three models are presented in Figure 3. It can be seen from the figures that the Jelinek-Mercer and Dirichlet features have similar data distribution, with a negative skew and few outliers having small values. They are also strongly correlated with Pearson correlation coefficient (PCC) of 0.95. The least correlation is obtained between BM25 and Dirichlet models. Concerning retrieval effectiveness, all the three models have positive correlations on the Amharic IR test collection.

5.2 Correlation Between Indri Query-IDF Features

To examine the relation among IDF features, the correlation between some query-IDF features was computed from the Indri index. Figure 4 shows PCC between IDF feature variants. Since the query-IDF features were computed on queries, the data count of query-IDF features is smaller than the query-document features that are presented above.

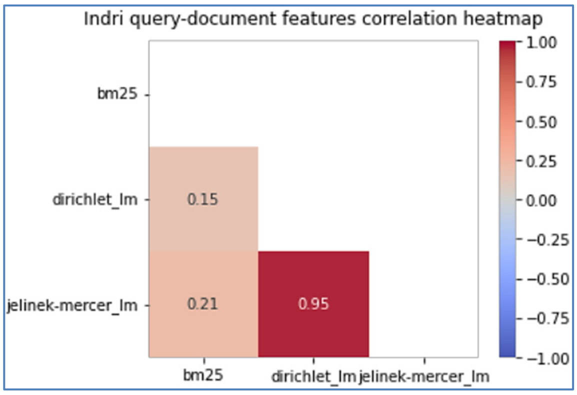

Figure 2: Correlation of the three retrieval models.

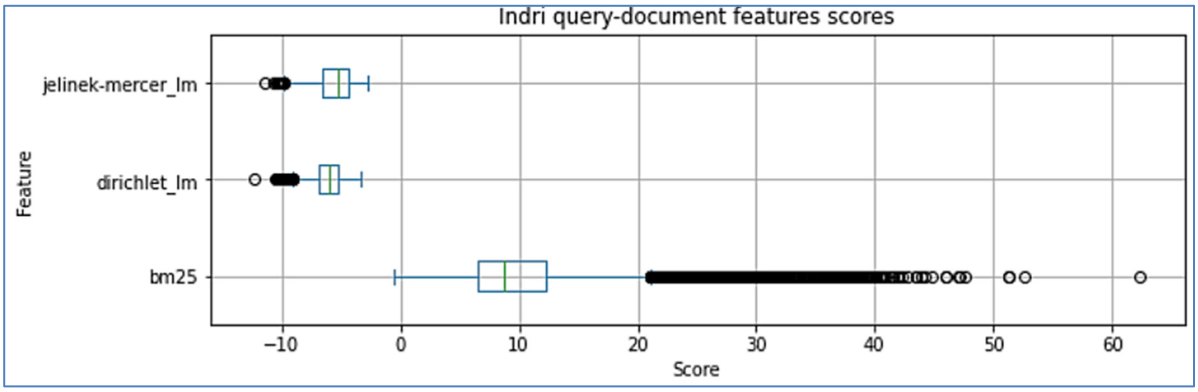

Figure 3: Boxplots of min and max extremes for the three retrieval models.

