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/**

# INTISARI

Klastering teks *closed-domain* berbahasa Indonesia pada bidang Studi Islam menghadapi tantangan karakteristik teks yang unik, meliputi dominasi kosakata teologis, banyaknya variasi morfologis, dan inkonsistensi transliterasi serapan Bahasa Arab. Penelitian ini mengatasi ketiadaan *dataset benchmark* dan keterbatasan representasi tradisional dengan tiga kontribusi utama. Pertama, ketersediaan *dataset* MAQASID yang berisi 52.374 dokumen literatur Islam dari 13 portal web Islam berbahasa Indonesia dengan validasi melalui *Inter-Annotator Agreement*. Kedua, arsitektur *Deep Constrained Islamic Text Clustering* (DCITC) sebagai *end-to-end pipeline* yang mengintegrasikan tiga komponen: (a) *preprocessing* teks; (b) *feature extraction* berbasis Bi-LSTM pada level karakter dan kata, *pairwise constraint Must-Link* dan *Cannot-Link* yang diturunkan dari label entitas IndQNER, serta *Adaptive Gate Fusion Layer* yang memadukan representasi karakter, kata, dan entitas menjadi vektor laten $z$ melalui *gating* adaptif; dan komponen (c) *joint optimization* berbasis divergensi Kullback-Leibler dan *entity pairwise constraint loss margin-based* secara simultan dengan bobot λ. Ketiga, analisis empiris dampak *pairwise constraint* berbasis NER pada ruang laten yang menghasilkan fenomena baru bagi khazanah *constrained deep clustering* pada teks *closed-domain*.

Evaluasi dilakukan pada 128 skenario eksperimen DCITC dengan variasi *preprocessing*, *embeddings* (*Standard Embeddings* dan *fine-tuning* IndoBERT), *sequence length word* (SLW), *sequence length character* (SLC), bobot λ, dan *learning rate* (LR), serta dibandingkan dengan *baseline* TF-IDF + K-Means pada 16 konfigurasi. Konfigurasi optimal DCITC diperoleh pada Eksperimen 106 (*fine-tuning* IndoBERT, *preprocessing stemming*, SLW = 512, SLC = 2.048, λ = 0,1, LR = 0,0001) dengan *test* NMI 0,1018 dan ARI 0,0465. DCITC unggul secara struktural dengan *Silhouette Coefficient* (SC) maksimum 0,8406 dibandingkan *baseline* (0,0120). Namun, tidak ada dari 128 konfigurasi DCITC yang melampaui capaian ekstrinsik *baseline* (NMI 0,1764; ARI 0,1308).

Analisis mendalam mengungkap tiga fenomena kritis yang menjelaskan kesenjangan tersebut. *Ubiquitous Entity Trap* menunjukkan bahwa dominasi entitas B-Messenger mempengaruhi *Must-Link constraint* dan memaksa dokumen lintas kategori mengelompok ke klaster yang sama. *Semantic Mode Collapse* ditandai oleh ko-eksistensi SC tinggi (0,8406) dengan NMI mendekati nol (0,0004), bukti pemadatan geometrik ruang laten tanpa keselarasan semantik. *U-shape divergence* pada 120 dari 128 eksperimen menandai *feature collapse* akibat ketiadaan *reconstruction loss* pada arsitektur. Temuan ini menegaskan bahwa kesenjangan tersebut bersumber pada karakteristik *hard-constraint* biner dari entitas dominan, bukan dari limitasi ekstraksi fitur *dual-stream* Bi-LSTM, sehingga integrasi pengetahuan domain memerlukan mitigasi berupa *soft constraint*, *entity whitelisting*, maupun *reconstruction loss*.

Kata kunci: *closed domain*, *deep constrained clustering*, *k-means, Bi-LSTM*, *pairwise constraint*, representasi teks, *ubiquitous entity trap.*

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/**

# ABSTRACT

Clustering closed-domain Indonesian Islamic texts faces challenges like dominant theological vocabulary, morphological variations, and inconsistent Arabic transliterations. This study addresses the lack of benchmark datasets and traditional representation limits through three contributions. First, the MAQASID dataset, containing 52,374 validated documents. Second, the Deep Constrained Islamic Text Clustering (DCITC) architecture as an end-to-end pipeline that integrates three components: (a) text preprocessing; (b) Bi-LSTM-based feature extraction at the character and word levels, Must-Link and Cannot-Link pairwise constraints derived from IndQNER entity labels, and an Adaptive Gate Fusion Layer that fuses character, word, and entity representations into a latent vector $z$ through adaptive gating; and component (c) joint optimization based on Kullback-Leibler divergence and a margin-based entity pairwise constraint loss simultaneously with weight $\lambda$. Third, an empirical analysis of the impact of NER-based pairwise constraints on the latent space, yielding a new phenomenon for the constrained deep clustering literature on closed-domain texts.

Evaluation was performed on 128 DCITC experimental scenarios varying preprocessing, embedding types (Standard vs. fine-tuned IndoBERT), sequence length word (SLW), sequence length character (SLC), λ, and learning rate (LR), and compared against a TF-IDF + K-Means baseline over 16 configurations. The optimal DCITC configuration was obtained in Experiment 106 (fine-tuning IndoBERT, stemming preprocessing, SLW = 512, SLC = 2,048, λ = 0.1, LR = $1 \times 10^{-4}$) with test NMI 0.1018 and ARI 0.0465. DCITC excels structurally with a maximum Silhouette Coefficient (SC) of 0.8406 versus the baseline's 0.0120, yet none of the 128 configurations surpassed the baseline's extrinsic performance (NMI 0.1764; Adjusted Rand Index 0.1308).

In-depth analysis revealed three critical phenomena that explain this gap. Ubiquitous Entity Trap shows that the B-Messenger entity, occurring in 60.74% of MAQASID documents, contaminates Must-Link constraints and forces documents from different categories into the same cluster. Semantic Mode Collapse is marked by the co-existence of a high SC (0.8406) and near-zero NMI (0.0004), evidencing geometric compaction of the latent space without semantic alignment to the ground truth. U-shape divergence, observed in 120 of 128 experiments, signals feature collapse due to the absence of a reconstruction loss. These findings establish that the gap stems from the binary hard-constraint derived from dominant entities, not from the feature extraction capacity of the dual-stream Bi-LSTM. Integrating domain knowledge into deep constrained clustering on closed-domain text therefore requires mitigation mechanisms such as confidence-based soft constraints, entity whitelisting, and a reconstruction loss.

*Keywords*: *closed domain*, *deep constrained clustering*, *k-means, Bi-LSTM*, *pairwise constraint*, *text representation*, *ubiquitous entity trap.*