Advances in Electrical Engineering and Informatics

VOLUME II

Pusat Penelitian Teknologi Informasi dan Komunikasi ITB
Sekolah Teknik Elektro dan Informatika ITB

Advances in Electrical Engineering and Informatics

Volume II

ADVANCES IN ELECTRICAL ENGINEERING AND INFORMATICS
VOLUME II

Editor: Armein Z. R. Langi dan Andriyan B. Suksmono
ISBN 978-979-15509-5-6

Diterbitkan pertama kali oleh PPTIK ITB
Gedung PAU Lantai 4
Institut Teknologi Bandung
Jl. Ganeca 10 Bandung

Cetakan pertama: November 2009

Kata Pengantar

Puji syukur kami panjatkan kepada Tuhan Yang Maha Kuasa atas selesainya buku Advances in Electrical Engineering and Informatics volume kedua ini. Buku ini merupakan kumpulan hasil beberapa kegiatan Riset ITB yang dilakukan di lingkungan Sekolah Teknik Elektro dan Informatika (STEI ITB) dan Pusat Penelitian Teknologi Informasi dan Komunikasi (PPTIK ITB) sepanjang tahun 2009.

Selama tahun ini, telah dilaksanakan 42 buah kegiatan riset ITB di lingkungan STEI dan PPTIK, yaitu terdiri dari 2 buah Riset Unggulan, 1 buah Riset Internasional, 10 buah Riset Kelompok Keahlian, dan 29 Riset Hibah Strategis Nasional DIKTI. Setiap kegiatan tersebut merupakan bagian dari suatu kerangka riset yang berkelanjutan. Di samping merupakan indikator kinerja riset, kehadiran buku ini diharapkan mampu memberikan terobosan dalam bidang ilmu teknik elektro dan informatika, serta memperoleh tempat strategis dalam upaya pembangunan Teknologi Informasi dan Komunikasi (TIK).

Buku ini tidak mungkin dapat terwujud tanpa kontribusi begitu banyak pihak dari dalam ITB maupun dari luar ITB. Ucapan terimakasih patut disampaikan kepada para peneliti di lingkungan PPTIK dan STEI, yang namanya tidak dapat disebutkan satu persatu. Terimakasih khusus kepada jajaran pimpinan ITB dan LPPM ITB yang tidak pernah berhenti mendukung kegiatan-kegiatan penelitian menuju universitas berbasis riset. Semoga kumpulan tulisan ini dapat memberi arti dalam upaya memperjuangkan pembangunan masyarakat pengetahuan di Indonesia.

Bandung, November 2009

Armein Z. R. Langi

PPTIK ITB

Daftar Isi

Vegetable Market Information Trend Extracted by SVM-Based Information Extraction Models

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

[1]Computer Engingeering 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 Technolgy
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). In this paper will be discussed SVM-GATE performance in extracting information on Bahasa Indonesia Corpus of Vegetable Market. The experimental results show that SVM-GATE performance is increase as the training sample number growth. The best Performance of SVM-GATE obtained at the τ Margin = 0.5 and the Window Size = 4x4. And the best F-measure on the SVM-GATE on Indonesian corpus of Vegetable Market is 0.67. Performance of SVM-GATE tends to increase as Window Size increased, but the increased performance at Window Size greater than 10 is not significant. There is correlation between big/national events and vegetable price fluctuation.

Keywords: *Information Extraction, Support Vector Machine, Bahasa Indonesia Corpus, NLP, GATE, Market Information Trend*

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 the document collection. Development tool and intelligent methods that can access the content of the document is therefore 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 an approach where the text classification problem is split into tokens and grouped them 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. [3.4]

SVM classifier can predict where a type of tag (token classes) begins and ends in the text. Classifier 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 called features. The simplest classification problem is to distinguish between positive and negative examples of concepts. Problems in extraction of information is how to determine whether the text position is the beginning of a tag (token class) or not and the end of a tag or not.

In this paper will discuss how the performance of the SVM algorithm on extracting information from Indonesian language corpus and will show experimental results in detail. The experiment will see the influence of several parameters on the performance of SVM information extraction. SVM algorithms learning curve will also be evaluated through the experiment the effect of the number of documents examples of F-measures.