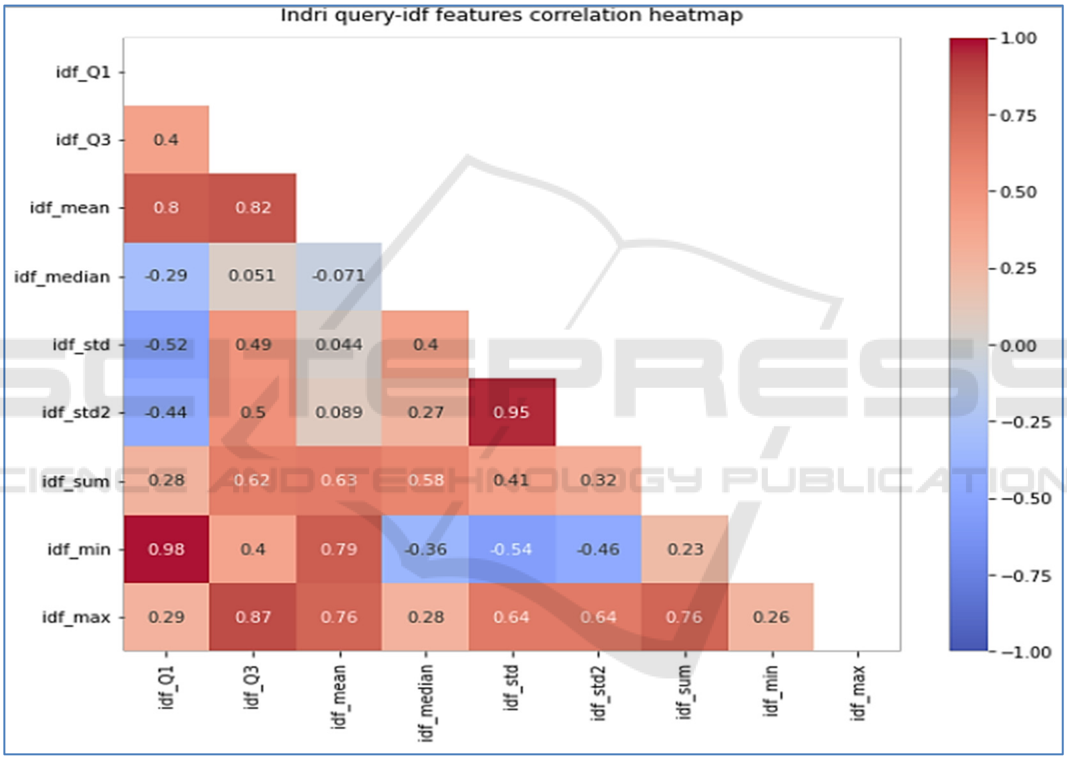

Figure 4: Correlation between query-IDF QPP predictors.

The data counts of query-document and query-IDF features are 2,334 and 240, respectively. The IDF_Q1 has negative correlations with IDF_std2, IDF_std and IDF_median predictors. Similarly the predictor IDF_min has a negative skew with IDF_median, IDF_std and IDF_std2. However, the correlation between any of the predictors is positive. The IDF_sum has the largest data distribution and the next most distributed Indri query-IDF feature variant is variance (IDF-std2). Most of the Indri query-IDF features have outliers having large values. There are also some weak inverse correlations between IDF_Q1 and IDF_median, IDF_std and IDF_std2, IDF_min and IDF_median, IDF_std and IDF_std2. Some of the strongly correlated features are:

- IDF_mean and IDF_Q3 with PCC of 0.82.
- IDF_std and IDF_std2 with PCC of 0.95.
- IDF_min and IDF_Q1 with PCC of 0.98.
- IDF_max and IDF_Q3 with PCC of 0.87.

For better understanding of the features, we plotted score extremes of Indri query-IDF features. We observed that IDF-median, IDF-std and IDF-std2 have positive values.

5.3 Correlation Between Letor Features

For Letor features, a subset of documents was selected since all the documents were not needed to create the training set. Then, feature extraction was performed to represent each query-document pair with the features. To investigate the correlation between Letor predictors, we computed query-document and document Letor features from the Terrier index. In this case, some features are poorly correlated while the others are strongly correlated. The correlations between weak features are shown in Figure 5 where PCCs range from 0.63 to 0.87. The Letor predictors Ml2, Mdl2, Inb2, parameter-free hypergeometric model with Popper's normalization (Dph) and divergence from independence model based on standardization (Dfiz) have relatively strong correlation with similar data distribution as shown in Figure 6. They are strongly correlated with PCCs ranging from 0.90 to 1.00, and we see that the data distributions are nearly the same for each feature which are positively skewed. As shown in Figures 6 and 7, the data distributions are nearly the same for each feature that is positively skewed, and they have more of large outlier values (with some small value outliers for the Hiemstra feature). They are strongly correlated with PCCs ranging from 0.96 to 1.00.

