Chapter 3

Background

This chapter describes different retrieval models used to evaluate the effect of different pre-processing strategies in the IR domain. Moreover, we describe the simulation setup, datasets, evaluation metrics and statistical tests used to assess the effectiveness of different pre-processing strategies.

3.1 Information Retrieval Framework

We use an open-source search engine called Terrier [1] IR platform for indexing and retrieval of the document collection. The main aim of indexing is to structure, organize, and store statistical information about the collection and support efficient search. The user expresses his information needs in terms of a query. The retrieval model matches the query term with the document term in the collection and retrieves a set of documents. For a particular query q, the similarity score of a retrieved document (d) is given by

$$score(d, q) = \sum_{t \epsilon q} score(t \epsilon d) \tag{3.1}$$

Where score $(t \epsilon d)$ represents the weight of a term calculated by a particular retrieval model. We evaluated the effectiveness of different pre-processing strategies on retrieval. Terrier supports various IR models, such as Probabilistic, DFR-based, and Language models. In this work, we used different retrieval models such as Probabilistic retrieval models

[1] http://terrier.org/

(BM25 and TF-IDF), DFR-based retrieval models (In_expB2, In_expC2, DLH, PL2, BB2, InL2, IFB2), and Hiemstra language model in particular. A detailed description of different retrieval models is given below.

3.1.1 BM25 Model

We consider a representative probabilistic model as BM25. BM stands for 'best matching'. For a given query term t, its score in document d is given by Equation 3.2.

$$w(t,d) = tf_d \cdot \frac{\log(\frac{N - tf_d + 0.5}{tf_d + 0.5})}{k_1((1-b) + b\frac{dl}{avdl}) + tf_d} \tag{3.2}$$

3.1.2 TF-IDF Model

In this model, a document's relevance score for a given query is calculated based on *term frequency* and *inverse document frequency*. The *term frequency* indicates the number of times a term is present in a given document, and *inverse document frequency* indicates the number of documents that contain the given term. TF-IDF weighting model within Terrier uses Robertson's tf and Sparck Jones idf [132].

$$w(t,d) = Robertson\_tf \cdot idf \tag{3.3}$$

where,

$$\begin{aligned} Robertson\_tf &= \frac{tf_d}{k_1((1-b)+b\frac{dl}{avdl})+tf_d} \\ idf &= \log(\frac{N}{df+1}) \end{aligned}$$

Notations used by different retrieval models are as follows [48].

dl : document length in number of terms

df : document frequency of term t

N : total number of documents in the collection

tf_{d} : number of times a term t present in the document (d)

k_1 : term-frequency parameter, a constant

b : document length normalization parameter

n_{t} : document frequency of term t

tfn : normalised term frequency

c : free parameter

$avdl$: average document length in the collection

F : frequency of term t in the collection

qtw : is the query term weight given by $\frac{qtf}{qtf_{\max}}$

qtf : is the query term frequency

$qtf_{\max}$: is the maximum query term frequency among the query terms

λ : is the variance and mean of a Poisson distribution.It is given by $\frac{F}{N}$

We also consider DFR-based models like In_expB2, In_expC2, DLH, PL2, BB2, IFB2 and InL2. These models come from the Divergence From Randomness (DFR) family [10]. The DFR models are based on the idea: The more the divergence of the within-document term-frequency from its frequency within the collection, the more information carried by the word (t) in the document (d) [111].

3.1.3 In_expB2 Model

In the Inverse expected document frequency model for randomness, a document's relevance score is given by the ratio of two Bernoulli's processes for first normalisation and normalisation 2 for term frequency normalisation shown in Equation 3.4.

$$w(t,d) = \frac{F+1}{n_t \cdot (tfn+1)}\big(tfn \cdot \log_2 \frac{N+1}{n_e+0.5}\big) \tag{3.4}$$

F : frequency of term(t) in the collection

N : total number of documents in the collection

n_t : document frequency of the term(t)

$n_e = N \cdot (1 - (1 - \frac{n_t}{N})^F)$

tfn : It is normalised term frequency and given by the normalisation 2 :

$$tfn = tf_d \cdot \log_2(1 + c \cdot \frac{avdl}{l}) \tag{3.5}$$

3.1.4 In_expC2 Model

In this model, Equation 3.6 calculates a document's relevance score.