2 Related Research

Using SVM based- information extraction systems, Isozaki [4] trains four classifier using sigmoid function to transfer the output of SVM into probabilities and applying Viterbi algorithm to determine the optimal sequence of labels for a sentence. The system is evaluated on the Japanese-language corpus using window size = 2. The results show this system has better performance than systems based on Maximum Entropy and Rule Learning. This system also describes an efficient implementation for the quadratic kernel SVM.

Mayfield [7] applied SVM with lattice-based approach to the cubic kernel for calculating the lattice transition probabilities. By using window size = 3, satisfactory results are obtained like [5].

GATE-SVM system is a variant of the SVM with uneven margins. In the usual SVM, positive and negative examples are treated the same way that margin hyperplane to the negative examples with a margin equal to the negative examples. However, the imbalanced training data where the positive examples is much less, then the SVM is not always appropriate representing the actual positive examples distribution. That why the positive margin greater than the negative margin is a better SVM model. Li [6] introduced the uneven margin parameter in the SVM algorithm. Uneven margin parameter is the ratio of negative margin to a positive margin. By using this parameter, SVM can handle imbalanced data better than normal SVM model.

3 Experiment Method

3.1 SVM Based Information Extraction

Formally, if given a training set $Z = ((x_1, y_1),...,(x_m, y_m))$, which is the input vector n-dimensional, and y_i (= +1 or -1) is the class label, and m is the amount of training data then the SVM with uneven margins are obtained by solving the quadratic optimization problem:

$$\min_{w,b,\xi} \langle w, w \rangle + C \sum_{i=1}^{m} \xi_i$$

$$\text{s.t } \langle w, x_i \rangle + \xi_i + b \geq 1 \quad \text{if } y_i = +1$$

$$\langle w, x_i \rangle - \xi_i + b \leq -\tau \quad \text{if } y_i = -1$$

$$\xi_i \geq 0 \quad \text{for } i = 1, \ldots, m$$

It can be seen from the above equation that there is an addition of parameter τ (tau margin). τ is the ratio of negative-class margin to positive class margin, and will be equal to 1 on the standard SVM. In imbalanced datasets, use a larger margin for the positive class than for the negative class, as can be seen in Figure 1. Therefore, in the SVM with uneven margins the value of τ is $0 < \tau < 1$.

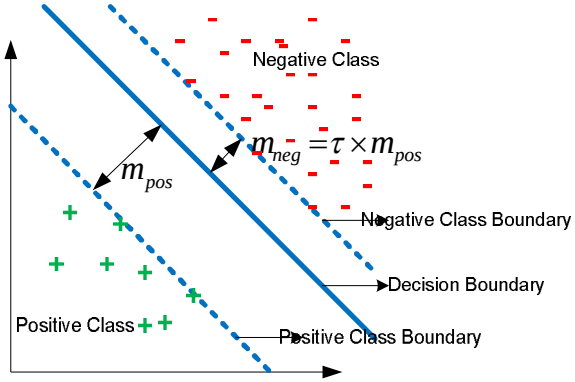

Figure 1 Illustration SVM with uneven margins.

The implementation of SVM for extracting information in this paper is using GATE (General Architecture for Texts Engineering) toolkit [3]. Machine learning process in GATE is based on SVM Light Wrapper [2]. Step-by-step process shown in Figure 3 and Figure 4.

Figure 3 Figure 3 explains step by step classifier training process for SVM. First corpus is stored in a GATE format and Annotated in accordance with the token type (e.g. *komoditi*) and this document is used as an intake to build SVM model (which is needed <*komoditi*> as start tag and </*komoditi*> as end tag on the object text / token in question). SVM models are generated and stored in an external file for later use.

In this experiment to build the features vector of tokens, several NLP features that are used: (1) Case/Orthography, namely the use of uppercase and lowercase letters by the token. (2) types of tokens: words, numbers, symbols, or punctuation and (3) entity, the output module of *named entity recognition* standards owned by GATE. The window size is the number of tokens before and after the target token. It is also used as an input for SVM.