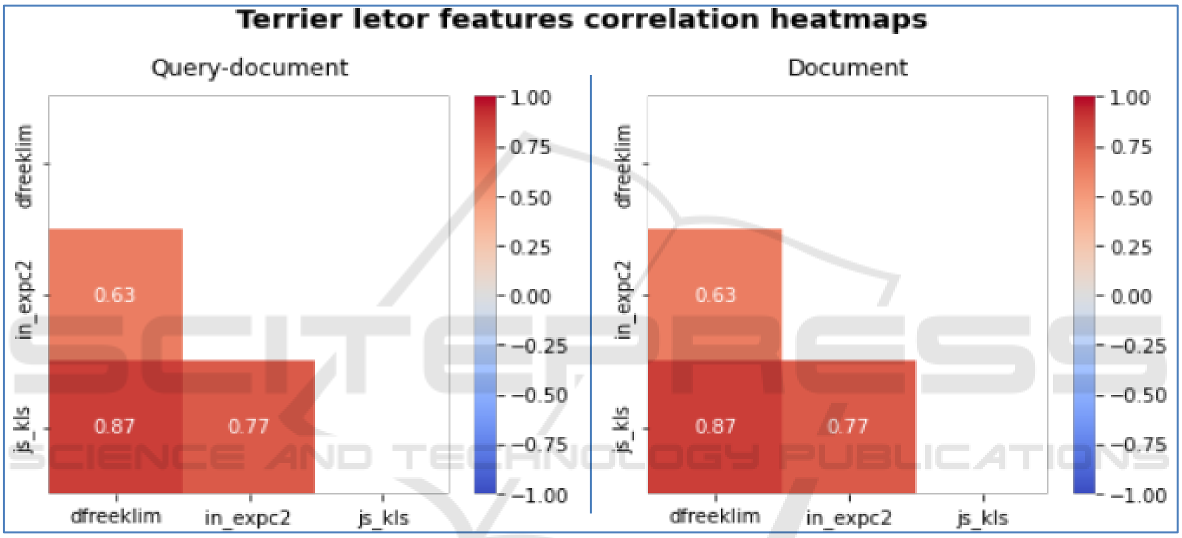

Figure 5: Letor correlated features with different data distribution.

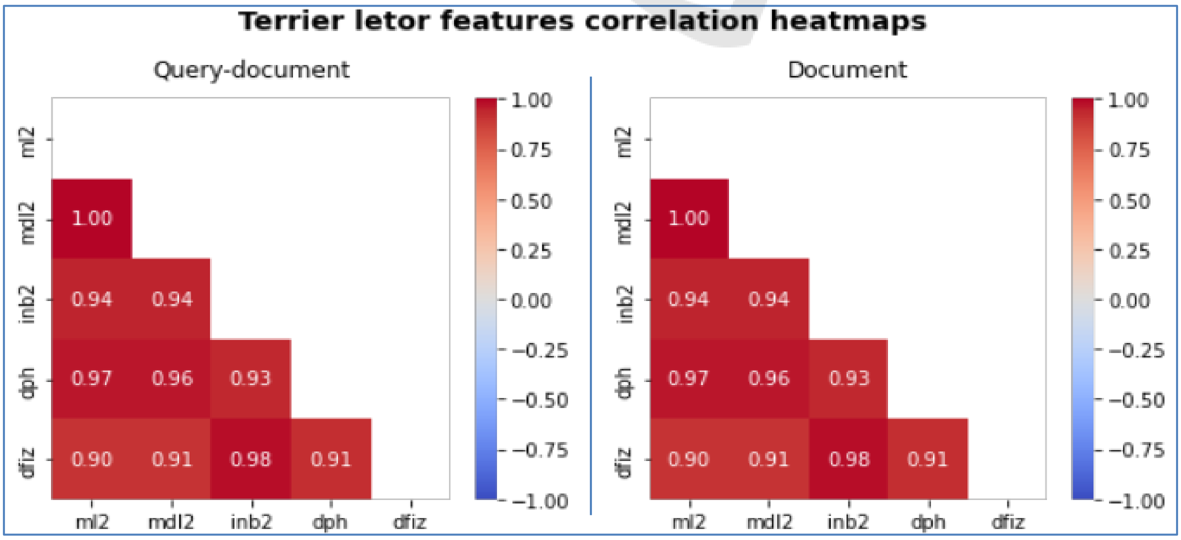

Figure 6: Strongly correlated Letor features with similar data distribution.