$$w(t,d) = \frac{F+1}{n_t \cdot (tfn_e + 1)}\big(tfn_e \cdot \log_2 \frac{N+1}{n_e + 0.5}\big) \tag{3.6}$$

tfn_e also denotes the normalised term frequency and is given by a modified version of normalisation 2:

$$tfn_e = tf_d \cdot \log_e(1 + c \cdot \frac{avdl}{l}) \tag{3.7}$$

3.1.5 BB2 Model

In the Bose-Einstein model for randomness, the weight of a query term in a given document is given by the ratio of two of Bernoulli's processes (first normalisation and second normalisation) for term frequency normalisation shown in Equation 3.8.

$$\begin{aligned} w(t,d) = \frac{F+1}{n_t \cdot (tfn+1)}\big[&(-\log_2(N-1) - \log_2(e) \\ &+ f(N+F-1, N+F-tfn-2) \\ &- f(F, F-tfn)\big)\big] \end{aligned} \tag{3.8}$$

3.1.6 IFB2 Model

In the IFB2 or inverse term frequency model for randomness, the weight of a query term in a document is given by Equation 3.9.

$$w(t,d) = \frac{F+1}{n_t \cdot (tfn+1)}\big(tfn \cdot \log_2 \frac{N+1}{F+0.5}\big) \tag{3.9}$$

3.1.7 DLH Model

DLH is a parameter-free weighting model. It will not affect the results even if the user specifies a parameter value. Equation 3.10 gives a query term's weight in a document.

$$w(t,d) = \sum_{t \epsilon Q} qtw \cdot ((\frac{1}{tf+0.5})) \cdot (\log_2[(\frac{tf \cdot avdl}{l}) \cdot (\frac{N}{F})]) + 0.5 \cdot log_2(2\pi tf(1-(\frac{tf}{l}))) \tag{3.10}$$

3.1.8 PL2 Model

In this model, Equation 3.11 gives a document's relevance score.

$$w(t,d) = \sum_{t \epsilon Q} qtw \cdot \frac{1}{tfn+1} \cdot (tfn(\log_2(\frac{tfn}{\lambda}) + (\lambda - tfn) \cdot \log_2(e) + 0.5 \cdot \log_2(2\pi \cdot tfn)) \tag{3.11}$$

3.1.9 InL2 Model

In this model, a document's relevance score is given by the ratio of two Laplace processes for first normalisation and normalisation 2 for term frequency normalisation as shown in Equation 3.12.

$$w(t,d) = \frac{1}{tfn+1}\big(tfn \cdot \log_2 \frac{N+1}{n_t+0.5}\big) \tag{3.12}$$

3.1.10 Hiemstra_Language Model

Finally, we explore a non-parametric probabilistic model, the language model proposed by Hiemstra [51]. The probability estimation depends upon the term frequency in the document d_i or the entire corpus. In this model, a smoothing parameter λ uses a default value 0.15. The similarity between a query and a document is represented by generation probability as given in Equation 3.13.

$$P(D, T_1, T_2, T_3, \cdots \cdots T_n) = P(D) \prod_{i=1}^{n} ((1 - \lambda_i) P(T_i) + \lambda_i P(T_i|D)) \tag{3.13}$$

where: D is a document and T_i are query terms.

3.2 Simulation Setup and Test Collections

3.2.1 Simulation Setup

In this thesis, we performed extensive simulations on different pre-processing strategies using the Terrier-version-4.1 IR system on a desktop Ubuntu 18.04 system with a core i3 processor and 8 GB RAM.

3.2.2 Test Collections

We used the Marathi, Bengali, Gujarati, Hindi and English language collections built during FIRE [2] evaluation campaign. We also built a small test collection in Sanskrit and experimented with it. The building of the Sanskrit test collection is presented in Chapter 6. These corpora consist of news articles extracted from different resources. The Marathi articles extracted from *'Maharashtra Times'* and *'Sakal'* (articles spanning the period April 2004 through September 2007), Bengali articles extracted from the *'CRI'* and *'Anandabazar Patrika'* (a newspaper edited by ABP Ltd.), Gujarati articles extracted from archives of the daily newspaper, *'Gujarat Samachar'* from 2001 to 2010, Hindi articles extracted from the daily newspaper, *'Dainik Jagran'* and English articles extracted from *The Telegraph*. Moreover, we extracted the Sanskrit news data from *'All India Radio News'* and *'Samprativartah'* news between 2015 and 2019. In the above collections, both topics and documents use the UTF-8 encoding system.

