The 5th ICIBA 2016, International Conference on
Information Technology and Engineering Application
Palembang-Indonesia, 19-20 February 2016

Vector Machine Learning Method for Text Mining Indonesian Social Media Named Entity Recognition

Agus Suryana, Sri Ipnuwati

Department of Information Science, STMIK Pringsewu Lampung
Jl. Wismarini No. 09 Pringsewu Lampung, Indonesia
e-mail: suryani64@yahoo.co.id

Abstract

Social media named entity recognition (SMNER) is one of the important tasks in social media information extraction, which involves the identification and classification of words or sequences of words denoting a concept or entity. With the extension of named entity recognition to new information areas, the task of identifying meaningful entities has become more complex as categories are more specific to a given domain. SMNER solutions that achieve a high level of accuracy in some language or domain may perform much poorly in a different context. Support Vector Machine (SVM) is rapidly emerging as a promising pattern recognition methodology due to its generalization capability and its ability to handle high-dimensional input. However, SVM is known to suffer from slow training especially with large input data size. In this paper, we explore the scalability issues for Indonesian social media named entity recognition using high-dimensional features and support vector machines.

Keywords : *Support Vector Machine, Machine Learning, Text Mining, Indonesian Social Media, Named Entity Recognition*

1 INTRODUCTION

Social development of the Indonesian media is very influential on the various joints public life of the nation Indonesia. Based on the classification of the influence of social media can be seen from all sides of life ideologies, social, cultural, political [1, 2], defense and security of the state (IPOLEKSUBUDHANKAM) [3].

Social media named entity recognition (SMNER) is one of the important tasks in Indonesian social media information extraction, which involves the identification and classification of words or sequences of words denoting a concept or entity. Examples of such information units are names of persons, organizations, or locations in the general context of social media, and the names of facebook and genes in the social media context. With the extension of named entity recognition to new information areas, the task of identifying meaningful entities has become more complex as categories are more specific to a given domain. SMNER solutions that achieve a high level of accuracy in some language or domain may perform much poorly in a different context [4].

Figure 1: Perspective Social Media Indonesia

From a technical perspective, social media analytics research faces several unique challenges. First, social media contains an enriched set of data or metadata, which have not been treated systematically in data- and text-mining literature. Examples include tags (annotations or labels using free-form keywords); user-expressed subjective opinions, insights, evaluation, and perspectives; ratings; user profiles; and both explicit and implicit social networks. Second, social media applications are a prominent example of human-centered computing with their own unique emphasis on social interactions among users. Hence, issues such as context-dependent user profiling and needs elicitation as well as various kinds of human computer interaction considerations must be reexamined. Third, although social media promises a new approach to tackling the noise and information-overload problem with Web-based information processing, issues such as semantic inconsistency, conflicting evidence, lack of structure, inaccuracies, and difficulty in integrating different kinds of signals abound in social media. Fourth, social media data are dynamic streams, with their volume rapidly increasing. The dynamic nature of such data and their sheer size pose significant challenges to computing in general and to semantic computing in particular [5].

Different approaches are used for carry out the identification and classification of entities. Statistical, probabilistic, rule-based, memory-based, and machine learning methods are developed. The extension of SMNER to specialized domains raise the importance of devising solutions that require less human intervention in the annotation of examples or the development of specific rules. Machine learning techniques are therefore experiencing an increased adoption and much research activity is taking place in order to make such solutions more feasible. Support Vector Machine (SVM) is rapidly emerging as a promising pattern recognition methodology due to its generalization capability and its ability to handle high-dimensional input. However, SVM is known to suffer from slow training especially with large input data size.

In this paper, we explore the scalability issues for Indonesian social media named entity recognition using high-dimensional features and support vector machines. We present the results of experiments using large Indonesia social media datasets and propose a plan to improve SVM scalability using new database-supported algorithms.

2 RESEARCH METHODOLOGY

2.1 Social Media Machine Learning Architecture

Constructing a social media named entity recognition solution using a machine learning approach requires many computational steps including preprocessing, learning, classification, and post-processing [6]. The specific components included in a given solution vary but they may be viewed as making part of the following groups summarized in Figure 2.