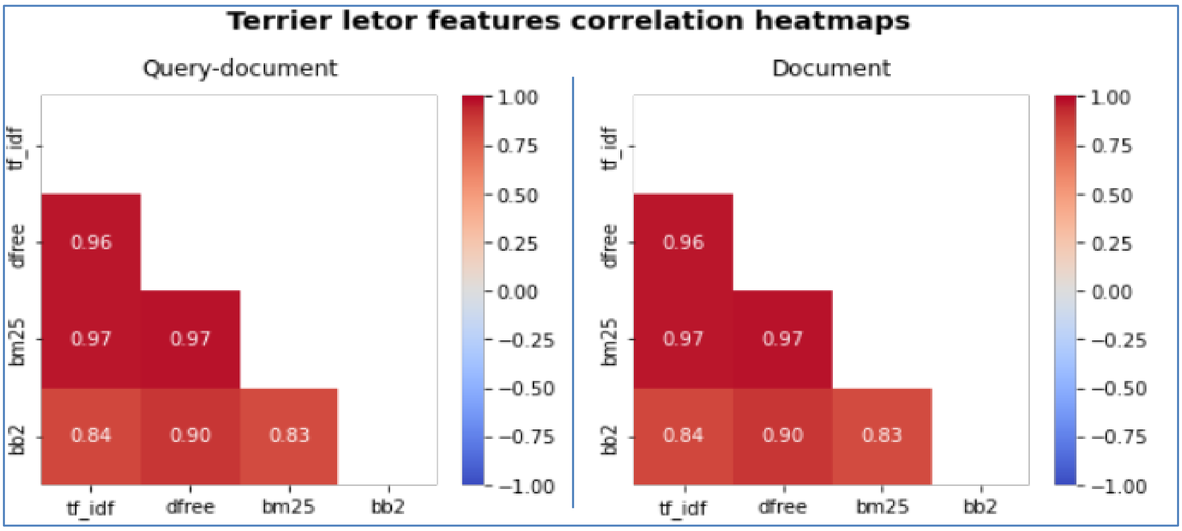

Figure 7: Correlation between Ml2, Mdl2, Inb2, Dph and Dfiz features.

