# Performance Evaluation of SVM-Based Information Extraction using τ Margin Values

Kuspriyanto[1,2], Oerip S Santoso[2], Dwi H Widyantoro[2], Husni S Sastramihardja[2], Kurnia Muludi[2,3], and Siti Maimunah[2,4]

[1]Computer Engineering Research Group,
[2]School of Electrical Engineering and Informatics,
Bandung Institute of Technology, Jl. Ganeca 10 Bandung, Indonesia
[3]Soil Science Department, Agriculture Faculty - University of Lampung, Indonesia
Jl. Sumantri Brojonegoro No. 1 Bandar Lampung 35145
[4]Information System Dept., Information Tech. Faculty,
Surabaya Adhitama Institute of Technology
Jl. A.R. Hakim No.100 Surabaya, Indonesia

**Abstract:** The rapid growth of Internet causes the abundance of textual information. It is necessary to have smart tools and methods than can access text content as needed. One of the success methods is Support Vector Machine (SVM). This paper will discuss how the performance of the SVM-GATE algorithm on extracting information from Indonesian language corpus in response to $\tau$ margin variation. Experimental results show that there is optimum $\tau$ margin for both Indonesian corpus of Vegetable Market and Seminar Announcement Corpus. The best Performance of SVM-GATE obtained at the $\tau$ Margin of 0.5 and the Window Size of 4x4.

**Keywords:** Information Extraction, Support Vector Machine, Bahasa Indonesia Corpus, NLP, GATE, optimum margin.

## 1. Introduction

Along with the rapid Internet development, the volume of textual information is also incredibly growing. Currently Information Retrieval technology alone is not able to provide specific information needs because this technology only provides information on the level of document collection. Development tool and intelligent methods that can access the content of the document are therefore a crucial issue.

Information extraction is the process of getting information about the pre-specified events, entities or relationships in the text such as news articles (Newswire) and web pages. Many research of information extraction are focused on named entity recognition. In general information extraction task can be regarded as an entity recognition task in the text. Extraction of information is very useful in many applications such as business intelligence, automatic annotations on web pages, and knowledge management.

Extraction of information can be approached through a classification problem where the text is split into tokens and grouped into the appropriate class. *Hidden Markov Models* are a popular method for the task, but this method cannot handle multiple tokens with attribute [1].

One of successful machine learning methods in the extraction of information is the *Support Vector Machine* (SVM), which is part of the supervised machine learning algorithms. This algorithm has achieved the performance state-of-the-art in various classification tasks, including named entity recognition [1, 2].

SVM classifier can predict where a type of tag (token classes) begins and ends in the text. Classifier is trained from a text that has been annotated. SVM classifier is used to distinguish items of one class against another class based on attributes of training examples. These attributes are called features. The simplest classification problem is to distinguish between positive and negative examples of concepts. Problems in extraction of information is how to

Received: March 8, 2010. Accepted: October 12, 2010