To extract information from a new document, the system requires the SVM model produced in the learning process. SVM Wrapper then will annotated text with the initial tag and the end tag that match existing models. In the next stage, the start and the end tags are combined in an appropriate token (Figure 4).

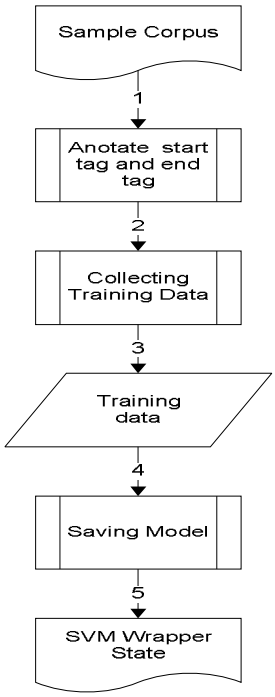

Figure 3 Flow chart showing the process of SVM training data in GATE.

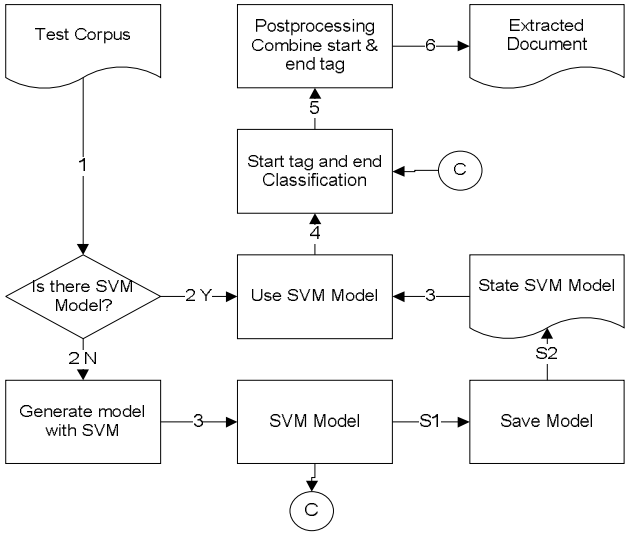

Figure 4 Flow chart showing the process of SVM training data in GATE.

To measure the performance of the algorithm in extracting the information, Precision, Recall and F-Measure are used.

$$precision = \frac{correct}{correct + falsePositive} \quad (1)$$

$$recall = \frac{correct}{correct + falseNegative} \quad (2)$$

$$F - measure = \frac{2 \times precision \times recall}{precision + recall} \quad (3)$$

Where is *correct* is the number of slots that are filled and true, *falsePositive* is the number of slots filled but wrong, and *falseNegative* is the number of slots that are not filled.

Partial correct is the prediction of extraction that is only partially true. In the case of formula for precision, recall, and F-measure (F-1), there are 3 approaches: (1) *Strict*, *partial correct* is consider wrong for both *falsePositive* and *falseNegative*, (2) *Lenient*, *partial correct* is considered true, and (3) *Average*, *partial correct* is weighted ½. In this paper an approach will be used (1) and (2).

To get better results, the experiment is repeated ten times (10-fold cross validation) for each variation of intake of the tested models. The model used by the SVM kernel is SVM linear with *reciprocal weighting*. While the chosen classification technique is *one-against-all* [2].

3.2 Dataset

Research dataset used is the corpus of Vegetable Market obtained from crawling the internet. Dataset consists of 210 documents (web pages) in Indonesian language that contains news about the change in vegetable prices in cities in Indonesia. For the purposes of learning and testing, each token corresponding to these documents manually annotated with eight classes using labels (slot) as follows:

- *Tanggal* : The date when the news was written on a web page
- *Lokasi*: Place of occurrence
- *Komoditi*: Commodity Type of vegetables
- *Harga sebelum*: vegetable commodity prices before the price changes
- *Harga terkini*: vegetable commodity prices after the price change (current)
- *Satuan*: Units that are used in trading
- *Perubahan harga*: commodity price fluctuations
- *Event*: Event-related causes commodity price changes.

Vegetable corpus statistics are presented in Table 1, while the sample web page that has been Annotated shown in Figure 5.