Table 3.1 shows the statistics of six text corpora. Bengali corpus is the largest (MB) with a good number of documents, and Sanskrit is the smallest, containing a small number of documents. Each document collection comprises 50 topics. We removed 11 and 4 topics from Marathi and Gujarati because no relevant documents were found in the

[2] http://fire.irsi.res.in/fire/static/data

Table 3.1: Shows the statistics of test collection

Year	Collection	Size	Number of documents	Number of terms in the lexicon	Number of queries
2012	Bengali	2.9 GB	5,00,122	13,24,951	50
2010	Marathi	514.9 MB	99,276	8,54,027	50
2011	Gujarati	2.2 GB	3,13,163	20,45,453	50
2011	Hindi	1.3 GB	3,31,608	4,43,243	50
2011	English	1.1 GB	3,92,577	3,93,351	50
2019	Sanskrit	11 MB	7,057	3,38,907	50

collection. Based on the TREC model, each topic comprises three logical sections: a brief title (under the <TITLE> tag) containing two to four words, followed by a description tag (<DESC>tag) containing a one-sentence user's information need, and narrative tag (<NARR> tag) describing relevance assessment criteria. An example of topic representation of Marathi and their English translation is shown in Figure 3.1 and 3.2. We used the Marathi topics [3] and qrels [4] for evaluation. Similarly, topic representation of Bengali, Gujarati, Hindi, English, Sanskrit and their English translation is shown in Figure 3.3, 3.4, 3.5, 3.6, 3.7, 3.8, 3.9 and 3.10. We used the Bengali topics [5] and qrels [6], Gujarati topics [7] and qrels [8], Hindi topics [9] and qrels [10], English topics [11] and qrels [12], Sanskrit topics [13] and qrels [14] for evaluation. An example of a document representation of Marathi is shown in Figure 3.11. In different evaluations of pre-processing strategies, we consider different sections of a query, i.e. title or title and description or title, description

[3]https://www.isical.ac.in/ fire/data/topics/adhoc/mr.topics.76-125.2010.txt
[4]http://www.isical.ac.in/ fire/data/qrels/adhoc/mr.qrels.76-125.2010.txt.gz
[5]https://www.isical.ac.in/ fire/data/topics/adhoc/bn.topics.176-225.2012.txt
[6]https://www.isical.ac.in/ fire/data/topics/adhoc/bn.qrels.176-225.2012-v1.txt
[7]https://www.isical.ac.in/ fire/data/topics/adhoc/gu.topics.176-225.2012.txt
[8]https://www.isical.ac.in/ fire/data/topics/adhoc/gu.qrels.176-225.2012-v1.txt
[9]https://www.isical.ac.in/ fire/data/topics/adhoc/hi.topics.126-175.2011.txt
[10]https://www.isical.ac.in/ fire/data/topics/adhoc/hi.qrels.126-175.2011.txt
[11]https://www.isical.ac.in/ fire/data/topics/adhoc/en.topics.126-175.2011.txt
[12]https://www.isical.ac.in/ fire/data/topics/adhoc/en.qrels.126-175.2011.txt
[13]https://github.com/cse-iitbhu/Sanskrit-Text-Collection/blob/main/query50.trec
[14]https://github.com/cse-iitbhu/Sanskrit-Text-Collection/blob/main/qrelm1-50.txt

and narrative.

