Automatic Extraction of Place Entities and Sentences Containing the Date and Number of Victims of Tropical Disease Incidence from the Web

Taufik Fuadi Abidin
Department of Informatics, College of Science, Syiah Kuala University
Banda Aceh, Aceh, 23111, Indonesia
Email: tfa@informatika.unsyiah.ac.id

Ridha Ferdhiana[1)] and Hajjul Kamil[2)]
[1)] Department of Statistics, College of Science, Syiah Kuala University
[2)] Department of Nursing, College of Medical, Syiah Kuala University
Banda Aceh, Aceh, 23111, Indonesia
Email: {ridha.ferdhiana, hajjul.kamil}@unsyiah.ac.id

***Abstract*—Many tropical disease incidences, such as leprosy, elephantiasis, malaria, dengue fever, are reported in online news portals. Online news portals are valuable data sources for creating a tropical disease repository if the information such as the location of the incidence, date of occurrence, and the number of victims can be automatically extracted from news articles. This paper describes approaches to extract that information from the Web. We introduce a rule-based algorithm to identify and extract the locations of the incidence and use Support Vector Machine (SVM) to determine the sentences containing the date of occurrence and the number of victims. Our experiments show that, the accuracy of the rule-based algorithm to identify the location entities is 99.8%, while the accuracy of the classifier to determine the sentences that contain one or more places of the incidence is 82%. The accuracy of SVM classifiers to classify the sentences that contain the date of occurrence and the number of victims are 96.41% and 93.38%, respectively.**

***Index Terms*—Entity Extraction from the Web, Support Vector Machine, Classification**

I. Introduction

Many tropical disease cases in Indonesia such as lymphatic filariasis, dengue fever, leprosy and malaria are reported every year. Between 2000 and 2009, a total of 11,914 chronic lymphatic filariasis cases have been reported nationally [1]. More than 17 provinces are reported to have malaria transmission with average transmission rate across the country around 5:1,000 population per annum [2]. Kompas online, one of the national Indonesian online newspapers headquartered in Jakarta, wrote that about 14,016 people were infected by *mycobacterium leprae* in 2001 and increased to 19,695 people (40.52%) in 2005 [3].

Due to the proliferation of internet technology, a large number of online news portals report the tropical disease incidence in Indonesia. If a keyword *kasus demam berdarah* (*dengue fever cases*) is searched on google.co.id, about 1,120,000 relevant results were returned.

Web pages, written in hypertext format and in a loosely structured text, are great sources of information in the modern age of internet-based technology today. Anyone can create web pages, and therefore, the size of the Web continues to grow. According to worldwidewebsize.com, the total number of web pages indexed by Google in June 2013 has reached approximately 47 billion pages [4]. A large number of web pages are being added to the Web every day, and thus, classifying web pages into interesting categories is an essential step and it is often treated as an initial step of mining the Web [5].

Classifying a large numbers of web pages into interesting classes is the goal of web classification. Web classification has been studied extensively and many research works in this field have been done, such as classifying web pages without negative examples which eliminates the requirement to manually collect negative training samples that tends to be biased [5], evaluating the capabilities of Bayesian algorithm for web classification and comparing its performance for both binary and multi-classification [6], surveying prominent web page classification methods [7], building SVM web classifiers and selecting web features [8], and learning to classify tropical disease web pages in a large Indonesian web documents [9].

To the best of our knowledge, this paper represents the first attempt to automatically recognize the locations where the tropical disease outbreaks occurred from Indonesian web pages and to identify the sentences that contain the occurrence date and the number of victims. We introduce a rule-based algorithm that incorporates morphological and contextual components as listed in Table 1 and a database of places [10] to recognize the location entities in the sentences, and then, use SVM classifier to classify the sentences and to determine which of those sentences contain the places where the tropical disease incidence occurred. We also build SVM models to identify the sentences that have occurrence date or the number of victims, or both. Previously classified web pages, described in [9], were used as the data source. A

doi:10.4304/jetwi.5.3.302-309

large number of sentences in the web pages were observed and manually annotated. The sentences that have the occurrence date or the number of victims, or both were labeled {+1}, and those that have no occurrence information or the number of victims were labeled {-1}. We took a portion of those labeled datasets for the training set to build SVM classifier and took the rest of the portion for the testing set. In summary, our contributions are twofold:

1. We introduced a rule-based approach to identify the place entities in the sentences and built an SVM classifier [11] to identify which of the sentences contain the place entities of tropical disease incidence.
2. We built SVM classifiers and selected the best classifiers to determine the sentences that contain the occurrence date and to identify the number of victims.
3. We organized the extracted entities and sentences into Keyhole Markup Language (KML) format and integrated them into Google Earth application.