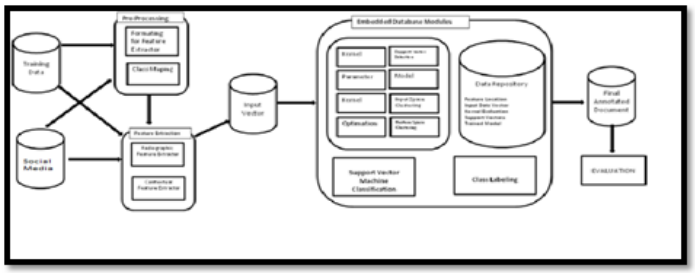

Figure 2: Social Media Machine Learning Architecture

2.2 Preprocessing Modules

Using a supervised machine learning technique relies on the existence of annotated training data. Such data is usually created manually by humans or experts in the relevant field. The training data needs to be put in a format that is suitable to the solution of choice. New data to be classified also requires the same formatting. Depending on the needs of the solution, the textual data may need to be tokenized, normalized, scaled, mapped to numeric classes, prior to being fed to a feature extraction module. To reduce the training time with large training data, some techniques such as chunking or instance pruning (filtering) may need to be applied.

2.3 Feature Extraction

In the feature extraction phase, training and new data is processed by one or more pieces of software in order to extract the descriptive information about it. The choice of feature extraction modules depends on the solution design and may include the extraction of orthographic and morphological features, contextual information about how tokens appear in the documents, linguistic information such as part-of-speech or syntactic indicators, and domain-specific knowledge such as the inclusion of specialized dictionaries or gazetteers (reference lists). Some types of information may require the use of other machine learning steps to generate it, for example, part-of-speech tagging is usually performed by a separate machine learning and classification software which may or may not exist for a particular language.

Preparing the data for use by the feature extractor may require special formatting to suit the input format of the software. Also, depending on the choice of machine learning software, one may need to reformat the output of the feature extraction to be compatible with whats expected by the machine learning module(s). Due to the lack of standardization in this area

and because no integrated solutions exist for named entity recognition, several compatibilities exist between the many tools one may use to build the overall architecture. In addition, one may also need to build customized interfacing modules to fit all the pieces of the solution together.

2.4 Learning and Classification

Most of the publicly available machine learning software use a two-phased approach where learning is first performed to generate a trained machine followed by a classification step. The trained model for a given problem can be reused for many classifications as long as there is no need to change the learning parameters or the training data.

2.5 Post-Processing Modules

The post-processing phase prepares the classified output for use by other applications and/or for evaluation. The classified output may need to be reformatted, regrouped into one large chunk if the input data was broken down into smaller pieces prior to being processed, remapped to reflect the string class names, and tested for accuracy by evaluation tools. The final collection of annotated documents may be reviewed by human experts prior to being used for other needs.

Social media application method named entity recognition for text mining interpretation and extracting information from social media web can use the web extract text mining methode to build domain models and rules of social media interpretation of named entity recognition dictionary. Extended the application of social media named entity recognition in the new domain lead to more adoption of supervised machine learning techniques that include Support Vector Machine (SVM) [7, 8].

With the growing adoption of machine learning techniques for SMNER, especially for specialized domains, the need for developing semi-supervised or unsupervised solutions. Supervised learning methods rely on the existence of manually annotated training data, which is very expensive in terms of labor and time and a hindering factor for many complex domains with growing nomenclature. However, using unannotated training data or a mixture of labeled and unlabeld data requires the development of new SMNER machine learning solutions based on clustering and inference techniques.

2.6 Multi-class Support Vector Classification

For classification problems with multiple classes, different approaches are developed in order to decide whether a given data point belongs to one of the classes or not. The most common approaches are those that combine several binary classifiers and use a voting technique to make the final classification decision. These include: One-Against-All, One-Against-One, Directed Acyclic Graph (DAG), and Half-against-half method [9, 10]. A more complex approach is one that attempts to build one Support Vector Machine that separates all classes at the same time. In this section we will briefly introduce these multi-class SVM approaches. Figure 4 compares the decision boundaries for three classes using a One-Against-All SVM, a One-Against-One SVM, and an All-Together SVM. The interpretation of these decision boundaries will be discussed as we define the training and classification techniques using each approach.

