UNIVERSITAS
GADJAH MADA
Arsitektur Deep Constrained Clustering Berbasis Dual-Stream Rnn untuk Closed-Domain Teks Keislaman
Berbahasa Indonesia
Muhammad Zidny Naf'an, Prof. Drs. Edi Winarko, M.Sc., Ph.D.; Dr. Sigit Priyanta, S.Kom., M.Kom.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

DAFTAR ISI

PRAKATA v
DAFTAR ISI vii
DAFTAR TABEL ix
DAFTAR GAMBAR x
DAFTAR LAMPIRAN xi
INTISARI xii
ABSTRACT xiii
BAB 1 PENDAHULUAN 1
1.1 Latar Belakang 1
1.2 Rumusan Masalah 5
1.3 Batasan Penelitian 6
1.4 Tujuan Penelitian 7
1.5 Manfaat Penelitian 7
1.6 Kontribusi Penelitian 8
1.7 Sistematika Penulisan 9
BAB 2 TINJAUAN PUSTAKA 11
2.1 *Constrained Clustering* 11
2.2 Penelitian-penelitian pada *Deep Clustering* 13
2.3 Representasi Teks untuk *Deep Clustering* 16
2.3.1 Representasi Tradisional: *Bag-of-Words* dan TF-IDF 16
2.3.2. *Word Embeddings* untuk Klastering 18
2.3.3. Representasi Teks berbasis RNN 19
2.3.4. Representasi *Multi-level*: *Character* dan *Word Level* 22
2.4 *Deep Text Clustering* 25
2.5 *Deep Constrained Clustering* 30
2.6 NER dalam Domain Studi Islam 35
2.7 Identifikasi Celah Penelitian 36
BAB 3 LANDASAN TEORI 42
3.1 Algoritma Klastering Klasik 42
3.2 *Recurrent Neural Network* 44
3.3 *Long-Short Term Memory* 49
3.4 *Bidirectional Long-Short Term Memory* (Bi-LSTM) 50
3.5 *Word Embeddings* 51
3.5.1. Non-contextual Embeddings 51
3.5.2. Contextual Embeddings 53
3.6 Metrik Evaluasi Klastering 55
BAB 4 METODOLOGI PENELITIAN 58
4.1 Deskripsi Umum dan Tahapan Penelitian 58
4.2 Pengembangan *Dataset* MAQASID 60
4.2.1. Sumber dan Proses Pengumpulan Data 60
4.2.2. Normalisasi Kategori dan Validasi (IAA) 65
4.2.3. Analisis Distribusi *Dataset* 65
4.3. Tahap *Preprocessing* dan Representasi Teks 67
4.3.1. Skenario *Preprocessing* 67

UNIVERSITAS
GADJAH MADA
Arsitektur Deep Constrained Clustering Berbasis Dual-Stream Rnn untuk Closed-Domain Teks Keislaman
Berbahasa Indonesia
Muhammad Zidny Naf'an, Prof. Drs. Edi Winarko, M.Sc., Ph.D.; Dr. Sigit Priyanta, S.Kom., M.Kom.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

4.3.2. Pelatihan *Word Embeddings* 68
4.3.3. Mekanisme Pembangkitan *Constraint* (ML/CL) dari IndQNER.. 70
4.4. Arsitektur *Deep Constrained Islamic Text Clustering* (DCITC)..... 73
4.4.1. Tahap *Preprocessing* untuk *Dataset* MAQASID 75
4.4.2. Feature Extraction dan *Adaptive Gate Fusion Layer* 76
4.4.3. *Joint Optimization and Entity Pairwise Constraint Loss*..... 79
4.5. Strategi Evaluasi 82
4.5.1. Evaluasi *Word Embeddings* (Intrinsik dan Ekstrinsik) 82
4.5.2. Evaluasi *Deep Constrained Islamic Text Clustering* 83
4.5.3. Konfigurasi Eksperimen dan Parameter..... 84
5. HASIL DAN PEMBAHASAN 86
5.1. Eksplorasi *Dataset* MAQASID 86
5.1.1. Statistik Korpus dan Distribusi Kategori 86
5.1.2. Analisis Frekuensi *Term* 88
5.1.3. Karakterisasi Entitas NER pada MAQASID 89
5.2. Hasil Pelatihan *Word Embeddings* 91
5.2.1. Evaluasi Intrinsik *Non-Contextual Embeddings* 91
5.2.2. Evaluasi Ekstrinsik *Word Embeddings* 95
5.2.3. Sintesis: Pemilihan Konfigurasi *Embedding* Optimal 97
5.3. Evaluasi *Baseline:* TF-IDF + K-Means..... 98
5.4. Evaluasi Arsitektur *Deep Constrained Islamic Text Clustering*..... 100
5.4.1. Pengaruh Strategi *Preprocessing* 100
5.4.2. Analisis Konvergensi dan Stabilitas Pelatihan..... 103
5.4.3. Perbandingan Paradigma Embedding: *Standard Embeddings* dan IndoBERT..... 106
5.4.4. Optimasi *Hyperparameter* 107
5.4.5. Ringkasan Konfigurasi Optimal..... 110
5.5. Temuan Kritis: Anomali *Latent Space* 111
5.5.1. Disparitas Metrik Intrinsik dan Ekstrinsik 112
5.5.2. Fenomena *Ubiquitous Entity Trap* 113
5.5.3. *Semantic Mismatch* dan *Feature Collapse*..... 116
5.6. Analisis Komparatif: DCITC vs *Baseline* 119
5.6.1. Perbandingan Metrik Ekstrinsik..... 119
5.6.2. Perbandingan Metrik Intrinsik 121
5.6.3. *Trade-off* Struktural vs Semantik..... 123
5.7. Analisis Kualitatif Hasil Klastering DCITC..... 125
5.7.1. Kemurnian Klaster dan Representatif Klaster..... 126
5.7.2. Fenomena Klastering Berbasis Entitas Dominan..... 129
5.7.3. Pengaruh Distribusi Kategori terhadap Dinamika Klaster..... 131
5.8. Sintesis Hasil dan Pembahasan..... 132
6. PENUTUP 135
6.1. Kesimpulan..... 135
6.2. Generalisasi dan Keterbatasan Penelitian..... 138
6.3. Rekomendasi Penelitian Lanjutan 139
DAFTAR PUSTAKA 140