<TOP lang='Mar'>
<NUM>79</NUM>
<TITLE>चीन आणि माऊंट एव्हरेस्ट दरम्यान रस्ता बांधणे</TITLE>
<DESC> चीनपासून माऊंट एव्हरेस्टपर्यंत रस्ता बांधण्याची योजना</DESC>
<NARR> संबंधित कागदपत्रात चीनपासून माऊंट एव्हरेस्टपर्यंत रस्ता बांधण्याच्या योजनेचे चित्रण करायला हवे. भारतीय आणि चीनी अधिकाऱ्यांच्या ह्या मुद्याबाबच्या चर्चादेखील संबंधित आहेत. </NARR>
</TOP>

Figure 3.1: An example of topic in Marathi

<NUM > 79 </NUM>
<TITLE> Building a road between China and Mount-Everest </TITLE>
<DESC> Road from China to Mount-Everest < /DESC>
<NARR> Relevant documents should outline a plan to build a road from China to Mount Everest. Discussions between Indian and Chinese officials on the issue are also relevant. </NARR>
</TOP>

Figure 3.2: Marathi topic translation in English

<TOP lang='bn'>
<NUM> 176 </NUM >
<TITLE>ওয়াই এস আর রেড্ডির মৃত্যু</TITLE>
<DESC>অন্ধ্র প্রদেশের মুখ্যমন্ত্রী ওয়াই এস আর রেড্ডির মৃত্যু</DESC>
<NARR>অন্ধ্র প্রদেশের মুখ্যমন্ত্রী ওয়াই এস আর রেড্ডির মৃত্যু হেলিকপ্টার দুর্ঘটনায় হয়েছে , প্রাসঙ্গিক নথিতে এই সংক্রান্ত তথ্য প্রয়োজনীয় </NARR>
</TOP>

Figure 3.3: An example of topic in Bengali

<TOP lang='Guj'>
<NUM>176</NUM>
<TITLE></TITLE>
<DESC></DESC>
<NARR></NARR>
</TOP>

Figure 3.4: An example of a topic in Gujarati

<TOP>
<NUM>176</NUM>
<TITLE> YSR Reddy death</TITLE>
<DESC> Death of Andhra Pradesh Chief Minister YSR Reddy </DESC>
<NARR> Relevant documents should contain information about Andhra Pradesh Chief Minister YSR Reddy's death in a helicopter crash. </NARR>
</TOP>

Figure 3.5: Gujarati and Bengali topic translation in English

<TOP>
<NUM>126</NUM>
<TITLE> स्वाइन फ्लू के टीके</TITLE>
<DESC>स्वाइन फ्लू के प्रतिरोध में भारतीय टीके</DESC>
<NARR>प्रासंगिक प्रलेख में भारत में स्वाइन फ्लू के टीके की तैयारी, मनुष्य और जीव–जंतुओं पर टीके प्रयोग करना, टीके के अभाव को दूर करने की व्यवस्था एवं टीके के माध्यम से जीवन बचाव से सम्बंधित सूचनाएं होनी चाहिए I</NARR>
</TOP>

Figure 3.6: An example of topic in Hindi

<TOP>
<NUM>126</NUM>
<TITLE> Swine flu vaccine</TITLE>
<DESC> Indigenous vaccine made in India for swine flu prevention </DESC>
<NARR> Relevant documents should contain information related to making indigenous swine flu vaccines in India, the vaccine's use on humans and animals, arrangements that are in place to prevent scarcity/unavailability of the vaccine, and the vaccine's role in saving lives.</NARR>
</TOP>

Figure 3.7: Hindi topic translation in English

<TOP>
<NUM>177</NUM>
<TITLE> Musicians Bharat Ratna </TITLE>
<DESC> Information about the Bharat Ratna is awarded to musicians </DESC>
<NARR> Relevant documents should contain information about famous musicians (including vocalists and instrumentalists such as Ravi Shankar, M.S. Subbalakshmi and Lata Mangeshkar) being awarded the Bharat Ratna. Articles about these musicians (e.g. brief biographies, concert reviews) are irrelevant unless they mention that the musician received (or will be receiving) the Bharat Ratna. </NARR>
</TOP>