Table 1 Vegetable corpus statistics that are used as a dataset

Slot	frequency	Example
Tanggal	282	12 November 2008, 10/12/2009
Lokasi	341	Pasar Induk Keramat Jati, Kabupaten Bandung
Event	182	Banjir, kemarau
harga_sebelum	561	5000
harga_terkini	1427	6000
komoditi	1409	Kacang panjang, mentimun, kol
perubahan_harga	106	Naik seribu rupiah
satuan	1088	Kg, Ikat, Liter

The total number of tokens in the dataset is 390,226. Only 5396 (1.4%) of the token is positive example. The ratio of positive and negative data in the dataset shown in Figure 6.

MAGELANG, SELASA - Harga sayur mayur di **<lokasi>**Kabupaten Magelang**</lokasi>,** kini turun secara signifikan. Pada berbagai jenis sayuran, penurunan harga terjadi bervariasi, mulai dari Rp 500 per kilogram (kg), hingga Rp 1.500 per kg.

Sumartini, salah seorang pedagang sayur di Pasar Muntilan, mengatakan, **<komoditi>**kacang panjang **</komoditi>** misalnya mengalami penurunan harga dari Rp **<harga_Sebelum>**3.500**</harga_sebelum>** per **<satuan>**kg**</satuan>.** menjadi Rp **<harga_terkini>**2.500**</harga_terkini>** per **<satuan>**kg**</satuan>.** Begitupun, harga **<komoditi>**seledri **</komoditi>** yang semula Rp **<harga_Sebelum>**2.500**</harga_Sebelum>** per **<satuan>**kg**</satuan>** sekarang menjadi Rp **<harga_terkini>**1.500 **</harga_terkini>** per **<satuan>**kg**</satuan>.**. Untuk **<komoditi>**tomat**</komoditi>** dan **<komoditi>**wortel**</komoditi>,** masing-masing turun harga Rp **<harga_Sebelum>**500**</harga_Sebelum>** per **<satuan>**kg**</satuan>** menjadi Rp **<harga_terkini>**1.500 **</harga_terkini>** per **<satuan>**kg**</satuan>** dan Rp **<harga_terkini>**1.000**</harga_terkini>** per **<satuan>**kg**</satuan>**.

Menurutnya, kondisi ini dimungkinkan terjadi karena melimpahnya persediaan sayur di **<event>**musim panen **</event>**. Namun, karena pasar sedang sepi, saya pun tetap membatasi

Figure 5 An Example of documents that have been annotated for the SVM machine learning.

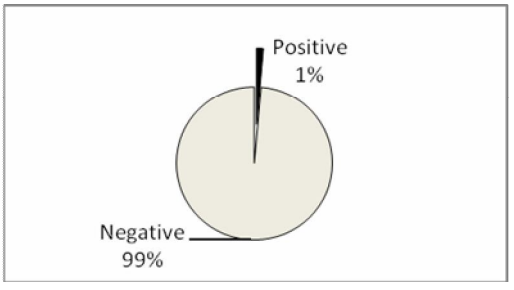

Figure 6 The ratio of positive and negative data in the dataset (Vegetable Market Corpus).

4 Result And Discussion

The relationship of the number of sample documents for learning and its performance of SVM-GATE on some τ margin and Window Size can be seen in Figure 7, 8 and 9.

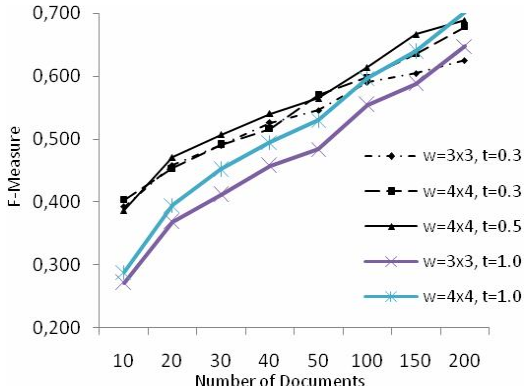

Figure 7 The relationship between the number of sample documents for learning and F-Measure on several composition of window size (w) and t (τ margin) using SVM GATE.

