

DAFTAR ISI

HALAMAN JUDUL	ii
HALAMAN PENGESAHAN	iii
HALAMAN PERNYATAAN	iv
HALAMAN PERSEMBAHAN	v
HALAMAN MOTTO	vi
PRAKATA	vii
LISTING	xii
INTISARI	xiii
ABSTRACT	xiv
I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Perumusan Masalah	3
1.3 Batasan Penelitian	4
1.4 Tujuan Penelitian	5
1.5 Manfaat Penelitian	5
II PENELITIAN TERKAIT	7
2.1 <i>Cross-Language Information Retrieval</i>	7
2.2 Metode CLIR Modern	7
2.3 Penelitian CLIR yang Relevan untuk Pasangan Bahasa Indonesia–Arab	11
III METODOLOGI PENELITIAN	20
3.1 Alur Penelitian	20
3.2 Data Penelitian	20
3.2.1 Pengumpulan Data Korpus Fikih Berbahasa Arab	20
3.2.2 Penyusunan Kueri Bahasa Indonesia	21
3.2.3 <i>Relevance Judgement</i>	24

3.2.4	Pembangkitan Data Latih Sintetis	24
3.2.5	Representasi dan Skema Data	27
3.3	Prapemrosesan Data dan Penerjemahan Kueri	29
3.3.1	Prapemrosesan Kueri	31
3.3.2	Penerjemahan Kueri	32
3.3.3	Prapemrosesan Dokumen	33
3.4	Arsitektur Sistem dan Skenario Eksperimen	36
3.4.1	Skenario 1: <i>Sparse Retrieval</i> berbasis Penerjemahan	37
3.4.2	Skenario 2: <i>Dense Retrieval</i> Multibahasa	40
3.4.3	Skenario 3: <i>Dense Retrieval</i> Ekabahasa Berbasis Penerjemahan	45
3.4.4	Skenario 3: <i>Dense Retrieval</i> Ekabahasa Berbasis Penerjemahan	46
3.4.5	Eksperimen Tambahan: <i>Hybrid Retrieval</i> Menggunakan <i>Reciprocal Rank Fusion</i> (RRF)	47
3.5	Rancangan Pengujian dan Evaluasi	49
3.5.1	Metrik Evaluasi	49

IV IMPLEMENTASI 52

4.1	Lingkungan Perangkat Lunak dan Keras	52
4.2	Implementasi Prapemrosesan	53
4.2.1	Implementasi Validasi Korpus	53
4.2.2	Implementasi Pemotongan Dokumen (<i>Chunking</i>)	54
4.2.3	Implementasi Pembersihan Teks (<i>Cleaning</i>)	55
4.2.4	Implementasi Normalisasi dan <i>Light Stemming</i>	57
4.3	Implementasi Pembangkitan Data Latih Sintetis (JH-POLO)	59
4.3.1	Implementasi Seleksi Pasangan Dokumen (<i>Hard Negatives</i>)	59
4.3.2	Penyelesaian Kendala Konteks LLM (<i>Chunking</i>)	60
4.3.3	Implementasi Pembangkitan Kueri dengan LLM	60
4.3.4	Implementasi Validasi Tripel Kueri Menggunakan <i>Cross-Encoder</i>	61
4.4	Implementasi Skenario 1: <i>Sparse Retrieval</i> Berbasis Penerjemahan	62
4.4.1	Persiapan Data dan Keselarasan Prapemrosesan Leksikal	62
4.4.2	Pembangunan Indeks Dokumen (<i>Inverted Index</i>)	64
4.4.3	Pencarian Parameter Optimal BM25 dan RM3	65
4.4.4	Evaluasi Pencarian Terintegrasi	66
4.5	Implementasi Skenario 2: <i>Dense Retrieval</i> Multibahasa	67
4.5.1	Persiapan Data Latih dan Pemuatan Model	68

4.5.2	Pelatihan Adaptasi Domain (<i>Targeted K-Fold</i> dan MNRL) . . .	69
4.5.3	Pembangunan Indeks Vektor (FAISS)	73
4.5.4	Evaluasi Pencarian Terintegrasi	75
4.6	Implementasi Skenario 3: <i>Dense Retrieval</i> Ekabahasa	77
4.6.1	Persiapan Data dan Pemuatan Model Pralatih	77
4.6.2	Pembangunan Indeks Vektor (FAISS)	79
4.6.3	Evaluasi Pencarian Terintegrasi	80
4.7	Implementasi <i>Reciprocal Rank Fusion</i> (RRF)	82
4.7.1	Integrasi <i>Retrieval Pipeline</i>	83
4.7.2	Algoritma <i>Reciprocal Rank Fusion</i>	84
4.7.3	Eksekusi Fusi dan Evaluasi	85
V	HASIL DAN PEMBAHASAN	87
5.1	Rancangan Eksperimen dan Parameter Pengujian	87
5.2	Hasil Kinerja Keseluruhan (Analisis Kuantitatif)	88
5.3	Analisis Skenario 1: Pengaruh Kualitas Terjemahan dan Ekspansi Kueri (RM3)	89
5.4	Analisis Skenario 2: Dampak <i>In-Domain Fine-Tuning</i> pada <i>Dense Retrieval</i> Multibahasa	92
5.5	Analisis Skenario 3: Evaluasi Model <i>Dense Retrieval</i> Ekabahasa AraDPR pada Korpus Lintas Domain	94
5.6	Analisis Kinerja Model Berdasarkan Tipe Kueri	96
5.7	Evaluasi Kualitatif dan Analisis Kegagalan Sistem	99
5.8	Eksperimen Tambahan: <i>Hybrid Retrieval</i>	103
VI	SIMPULAN DAN SARAN	105
6.1	Simpulan	105
6.2	Saran	106
A	Prompt JH-POLO	115
B	Instruksi Prompt (System Role) Generative AI	118
C	Hasil Evaluasi Pencarian Hyperparameter (Targeted K-Fold)	119
3.1	DistilBERT-base-multilingual	119
3.2	mmBERT-small	121

3.3	mmBERT-base	122
D	Hasil Penarikan Dokumen (<i>Top-5 Retrieval</i>) pada Kasus Anomali AraDPR123	
E	Peringkat Kueri per Model	129
F	Daftar Kueri Gagal Total pada Seluruh Model	146