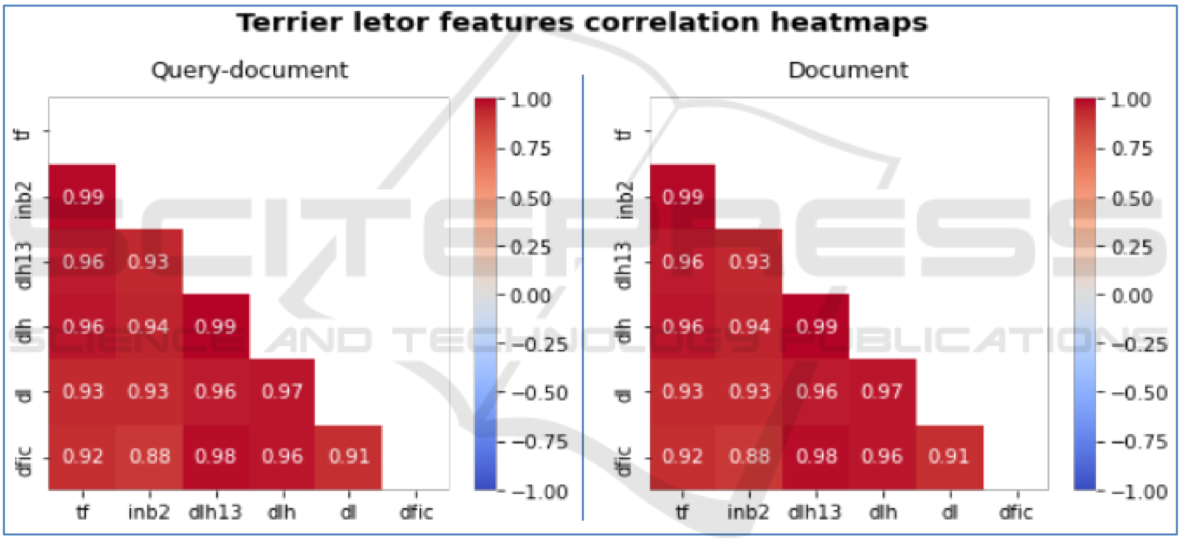

Figure 8: Correlation of TF, inb2, dhl3, dl and dfic features.

The correlations between TF, inb2, dhl3, dl and divergence from independence model based on Chi-square statistics (dfic) features are also computed and shown in Figure 8 where the data distributions are nearly the same for each feature that is positively skewed, and they have more of large outlier values with a few outliers having small values for the document length (DL), parameter-free weighting model (Dlh) and Dlh13 features. They are strongly correlated with PCCs ranging from 0.88 to 0.99.

Moreover, we examined the correlations between TF, Inb2, Dlh13, Dlh, Dl, Dfic, LemurTF-IDF and BM25 divergence from randomness (Dfr) features. Experimental results show that data distributions are nearly the same for each feature. TF, Inb2 and Dfic have more of large outlier values whereas the Dl, Dlh and Dlh13 features have few outliers having small values. All of these features are strongly correlated with PCCs ranging from 0.88 to 0.99. Using box plots of Terrier query-document and document features, we observed that the query-document and document features have the same trend when considering their different variants.

5.4 Performance of Selected Predictors

The correlations between some selected QPP methods and the retrieval scores were computed and analysed. We compared the prediction quality of the four pre-predictors, namely query length (QL), TF,

Table 3: Correlation of query performance predictors and retrieval scores.

Predictor	MAP	NDCG
IDFstd	0.039501	0.03931
maxIDF	0.030262	0.028146
MeanIDF	0.250945	0.249829
TF	0.130696	0.164911
TF-IDF	0.339494	0.370119
QL	-0.15229	-0.1304

IDF and TF-IDF. The PCC of each query prediction feature score against MAP and NDCG for BM25 retrieval model are presented in Table 3. The table indicates the correlation of IDF of query terms and some other query performance prediction features with average precision and NDCG in 2AIRTC. A correlation score of 1 indicates a perfect agreement between a predictor score and the actual retrieval score in the rankings whereas -1 indicates opposite agreement between a predicator and a retrieval score (i.e. MAP and NDCG). All predictors except QL have weak positive correlation with MAP and NDCG of the queries. The correlation between QL and the retrieval scores (i.e. MAP and NDCG) has opposite association. The average query length is 4. The top three well performing predictors are TF-IDF, TF and meanIDF.

5.5 Analyzing Retrieval Scores

The performance measures of various retrieval scoring metrics were computed and analyzed as shown in Figures 9 and 10. In Figure 9, we can see strong correlations between Set_F, Set_P and set_MAP measures with PCCs ranging from 0.95 to 0.97 for Terrier, and from 0.91 to 0.97 for both Indri Dirichlet and Indri Jelinek-Mercer. There is also correlation between the Set_relative_P and Set_recall measures with PCC of 0.99 for Terrier and 0.97 for both Indri Dirichlet and Jelinek-Mercer. In Figure 10, we see that the success at 5 and 10 recommendations are correlated with PCC of 0.85 for both Terrier and Indri Jelinek-Mercer, and 0.76 for Indri Dirichlet.

