UNIVERSITAS
GADJAH MADA
Medical Concept Normalization pada Data Teks Media Sosial Berbasis SapBERT dan Contextual Information Fusion Similarity
Dyandra Aulia, Yunita Sari, S.Kom., M.Sc., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

INTISARI

MEDICAL CONCEPT NORMALIZATION PADA DATA TEKS MEDIA SOSIAL BERBASIS SAPBERT DAN *CONTEXTUAL INFORMATION FUSION SIMILARITY*

Oleh

Dyandra Aulia 21/473733/PA/20423

Medical Concept Normalization (MCN) merupakan tugas pemetaan sebutan entitas medis dalam teks bebas menuju konsep medis baku pada suatu terminologi terstandar. Penelitian mengenai MCN telah berkembang dari pendekatan berbasis aturan hingga metode *deep learning* berbasis *transformer*. Namun, pada data teks media sosial, variasi bahasa kolokial, ambiguitas sebutan, dan keberadaan konsep medis yang tidak pernah muncul pada data pelatihan (*unseen concepts*) masih menjadi tantangan utama. Permasalahan *unseen concepts* penting dalam aplikasi dunia nyata karena sistem sering dihadapkan pada konsep medis baru yang tidak tersedia selama proses pelatihan. Meskipun beberapa penelitian sebelumnya telah menerapkan skenario *zero-shot,* kemampuan generalisasi terhadap *unseen concepts* masih terbatas.

Penelitian ini mengusulkan pendekatan *generate-and-rank* untuk tugas MCN pada teks media sosial yang dievaluasi menggunakan *dataset* COMETA dengan skema *zero-shot specific*. SapBERT yang telah dipralatih menggunakan pengetahuan medis UMLS dimanfaatkan untuk membentuk representasi semantik konsep dan sebutan medis pada tahap pembangkitan kandidat dan pemeringkatan ulang kandidat. Selanjutnya, tahap pemeringkatan ulang kandidat dilengkapi dengan *Contextual Information Fusion Similarity* (CIFS) untuk mengintegrasikan informasi konteks kalimat secara lebih mendalam pada proses pemilihan konsep akhir.

Hasil pengujian menunjukkan performa *Precision@1*, *Recall@1,* dan *Accuracy@1* sebesar 65,56%, serta *Recall@25* sebesar 87,70%, pada keseluruhan data uji, melampaui penelitian sebelumnya yang menggunakan *dataset* dan skenario evaluasi yang sama. Hal tersebut menunjukkan bahwa kombinasi SapBERT dan CIFS efektif dalam meningkatkan kemampuan generalisasi sistem terhadap variasi bahasa medis informal, ambiguitas sebutan, dan konsep medis yang belum pernah terlihat selama pelatihan.

Kata Kunci: *Medical Concept Normalization*; SapBERT; *Contextual Information Fusion Similarity;* SNOMED-CT; Teks Media Sosial; *Zero-Shot Evaluation*

UNIVERSITAS
GADJAH MADA
Medical Concept Normalization pada Data Teks Media Sosial Berbasis SapBERT dan Contextual Information Fusion Similarity
Dyandra Aulia, Yunita Sari, S.Kom., M.Sc., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

ABSTRACT

MEDICAL CONCEPT NORMALIZATION ON SOCIAL MEDIA TEXT USING SAPBERT AND CONTEXTUAL INFORMATION FUSION SIMILARITY

Oleh

Dyandra Aulia 21/473733/PA/20423

Medical Concept Normalization (MCN) is the task of mapping medical entity mentions in free text to standardized medical concepts within a standard terminology. Research in this area has progressed from rule-based methods to transformer-based deep learning approaches. Nevertheless, social media text presents persistent challenges, including colloquial language variation, mention ambiguity, and the presence of unseen concepts absent from training data. The unseen concepts problem is particularly critical in real-world deployment, as systems must generalize to medical concepts not observed during task-specific training. Although prior work has explored zero-shot evaluation, generalization to unseen concepts remains an open challenge.

This study proposes a generate-and-rank framework for MCN on social media text, evaluated on the COMETA dataset under the zero-shot specific scheme. SapBERT, pre-trained on UMLS medical knowledge, is employed to construct semantic representations of both medical mentions and concepts across the candidate generation and re-ranking stages. The re-ranking stage is further enhanced with Contextual Information Fusion Similarity (CIFS), which incorporates sentence-level contextual information to improve final concept selection.

Experiments demonstrate Precision@1, Recall@1, and Accuracy@1 of 65.56%, alongside Recall@25 of 87.70% on the full test set, outperforming prior systems evaluated on the same dataset and protocol. These results suggest that combining SapBERT with CIFS effectively improves generalization across informal medical language variation, ambiguous mentions, and concepts unseen during training.

Keywords: Medical Concept Normalization; SapBERT; Contextual Information Fusion Similarity; SNOMED-CT; Social Media Text; Zero-Shot Evaluation