UNIVERSITAS
GADJAH MADA
Arsitektur Deep Constrained Clustering Berbasis Dual-Stream Rnn untuk Closed-Domain Teks Keislaman
Berbahasa Indonesia
Muhammad Zidny Naf'an, Prof. Drs. Edi Winarko, M.Sc., Ph.D.; Dr. Sigit Priyanta, S.Kom., M.Kom.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

DAFTAR TABEL

Tabel 2.1. Daftar penelitian *deep constrained clustering* 33
Tabel 2.2. Kajian pustaka pada *deep clustering* dan *constrained clustering* 38
Tabel 4.1. Peringkat media *online* keislaman (RedaksiIB, 2020) dan afiliasinya 61
Tabel 4.2. Banyaknya artikel yang dikumpulkan dari sumber *website* 64
Tabel 4.3. Distribusi Dokumen per Kategori 66
Tabel 4.4. Statistik Panjang Dokumen 66
Tabel 4.5. Skenario *Preprocessing* 67
Tabel 4.6. Variasi transliterasi Bahasa Arab ke Bahasa Indonesia 68
Tabel 4.7. Konfigurasi *parameter* pelatihan Word2Vec dan fastText 69
Tabel 4.8. Konfigurasi parameter *fine-tuning* IndoBERT 70
Tabel 4.9. Parameter tahap pelatihan *word embeddings* 75
Tabel 4.10. Nilai parameter pada pelatihan *word embeddings* 84
Tabel 4.11. Nilai *parameter tuning* untuk arsitektur DCITC 84
Tabel 5.1. Jumlah kalimat dan kata setiap kategori 87
Tabel 5.2. Banyaknya kemunculan label NER IndQNER + MAQASID 89
Tabel 5.3 Contoh Hasil *Tagging* IndQNER pada Penggalan Dokumen ke-4 91
Tabel 5.4. Top-10 kata terdekat dengan kata “abadi” 92
Tabel 5.5. Hasil evaluasi klasifikasi untuk *non-contextual embeddings* 96
Tabel 5.6. Hasil macro-F1 pada *contextual word embeddings* 97
Tabel 5.7. Konfigurasi *hyperparameter clustering baseline* 98
Tabel 5.8. Hasil eksperimen dari model dasar klastering 99
Tabel 5.9. Statistik Deskriptif *Test* NMI Berdasarkan *Preprocessing* 101
Tabel 5.10. Statistik Deskriptif *Test* ARI Berdasarkan Metode *Preprocessing* . 102
Tabel 5.11. Statistik Deskriptif nilai *Loss* Berdasarkan *Learning Rate* 104
Tabel 5.12. Perbandingan Metrik Evaluasi pada *Standard Embeddings* dan *Fine-Tuning* IndoBERT 106
Tabel 5.13. Dampak *hyperparameter* terhadap metrik evaluasi klastering 108
Tabel 5.14. Konfigurasi optimal per metrik evaluasi 111
Tabel 5.15. Perbandingan Metrik Evaluasi pada Model *Embeddings* 112
Tabel 5.16. Jumlah Dokumen untuk setiap Label IndQNER 114
Tabel 5.17. Nilai metrik ekstrinsik tertinggi berdasarkan arsitektur 120
Tabel 5.18. Ringkasan Nilai *Test* SC 121
Tabel 5.19. Eksperimen SC tertinggi pada *Standard Embeddings* dengan $SLC = 4096$, $\lambda = 0{,}1$ dan $LR = 0{,}001$ 123
Tabel 5.20. Ringkasan kemurnian klaster DCITC vs Baseline 126
Tabel 5.21. Dokumen representatif DCITC dengan nilai $qmax$ tertinggi 127
Tabel 5.22. Dokumen representatif Baseline TF-IDF + K-Means 128
Tabel 5.23. Klaster DCITC yang memuat kategori Fikih dan Sejarah 129
Tabel 5.24. Recall per kategori pada DCITC dan Baseline 131