Figure 3.8: An example of a topic in English

<TOPICS>
<TOP>
<NUM>3</NUM>
<TITLE> दक्षिण–अफ्रीकायाः दशम–ब्रिक्स–सम्मेलनम् </TITLE>
<DESC> दक्षिण–अफ्रीकायाः जोहान्सबर्गे दशम–ब्रिक्स–सम्मेलनं भविष्यति । </DESC>
<NARR> दक्षिण–अफ्रीकायाः जोहान्सबर्गे पञ्चानां ब्रिक्सराष्ट्रप्रमुखाणाम् अध्यक्षतायाम् आयोजितस्य दशम–ब्रिक्ससम्मेलनस्य सम्बन्धिनः विषयाः अत्र भवेयुः । भारतस्य सुदृढ–पारस्परिक–सम्बन्धार्थम् एतत् सम्मेलनम् अति–महत्वपूर्णं विद्यते । अन्यत् किमपि राष्ट्रियम् अन्ताराष्ट्रियं वा सम्मेलनम् अत्र प्रासङ्गिकं नास्ति । </NARR>
</TOP>
</TOPICS>

Figure 3.9: An example of a topic in Sanskrit

<TOPICS>
<TOP>
<NUM>3</NUM>
<TITLE> 10th BRICS summit at South Africa </TITLE>
<DESC> 10th BRICS summit will be held in Johannesburg of South Africa </DESC>
<NARR> Relevant documents should outline the 10th BRICS summit in Johannesburg, South Africa. Discussion about other international summits is irrelevant. </NARR>
</TOP>

Figure 3.10: Sanskrit topic translation in English

<DOC>
<DOCNO>Solapur61B5F4CF38.htm.txt</DOCNO>
<TEXT> जगदंबा सूत गिरणीतून पाच लाखाच्या मालाची चोरी
माढा, ता. १ – येथील जगदंबा सूत गिरणीच्या गोदामातून सुमारे पाच लाखाचा माल चोरीस गेल्याचा तक्रारी अर्ज मालेगाव येथील एस.एम. एन्टरप्राईजचे मालक महेंद्रकुमार शंकरलाल मोदी यांनी माढा पोलिसांना दिला असून पोलिस निरीक्षक सुरेश गुरव याबाबत चौकशी करीत असून अद्याप गुन्हा नोंद केला नसल्याचे सांगितले. ..... याबाबत दिलेल्या तक्रारी अर्जात श्री. मोदी यांनी सूत गिरणीला मी माल पुरवठा करून त्याचे सूत काढून ते देत होते. आता करार आमच्यात झाला होता. कराराप्रमाणे सूत गिरणीने माल ठेवण्यासाठी तीन गोदामे दिली होती. कामगारांचे भविष्य निर्वाह निधीचे पैसे न भरल्यामुळे गोदामांना सील केले. त्यानंतर हायकोर्टाच्या आदेशाप्रमाणे गिरणी पुन्हा चालू झाली. पुढे १० ते १५ दिवस सूत गिरणी चालली तो माल गोदामात ठेवला. हायकोर्टाच्या आदेशाशिवाय माल हलवायचा नव्हता. या मालाचा मी मालक असून त्याचा विमा उतरविला आहे. या सुमारे पाच लाखाचा माल ३० जूनला सायंकाळी मालट्रक (एमएच ०४– पी २६००) मध्ये गोदामाचे सील तोडून नेला. याबाबत चौकशी व्हावी असे श्री. मोदी यांनी तक्रार अर्जात नमूद केले आहे. या अर्जाबाबत चौकशी करत असून अद्याप गुन्हा नोंद केला नाही असे पोलिस निरीक्षक गुरव यांनी सांगितले. या बाबत सूत गिरणीचे अध्यक्ष संजय थोरात यांनी याबाबत अद्याप कोणी कळविले नसल्याचे सांगितले.
</TEXT>
</DOC>

Figure 3.11: An example of a document in Marathi

3.3 Evaluation Strategy

We adopt the following two kinds of evaluation to measure the effectiveness of pre-processing strategies against other state-of-the-art methods.

3.3.1 Direct Evaluation

When we have ready gold-standard data available for measuring the effectiveness, we call the evaluation *direct*. As the name suggests, it provides a raw-level quick idea of how the proposed pre-processing strategies affect the NLP and IR systems. We focus on prediction accuracy to see whether the pre-processing strategies are affected and, if yes, to what extent. Precision, recall, F_1- measure, over-stemming, under-stemming, and

index compression factor are the evaluation metrics that are used during the evaluation (detailed below).