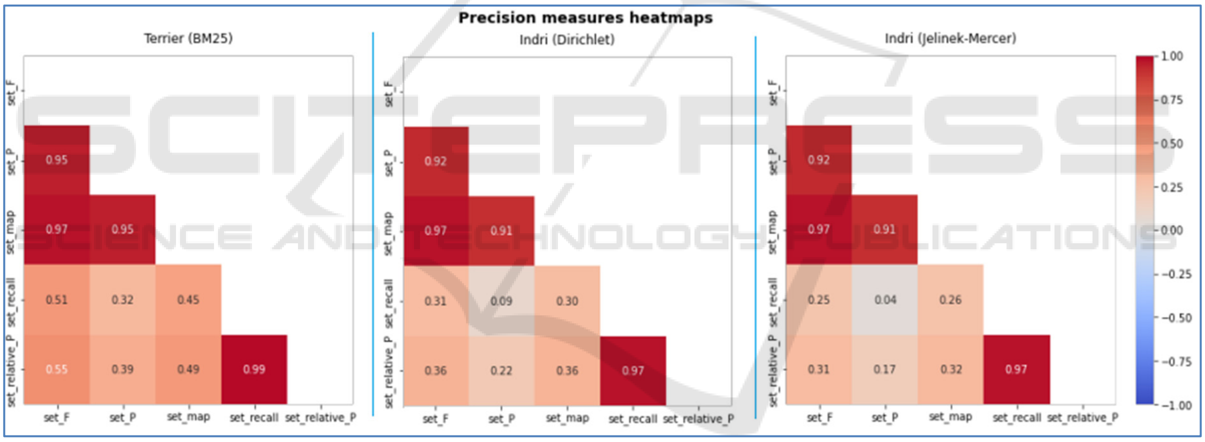

Figure 9: Set-based measures.

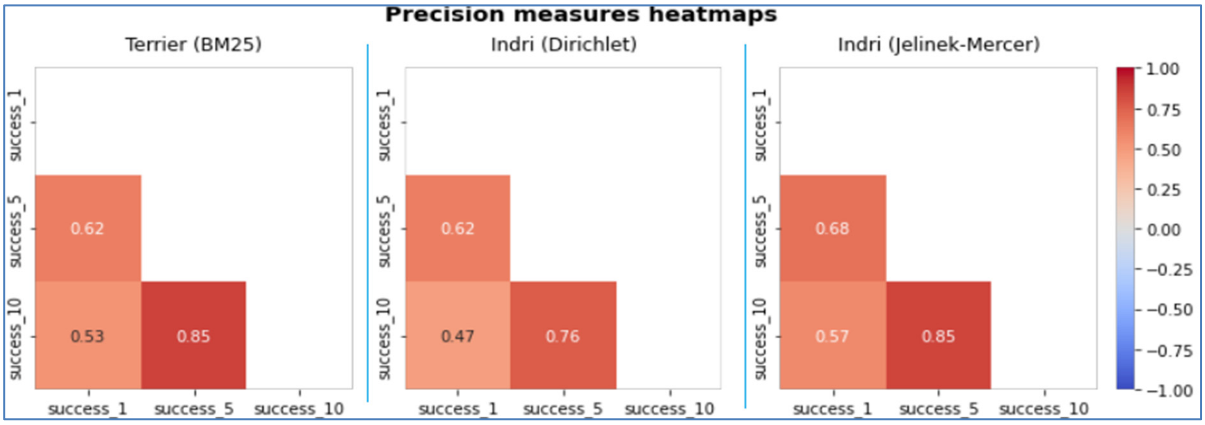

Figure 10: Success measures.

From the experiments carried out to evaluate various IR models, we see that Jelinek-Mercer model outperformed the other models on the 2AIRTC corpus. Regarding the extracted features from the Indri index, there is a high correlation between Jelinek-Mercer and Dirichlet query-document features, but generally low correlations between the different query-IDF features, with few of them strongly pairwise-correlated. Based on our results, it is clear that some queries perform better than others. Most of the query terms are more of specific terms and are able to discriminate retrieval results. Regarding the extracted features from the Terrier index, most of them have high correlations between the different features (except for three of them). An interesting point of discussion is that both query-document and document Letor features have the same correlation matrix as well as the same data distribution for each feature. This is because we nearly obtained the same results for both query-document and document Letor features. Regarding the computed evaluation metrics from both Terrier and Indri (for both Dirichlet and Jelinek-Mercer smoothing) indexes, we can see very similar data distributions and correlations for the precision, NDCG, MAP, recall, relative precision, set-based, success and number-based metrics. The correlation matrices between the query-IDF predictors and the evaluation measures show very low PCC values.