UNIVERSITAS
GADJAH MADA
Arsitektur Deep Constrained Clustering Berbasis Dual-Stream Rnn untuk Closed-Domain Teks Keislaman
Berbahasa Indonesia
Muhammad Zidny Naf'an, Prof. Drs. Edi Winarko, M.Sc., Ph.D.; Dr. Sigit Priyanta, S.Kom., M.Kom.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

DAFTAR GAMBAR

Gambar 2.1. Arsitektur umum *deep clustering* (Aljalbout dkk., 2018) 13
Gambar 2.2. Konteks kata "surat" dalam *closed-domain* dan *open-domain* dan.. 17
Gambar 3.1. Ilustrasi RNN dasar (Goodfellow dkk., 2016) 45
Gambar 3.2. Arsitektur RNN dengan output (Goodfellow dkk., 2016)..... 46
Gambar 3.3. Arsitektur RNN pada *language model* (Jurafsky & Martin, 2026).. 47
Gambar 4.1. Tahapan penelitian 59
Gambar 4.2. Contoh proses *web scraping* 63
Gambar 4.3. Sampel *dataset* MAQASID..... 64
Gambar 4.4. Metode penelitian pengembangan *word embeddings* 69
Gambar 4.5. Diagram Tahapan DCITC 74
Gambar 4.6. Mekanisme *feature extraction* pada arsitektur DCITC 77
Gambar 4.7. Alur *joint optimization* pada arsitektur DCITC 79
Gambar 4.8. Tangkapan layar Kateglo pada Laman Glosarium Bidang Agama Islam..... 82
Gambar 5.1. Distribusi normalisasi kategori..... 87
Gambar 5.2. Distribusi 20 kata terbanyak dalam MAQASID 88
Gambar 5.3. Frekuensi 20 kata terbanyak tanpa *stopword* 89
Gambar 5.4. Persentase kata berelasi dalam top-10 kata terdekat untuk setiap skenario eksperimen..... 92
Gambar 5.5. Persentase keberhasilan model dalam *analogy task*..... 94
Gambar 5.6. Perbandingan *Test* NMI Berdasarkan Metode *Preprocessing* 100
Gambar 5.7. Perbandingan *Test* ARI Berdasarkan Metode *Preprocessing* 102
Gambar 5.8. Kurva *training loss* dan *validation loss* pada Eksp-1..... 103
Gambar 5.9. Perbandingan *Loss* pada LR yang Berbeda..... 105
Gambar 5.10. Perbandingan *Test* NMI Berdasarkan *Learning Rate* 105
Gambar 5.11. Proyeksi t-SNE dua dimensi atas ruang laten DCITC Eksp-106 (Panel A dan B) dan ruang TF-IDF (Panel C dan D) pada *test set* MAQASID..... 117
Gambar 5.12. Nilai SC tertinggi untuk setiap skenario pada *embeddings*..... 122
Gambar 5.13. *Scatter Plot Test* SC vs *Test* NMI berdasarkan *Embeddings* 124
Gambar 5.14. Komposisi kategori *ground truth* per klaster: DCITC Eksp-106 vs *Baseline* TF-IDF + K-Means 127
Gambar 5.15. *Recall* per kategori vs proporsi *ground truth* pada *test set*..... 132

UNIVERSITAS
GADJAH MADA
Arsitektur Deep Constrained Clustering Berbasis Dual-Stream Rnn untuk Closed-Domain Teks Keislaman
Berbahasa Indonesia
Muhammad Zidny Naf'an, Prof. Drs. Edi Winarko, M.Sc., Ph.D.; Dr. Sigit Priyanta, S.Kom., M.Kom.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

DAFTAR LAMPIRAN

Lampiran 1. Pemetaan Kategori Asli ke Daftar Tajuk Islam 153
Lampiran 2. Metrik Evaluasi Baseline (TF-IDF + K-Means) 156
Lampiran 3. Konfigurasi Eksperimen DCITC-Standard Embeddings 157
Lampiran 4. Waktu Komputasi dan *Maximum Epochs* pada Eksperimen DCITC-Standard Embeddings..... 159
Lampiran 5. Hasil Evaluasi DCITC-Standard Embeddings 161
Lampiran 6. Konfigurasi Eksperimen DCITC-IndoBERT 164
Lampiran 7. Waktu Komputasi dan *Maximum Epochs* pada Eksperimen DCITC-IndoBERT 166
Lampiran 8. Hasil Evaluasi DCITC-IndoBERT 168