3.3.2 Indirect Evaluation

Often, direct evaluation is not possible as appropriate ground truths are not readily available. We can indirectly measure the effectiveness of a given technique through the change in performance scores in an application when the technique is applied vis-a-vis when not applied, keeping all other parameters and experimental settings the same.

At places, we also conduct indirect evaluation through IR experiments. In IR experiments, one gets a ranked list of documents from a large collection of documents in response to a query. The system provides the ordered list of documents that human experts assess in terms of topical relevance and/or usefulness. This assessment is treated as ground truth for IR experiments, which are used to compute different evaluation metrics like precision, recall, Precision@k, R-precision, average precision, and mean average precision (Eqn. 3.28). When these metrics are used for purely IR tasks, they are direct evaluations. However, when changes in these metric values are seen as the effect of some other task, such evaluations are called *indirect*.

3.4 Evaluation Metrics

Evaluation is done for different pre-processing tasks. Some evaluations are *direct*, using raw statistics. Some are conducted in an *indirect* way by measuring the change in retrieval effectiveness in an IR setting.

3.4.1 Direct Evaluation

Table 3.2 summarizes some definitions related to stemming and helps us define and understand relevant metrics used in the thesis.

(a) **Over-stemming:** It occurs when words are not indeed morphological variants but conflated to a single stem. An example of over-stemming would be that 'universal' and 'university' are conflated to 'univers'. From a stemming perspective, an over-stemming can be treated as a 'false positive' (FP) as a word that should not be

Table 3.2: Confusion Matrix

	Words Stemmed	Not Stemmed
Words to be stemmed	True Positive (TP)	False Negative (FN) (Under stemming)
Words not to be stemmed	False Positive (FP) (Over stemming)	True Negative (TN)

stemmed is unduly being stemmed. The percentage of over-stemming is given by:

$$\% \textit{ of over-stemming} = \frac{\textit{Number of nonvariants conflated}}{\textit{Total number of words conflated}} \tag{3.14}$$

$$= \frac{FP}{TP + FP} \tag{3.15}$$

(b) **Under-stemming:** It occurs when the word is indeed a morphological variant but is not conflated to its stem by the stemmer. For example, if 'India' and 'Indian' are not conflated to the same root, then it is a case of under-stemming. Under-stemming can be treated as a 'false negative' (FN). The percentage of under-stemming is given by:

$$\% \textit{ of under-stemming} = \frac{\textit{Number of variants not conflated}}{\textit{Total number of morphological variants}} \tag{3.16}$$

$$= \frac{FN}{TP + FN} \tag{3.17}$$

(c) **Index compression factor (ICF):** It is defined as the percentage reduction in index size accomplished through stemming. A good stemmer should have the largest index compression factor. ICF is given by:

$$ICF = (\frac{n - s}{n}) \times 100 \tag{3.18}$$

Where n is the total number of words in the corpus, and s is the number of distinct words after stemming.

(d) **Precision:** It is the ratio between the number of correct stems produced by a stemmer (True Positives or TP) and the total number of words stemmed by the stemmer (including the wrongly stemmed ones, called False Positives or FP).

$$Precision = \frac{\text{TP}}{\text{TP+FP}} \tag{3.19}$$

(e) **Recall:** The ratio between the number of correct stems produced by a stemmer and the total number of words that should be stemmed in the collection (the stemmer may or may not stem).

$$Recall = \frac{\text{TP}}{\text{TP+FN}} \tag{3.20}$$

(f) F_1**-measure:** It is the balanced harmonic mean of precision and recall.

$$F_1\text{-}measure = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision+recall}} \tag{3.21}$$

(g) **Accuracy:** The accuracy of a stemmer can be defined as the ratio of correctly identifying the words to be stemmed and not to be stemmed over the total number of words that are stemmed or not stemmed. Mathematically, it is given by Equation 3.22.

$$Accuracy = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \tag{3.22}$$

3.4.2 Indirect Evaluation

The effect of different pre-processing techniques is studied through the prism of change in retrieval effectiveness in an IR setting. Retrieval is done and evaluated after applying a given pre-processing technique and compared with when it is not applied.

We used the following evaluation metrics to measure retrieval effectiveness.