6 CONCLUSION

QPP is the task of estimating the retrieval quality of an IR system for a given query. In this study, we investigated the retrieval effectiveness of various IR models, examined the quality of some query performance predictors in IR task, and explored the correlations between some predictors. We carried out a range of experiments to analyze the effect of QPP methods on 2AIRT based on different IR models and IR metrics. The performances of IR models for queries were tested on 2AIRTC test collection. Our findings reveal Jelinek-Mercer model outperformed over the other models and more correlated to Dirichlet than BM25. Strong correlation is observed between TF, inb2, dhl3, dl and dfic features. Future work needs to focus on developing large test collections for generalizing the quality of a retrieval model on a given query and for investigating the consistence of each query performance predictors across test collections and query types. Further investigation can be made on the quality of more post-retrieval strategies and automatic query expansion on all queries and selective queries using QPP methods.

REFERENCES

Abate, M. and Assabie, Y. (2014). Development of Amharic morphological analyzer using memory based learning. In *Proc. of the 9th Int. Conf. on Natural Language Processing*, Warsaw, pp. 1-13.

Akdere, M., Çetintemel, U., Riondato, M., Upfal, E., Zdonik, S. (2012). Learning-based query performance modelling and prediction. In *Proceedings of 2012 IEEE 28th International Conference on Data Engineering* Arlington, VA, USA, pp. 390-401.

Alemayehu, N. and Willett, P. (2002). Stemming of Amharic words for information retrieval. *Literary and Linguistic Computing*, 17(1), pp.1–17.

Argaw, A.A and Asker, L. (2006). Amharic-English information retrieval. In: Peters, C., *et al.* Evaluation of Multilingual and Multi-modal Information Retrieval. *CLEF 2006. Lecture Notes in Computer Science,* vol 4730. Springer, Berlin, Heidelberg.

Carmel, D. and Yom-Tov, E. (2010). Estimating the query difficulty for information retrieval. Synthesis Lectures Inf. *Concepts Retrieval Serv*, 2(1).

Cronen-Townsend, S., Zhou, Y. and Croft, W. B. (2002). Predicting query performance. In: J¨arvelin, K., Beaulieu, M., Baeza-Yates, R.A., Myaeng, S. (eds.) SIGIR 2002: In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Tampere, Finland, pp. 299-306.

Datta,S., Ganguly,D., Mitra,M. and Greene, D. (2022). A relative information gain-based query performance prediction framework with generated query variants. *ACM Transactions on Information Systems,* 1(1).

Eshetu, A., Teshome, G. and Abebe, T. (2020). Learning word and sub-word vectors for Amharic (Less Resourced Language). *Int. J. Adv. Eng. Res. Sci. (IJAERS)*, 7(8), pp. 358-366.

Gambäck, B. (2012). Tagging and verifying an Amharic news corpus. *Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL8/AfLaT2012),* Istanbul, Turkey,pp. *79-84.*

Ganguly, D., Datta, S., Mitra, M. and Greene, D. (2022). An analysis of variations in the effectiveness of query performance prediction. In *44th European Conference on Information Retrieval (ECIR 2022)*, Stavanger, Norway, pp. 215-229.

Gasser, M., (2011). Hornmorpho: A system for morphological processing of Amharic, Afaan Oromo, and Tigrinya. In: *conference on Human language technology for development*, Alexandria, Egypt, pp. 94-99.

He, B. and Ounis, I. (2006). Query performance prediction. *Information Systems*, 31(7), pp. 585–594.

Mothe, J., Tanguy, L. (2005). Linguistic features to predict query difficulty. In *ACM Conference on research and*

Development in Information Retrieval, SIGIR, Predicting query difficulty-methods and applications workshop, Salvador de Bahia, Brazil, pp.7-10.