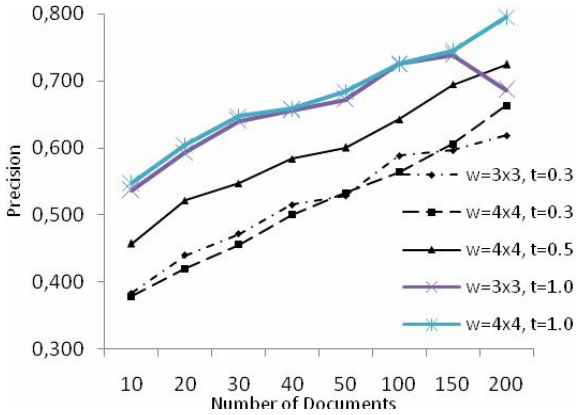

Figure 8 The relationship between the number of sample documents for learning and Precision on several composition of window size (w) and t (τ margin) using SVM GATE.

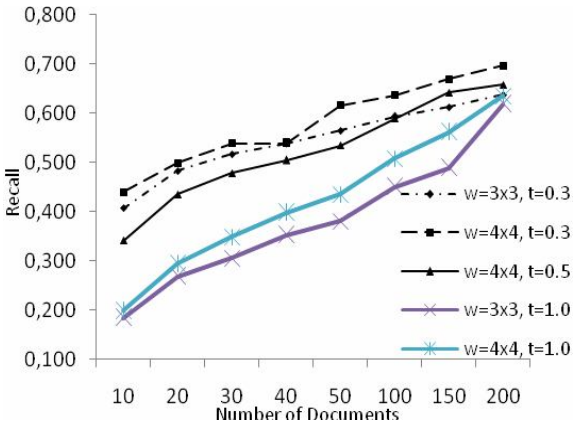

Figure 9 The relationship between the number of sample documents for learning and Recall on several composition of window size (w) and t (τ margin) using SVM GATE.

Figure 7, 8, and 9 show that with the increasing number of documents samples for learning; Precision, Recall and F-Measure are increasing as well. These results are also similar to that result obtained by Li [5] on the Job Corpus for small samples. Increasing the number of sample documents cause classifier quality is getting better so that their performance in the extraction of information also become better.

In general, the best F-measure and precision is obtained at window size = 4 and the τ margin of 0.5 followed by a window size = 3 and the τ margin of 0.3, while the best recall obtained at window size = 4 and the τ margin = 0, 3.

In the experiment the influence of τ Margin, Precision seen an increase in line with the increase in τ margin. Instead, Recall decline with increasing τ Margin. The highest F-Measure is obtained on τ Margin = 0.5. This result is different from what Li [5] get where the best value of τ margin is 0.4 for *job corpus.* This result is also different from that obtained by Paramita [8] who get the best performance of SVM GATE at τ Margin = 0.6 in the Indonesian and English mixed corpus. The difference is expected due to the differences in the distribution of positive and negative examples that differ between the corpus in [8] and that we use.

In the experiment the influence of Window Size, Window Size = 4 showed the best performance. This result is different from the results of [5] and [8] which recommends Window Size = 3.

Based on the above experiment, the highest F-measure is achieved at composition of Window Size = 4 and τ Margin = 0.5.

The influence of window size on the performance of SVM to extract the information shown in Figure 10.

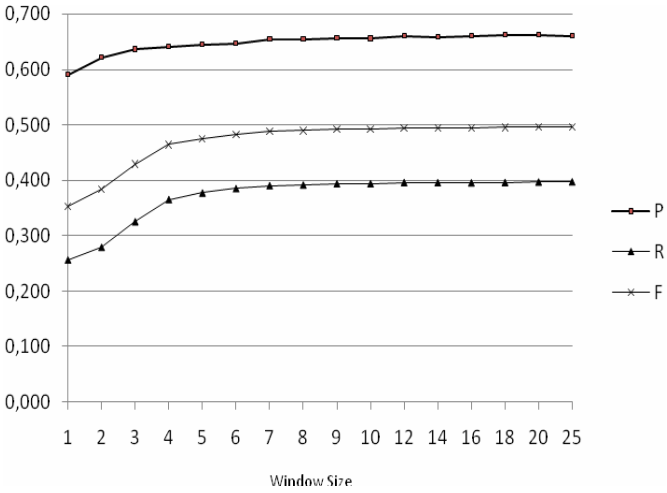

Figure 10 Window Size influences the performance of SVM in extracting information.

