ABSTRACT

The problem of utilizing machine learning approach in Indonesia Named Entity Recognition system is the limited amount of labelled data for training process. However, unlike the limited availability of labelled data, unlabelled data is widely available from many sources. This enables a semi-supervised learning approach to solve this NER system problem.

This research aims to design a semi-supervised learning model to solve NER system problem. Co-training algorithm is used to utilize unlabelled data in NER learning process to produce new labelled data that can be applied to enhance a new NER classification system.

This research uses two kinds of data, Indonesian DBPedia data as labelled data and news article text from Indonesian news sites (kompas.com, cnnindonesia.com, tempo.co, merdeka.com and viva.co.id) as unlabelled data. The pre-processing steps applied to analyze unstructured text are sentence segmentation, tokenization, stemming, and PoS Tagging. The results of this pre-process are entity candidates and its context, these used as unlabelled data for the co-training process. The SVM algorithm is used as a classification algorithm in this process. 10 Cross Fold Validation is used as the system testing approach. Based on the result of the NER testing system, the **precision** score is **79.15%,** the **recall** score is **75.13%** and **f1 mean score** is **77.05%**.

Keywords : NER, Bahasa Indonesia, Semi-Supervised Learning, Co-Training, SVM, Precision, Recall, F1Score

INTISARI

Pendekatan *machine learning* dalam penyelesaian permasalahan NER berbahasa Indonesia memiliki kendala terbatasnya data berlabel sebagai data training. Terbatasnya data training berbanding terbalik dengan melimpahnya data tanpa label. Kondisi terbatasnya data berlabel dan ketersediaan data tanpa label memberikan peluang model pembelajaran *semi-supervised learning* dalam penyelesaian masalah NER.

Penelitian ini bertujuan untuk merancang sebuah model pembelajaran *semi-supervised learning* dalam membantu penyelesaian permasalahan NER. Model pembelajaran *semi-supervised* dengan menggunakan algoritme *co-training* digunakan untuk memanfaatkan data tanpa label dalam proses pembelajaran sistem NER. Model ini diharapkan dapat menghasilkan data berlabel baru yang dapat digunakan untuk membentuk mesin klasifikasi NER.

Dalam proses jalannya penelitian, terdapat dua tipe *dataset* yang digunakan yaitu data berlabel DBPedia Indonesia dan data tanpa label yang berasal dari artikel situs berita nasional di Indonesia antara lain: Kompas.com, Cnnindonesia.co.id, Tempo.co, Merdeka.com dan Viva.co.id. Tahap *pre-processing* yang dilakukan pada *unstructured text* yaitu *sentence segmentation, tokenization, stemming* dan *PoS Tag*. Proses *pre-processing* menghasilkan *view1* (*named entity*) dan *view2* (konteks) sebagai data tanpa label untuk proses pembelajaran *semi-supervised*. Dalam pembelajaran *semi-supervised* ini SVM digunakan sebagai algoritme klasifikasinya. Proses pembelajaran dan pengujian sistem NER dilakukan dengan pendekatan 10 *Cross Fold Validation*. Hasil pengujian sistem NER menunjukkan nilai ***precision*** sebesar **79,15%**, ***recall*** sebesar **75,13%** dan rata-rata ***f1 score*** sebesar **77,05%**.

Kata kunci – NER, Bahasa Indonesia, Semi-Supervised Learning, Co-Training, SVM, Precision, Recall, F1Score