Muhie, S., Ayele, A., Venkatesh, G., Gashaw, I. and Biemann, C. (2021). Introducing various semantic models for Amharic: Experimentation and evaluation with multiple tasks and datasets. *Future Internet*, *13(11)*.

Shtok, A., Kurland, O., Carmel, D., Raiber, F., and Markovits, G. (2012). Predicting query performance by query-drift estimation. *ACM Trans. Inf. Syst*. 30(2).

Pérez-Iglesias, J. and Araujo, L. (2010). Evaluation of query performance prediction methods by range. E. Chavez and S. Lonardi (Eds.): *SPIRE 2010, LNCS 6393*, pp. 225–236.

Tao, Y. and Wu, S. (2014). Query performance prediction by considering score magnitude and variance together. In: Li, J., Wang, X.S., Garofalakis, M.N., Soboroff, I., Suel, T., Wang, M. (eds.). In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM 2014, Shanghai, China, pp. 1891-1894.

Teufel, S. (2007). An overview of evaluation methods in TREC adhoc information retrieval and TREC question answering. In title of Evaluation of Text and System speech, Dybkjær et al. (eds.),163-186, University of Cambridge, United Kingdom, Springer.

Woldeyohannis, M. and Meshesha, M. (2020). Usable Amharic text corpus for natural language processing applications. *Applied Corpus Linguistics, 2(3).*

Yeshambel, T., Mothe, J. and Assabie, Y. (2020a). Evaluation of corpora, resources and tools for Amharic information retrieval. In *Proceeding of ICAST2020,* springer, Bahir Dar, Ethiopia,pp.470-483.

Yeshambel, T., Mothe, J. and Assabie, Y. (2020b). Construction of morpheme-based Amharic stopword list for information retrieval system. In *Proceeding of ICAST2020*, Springer, Bahir Dar, Ethiopia, pp. 484-498.

Yeshambel, T., Mothe, J. and Assabie, Y. (2020c). 2AIRTC: The Amharic Adhoc information retrieval test collection. In *proceeding of CLEF 2020*, LNCS 12260, Thessaloniki, Greece, pp.55-66.

Yeshambel, T., Mothe, J. and Assabie, Y. (2020d). Amharic document representation for adhoc retrieval. In *procedding of the 12th internatinal conference on knowledge discovery,* knowledgey enginerring and management, online, pp. 123-134.

Yeshambel, T., Mothe, J. and Assabie, Y. (2021). Morphologically annotated Amharic text corpora. In *Proc. of 44th ACM SIGIR Conf. on research and development in information retrieval*, online conference, Canada, pp. 2349–2355.

Yeshambel, T., Mothe, J., Assabie, Y. (2022) Amharic *Adhoc* information retrieval system based on morphological features. *Applied Sciences*; 12(3).

Yeshambel, T., Mothe, J., Assabie, Y. (2023). Learned text representation for Amharic information retrieval and natural language processing. *Information* ,14(3).

Zendel, O., Liu, B., Culpepper, J. and Scholer, F. (2023). Entropy-based query performance prediction for neural information retrieval systems. *Query Performance Prediction and Its Evaluation in New Tasks, co-located with the 45th European Conference on Information Retrieval (ECIR),* Dublin, Ireland.

Zhao, Y., Scholer, F. and Tsegay, Y. (2008). Effective pre-retrieval query performance prediction using similarity and variability evidence. C. Macdonald et al. (Eds.): *ECIR 2008, LNCS 4956*, pp. 52–64,

Zhou, Y. and Croft, W.B. (2006). Ranking robustness: a novel framework to predict query performance. In: Yu, P.S., Tsotras, V.J., Fox, E.A., Liu, B. (eds.) In *Proceedings of the 2006 ACM CIKM International Conference on Information and Knowledge Management*, Arlington, Virginia, USA, 2006, pp. 567-574.

Yimam, B., (2001). Yamarigna sewasiw *(Amharic grammar)*. CASE. Addis Ababa, 2nd edition.