From Figure 10, it can be seen that the performance of SVM to extract information increases with the increase of Window Size. However, increasing getting less and less and tend to be not significant at Window Size greater than 10.

Vegetable market information trend extracted from corpus is shown in Table 2. From that table it shows that there are relations between events that occur at certain time and vegetable price fluctuation, for example the increase of fuel price is correlated with the increase of vegetable price on the market. Big national event such as new year and idiil fitr festival are also significantly correlated with vegetable price fluctuation.

Table 2 Vegetable market information trend extracted from the corpus (relation between field *event* and field *perubahan harga*)

Event	*Perubahan Harga*
Panen raya	Kenaikannya mencapai 100 persen
Maulid Nabi Muhammad SAW	Kenaikannya mencapai 100 persen
Kenaikan harga bahan bakar minyak	naik rata-rata Rp.50 hingga Rp.1000/kg
Natal dan Tahun Baru	naik mencapai 100 persen
Virus avian influenza	penurunan 20 sen
Kenaikan gaji	meningkat 50 sen
Musim penghujan	kenaikan Rp2 ribu/kg
Terhambatnya distribusi	naik Rp600
Usai Lebaran	naik 15,38 persen
Pemerintah menurunkan harga BBM	penurunan 10-30 persen

5 Conclusion And Suggestions

5.1 Conclusion

From the above experiment, it can be summarized as follows:

1 The best F-measure on the SVM-GATE on Indonesian corpus of Vegetable Market is 0.67.
2 The more the number of training samples, the better SVM-GATE performance.
3 The best Performance of SVM-GATE obtained at the τ Margin = 0.5 and the Window Size = 4x4.
4 Performance of SVM-GATE tends to increase as *Window Size* increased, but the increased performance at Window Size greater than 10 is not significant.
5 There is correlation between big/national events and vegetable price fluctuation.

5.2 Suggestions

1 To further improve the performance of SVM-GATE, NLP enrichment is needed such as the use of Part of Speech Tagger for Indonesian.
2 It should be further investigated whether the performance of SVM-GATE continues to increase with increasing the number of sample documents, and how the relationship between *Window Size* and τ Margin to the corpus statistics.

6 References

1. Bouckaer, R.R. 2002. Low level information extraction. A Bayesian network based approach. In *Proc. TextML* 2002.
2. Cunningham, Hamish et al., 2007. *Developing Language Processing Components with GATE Version 4 (a User Guide)*. The University of Sheffield 2001-2007. URL: http://gate.ac.uk/sale/tao/
3. Cunningham, Hamish, D. Maynard, K. Bontcheva, V. Tablan. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics* (ACL'02). URL: http://gate.ac.uk/sale/acl02/acl-main.pdf.
4. Isozaki, H., Kazawa, H.: Efficient Support Vector Classifiers for Named Entity Recognition. In *Proceedings of the 19th International Conference on Computational Linguistics* (COLING'02), pages 390{396, Taipei, Taiwan, 2002.
5. Li, Yaoyong, Kalina Bontcheva, and Hamish Cunningham. 2005. SVM Based Learning System For Information Extraction. Sheffield Machine Learning Workshop, *Lecture Notes in Computer Science*, Springer Verlag
6. Li, Y., Shawe-Taylor, J.: The SVM with uneven margins and Chinese document categorization. In Proceedings of The 17th Pacific Asia Conference on Language, Information and Computation (PACLIC17), pages 216{227, Singapore, Oct. 2003.
7. Mayfeld, J., McNamee, P., Piatko, C.: Named entity recognition using hundreds of thousands of features. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003*, pages 184{187. Edmonton, Canada, 2003.
8. Paramita. 2008. *Penerapan Support Vector Machine untuk Ekstraksi Informasi dari Dokumen Teks*. Tugas Akhir Program Studi Teknik Informatika STEI Institiut Teknologi Bandung.