(h) **Precision:** It is the fraction of relevant documents retrieved among the retrieved documents.

$$Precision = \frac{Number\ of\ relevant\ documents\ retrieved}{Number\ of\ retrieved\ documents} \tag{3.23}$$

(i) **Recall:** It is the fraction of relevant documents retrieved from a set of relevant documents.

$$Recall = \frac{Number\ of\ relevant\ documents\ retrieved}{Number\ of\ relevant\ documents\ in\ the\ collection} \tag{3.24}$$

(j) **Precision @ k:** In web-scale IR, recall is not considered a meaningful metric because each query contains thousands of relevant documents, and very few users try to read all of them. Precision at k documents (P@k) is still a meaningful metric because it tries to read the top 10 or 20 documents (e.g., P@10 defines the number of relevant documents retrieved in the top 10 documents).

$$Precision@k = \frac{Number\ of\ relevant\ documents\ retrieved\ in\ top\ k\ documents}{k} \tag{3.25}$$

(k) **R-Prec:** It is the precision when the *Rel* number of documents are retrieved.

$$R\text{-}Prec = \frac{Number\ of\ relevant\ retrieved(r)}{|Rel|} \tag{3.26}$$

$|Rel|$: Number of relevant documents in the collection

(l) **Average Precision (AP):** It is the average of precision values whenever a relevant document is retrieved. R_k stands for the k-th relevant document among a ranked list of retrieved documents. Here, the total number of relevant documents for a given query is 'R':

$$Average\ Precision\ (AP) = \frac{1}{R}\sum_{k=1}^{R} precision(R_k) \tag{3.27}$$

For the relevant documents not retrieved (or not available in the ranked list), precision scores are considered to be zero (0), as if they are retrieved at rank ∞.

(m) **Mean Average Precision (MAP):** AP corresponds to a single query. However, the comparison between two IR models should not be based on a single query but

on a set of queries so that variation in retrieval effectiveness over individual queries is averaged out. Hence, MAP is defined for a set of queries and is computed as the simple arithmetic mean of AP scores of all such queries.

$$\textit{Mean Average Precision (MAP)} = \frac{1}{|Q|}\sum_{t=1}^{|Q|} AP(t) \tag{3.28}$$

$|Q|$: Number of queries

MAP has been the most standard evaluation measure in IR used across different evaluation forums, including TREC[15]. It is the mean of AP values for a number of queries. We computed the MAP values by TREC_EVAL software based on a maximum of 1,000 retrieved documents. Mean as an evaluation measure signifies that we give equal importance to all queries.

3.5 Statistical Tests

3.5.1 t-test

t-test is a popular test of significance that gives us a notion of confidence with which we can see the difference in performance between a pair of experiments or trials. We perform a two-sided t-test to measure the quality of different pre-processing strategies in the Indian language IR. In any test of significance, there are two hypotheses. The baseline or *null hypothesis (*H_0 used in our t-test assumes that the means of the two populations are equal. This is tested against an *alternative hypothesis (*H_1 that assesses whether the means of two populations are different with some pre-defined level of confidence based on the available data. In other words, whether the null hypothesis can be refuted or the difference of means observed is statistically significant [91]. Mathematically, H_0: μ_1 - μ_2=0. On the contrary, the alternate hypothesis comprises the means of two different populations are not equal H_1: μ_1 - $\mu_2 \neq 0$. Moreover, we assume that the hypothesized mean difference is zero (0) and the level of significance (α) is 0.05.

[15]https://trec.nist.gov/

3.5.2 Bonferroni correction

We implement Bonferroni correction [122] to counteract the multiple hypothesis testing. If multiple hypotheses are tested, the probability of observing a rare event increases; therefore, the likelihood of incorrectly rejecting a null hypothesis increases. The Bonferroni correction compensates for that increase by testing each hypothesis at a significance level (α/m), where α is the desired overall significance level and m is the number of hypotheses.

3.5.3 Confidence Interval

A confidence interval (CI) refers to the probability that a population parameter will fall between a set of values for a certain proportion of times [38].

CI=μ $\pm$z$\cdot\frac{\sigma}{\sqrt{s}}$

Where z is the Z-value (e.g., 1.96 for 95% confidence level), μ= sample mean, s is the number of samples, and σ = sample standard deviation.