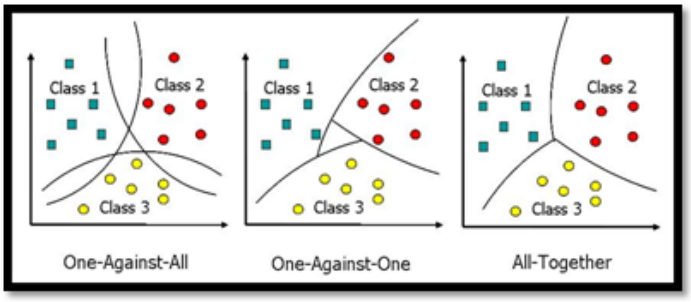

Figure 3: Comparison of Multi-Class Boundaries

3 SOCIAL MEDIA MULTI-CLASS SUPPORT VECTOR MACHINE RESULTS

The multi-class performance results are summarized in table 3.1.. The overall recall measure achieved is ideologies (65.27%), social (56.19%, cultural (55.56%), political (61.99%), economic (48.15%) and Defence and Security (63.03%). These results are compared to those obtained by the social media participating systems which used support vector machines either in isolation or in combination with other models using the same task data. The performance comparison results are reported in table 1. The language-independent approach used in this experiment performed very close than which both used Support Vector Machine as the only learning model. In this experiment used character n-grams, orthographic information, word shapes, gene sequences prior knowledge, word variations, part-of-speech tags, noun phrase tags, and word triggers.

Table 1: Multi-Class SVM Results 1998-2001 Set

Named entity	**Complete match**	**Right boundary match**	**Left boundary match**
Ideologies (3186)	2271 (71.28 / 60.19 / 65.27)	2666 (83.68 / 70.66 / 76.62)	2526 (79.28 / 66.95 / 72.60)
Social (588)	277 (47.11 / 69.60 / 56.19)	386 (65.65 / 96.98 / 78.30)	312 (53.06 / 78.39 / 63.29)
Cultural (70)	35 (50.00 / 62.50 / 55.56)	53 (75.71 / 94.64 / 84.13)	36 (51.43 / 64.29 / 57.14)
Political (1138)	570 (50.09 / 81.31 / 61.99)	802 (70.47 / 114.41 / 87.22)	612 (53.78 / 87.30 / 66.56)

4 CONCLUSIONS

The historical development of support vector machine learning methode and its applications in Indonesian social media shows that from simple and straigt forward to use algorithms, systems and methodology have emerged that enable advanced and sophisticated data analysis. In the future, social media intelligent data analysis will play even a more important role, due to the huge amount of information produced and stored by modern technology.

Current machine learning algorithms provide tools that can signicantly help social media practitioners to reveal interesting relationships in their data. Our experiments show that in social media domains various classiers perform roughly the same. So one of the important

factors when choosing which classifier to apply is its explanation ability.

References

[1] L. A. Abdillah, "Indonesian's presidential social media campaigns," in *Seminar Nasional Sistem Informasi Indonesia (SESINDO2014)*, ITS, Surabaya, 2014.

[2] L. A. Abdillah, "Social media as political party campaign in Indonesia," *Jurnal Ilmiah MATRIK*, vol. *16*, pp. 1-10, 2014.

[3] A. Suryana and S. Ipnuwati, "Perspective Text Mining Analytic Intelegent Information Extraction for Impect of Indonesian Social Media," in *Prosiding International conference on Information Technology and Business (ICITB) 2015*, 2015, pp. 100-113.

[4] A. Gattani, et al., "Entity extraction, linking, classification, and tagging for social media: a wikipedia-based approach," *Proceedings of the VLDB Endowment*, vol. *6*, pp. 1126-1137, 2013.

[5] D. Zeng, et al., "Social media analytics and intelligence," *Intelligent Systems, IEEE*, vol. *25*, pp. 13-16, 2010.

[6] S. Rping, Support vector machines in relational databases: *Springer*, 2002.

[8] T. Solorio, "Improvement of Named Entity Tagging by Machine Learning," Puebla, Mxico: *Coordinacin de Ciencias Computacionales, Instituto Nacional de Astrofsica*, 2004.

[8] T. Solorio and A. L. Lopez, "Adapting a Named Entity Recognition System for Spanish to Portuguese," in *Workshops on Artificial Intelligence: Workshop on Herramientas y Recursos Lingsticos para el Espaol y el Portugus*, Nov, at Puebla, Mexico, 2004, pp. 292-297.

[9] S. Abe, Support vector machines for pattern classification vol. 2: *Springer*, 2005.

[10] S.-H. Lee, Selection of gaussian kernel widths and fast cluster labeling for support vector clustering: University of Massachusetts Lowell, 2005.

[]