The paper is organized as follows: Section 2 discusses related work. Section 3 describes the proposed approaches, including the contextual and morphological components, the methodology to remove the conflicting words in dictionaries, the construction of features, and the evaluation metrics to measure the accuracy of the SVM classifiers. Section 4 reports the results, and finally, Section 5 concludes our discussion of the automatic extraction of place entities and targeted sentences.

II. Related Work

Research on named entity recognition (NER) that aims at recognizing person, place, organization, time, and numerical expressions from text corpus has become an interesting study since the last two decades. Zhao [12] proposed a Hierarchical Hidden Markov Model to automatically identify product named entity in Chinese text. The entity was constructed using word forms and part-of-speech (POS) features. The findings concluded that the proposed methods outperformed the cascaded maximum entropy model and worked well for electronic and cell phone products. Sari et al. [13] used part-of-speech (POS) and syntactical structure, combined with semi-supervised learning method, to recognize and categorize named entity. They used Natural Language Processing (NLP) software to produce syntactic structure and used Stanford tagger to get POS tags of the sentences. They introduced a new method to automatically extract the date and location patterns from the sentences which have been labeled as prepositional phrase and from the sentences which have not been labeled by the tagger as prepositional phrase. They claimed that the performance of their proposed NER system is in the range of 50-70%.

Chanlekha [14] proposed a methodology to recover the most specific location where the outbreak of infectious disease occurred. They incorporated various features for recognizing spatial attributes into the models and trained the models using machine learning techniques such as Conditional Random Fields (CFP), SVM, and Decision Tree. In that work, Chanlekha considered events as the expression of phrases or grammatical constituents. The drawback of Chanlekha's approach is that the expression of phrases must be entirely defined to ensure that all events reported in the news articles can be recovered.

While research on named entity recognition of location has been intensively studied in English domain, far less attention has been paid to NER of location in Indonesian domain. This paper represents the first attempt to automatically recognize the locations where the tropical disease outbreaks occurred and to identify the sentences that contain the occurrence date and the number of victims.

III. Proposed Approaches

We propose the following approaches: 1) A rule-based approach to identify whether place entities are found in sentences by incorporating contextual and morphological components, and a database of places. The sentences that contain place entities, then, are classified using SVM classifier to ensure that the place entities are the locations where the tropical disease occurred. If the classifier categorizes a sentence as {+1}, then the place entities are extracted; 2) Develop SVM models to categorize sentences containing the date and the number of victims of tropical disease incidence; and 3) Organize the extracted entities into KML to integrate them into Google Earth application. We will discuss the proposed approaches in the following sections.

A. Contextual and Morphological Components

Contextual is a reference component that forms a place entity or negates it. In a sentence, contextual component is commonly written adjacent to a place entity that can be a single-word term, a two-word term, or a three-word term positioned consecutively.

Morphology is a major component in the grammar and it is primarily concerned with the rules of the word formation [15]. For place entities, the words are formally written in title case or uppercase, e.g. Bali or BALI. For date entities, they are usually written as a combination of digits and strings in specific formats such as *dd/dd/dddd*, *dd-dd-dddd*, *dd/dd/dd*, *dd-dd-dd*, *d name-of-month dddd*, and several other forms. The morphology for the number of victims is formally written as a combination of digits and contextual words, such as the word *korban (victim)* or *meninggal (dead)*.

Contextual components such as location prefix (LPRE) [16], popular town (PT), sign of location (SILO), preposition followed by a location (LOPP), and sign of address (SIAD) help us identify a place entity in a sentence, whereas location leader (LLDR) assists us that after the LLDR, the following phrase must not be a place entity. It is a place where the leader leads, instead. Table 1 lists the contextual components for place entity. Let's

discuss a few examples:

Wabah malaria terjadi di Kota Jakarta Utara

(Malaria outbreaks in the City of North Jakarta)

The word *Kota (City)* in that sentence is a location prefix, labeled as LPRE in Table 1. LPRE is a contextual component that gives us a clue that the next adjacent words, *Jakarta Utara*, written in title case, positioned consecutively, found in the database of zip codes, and morphologically true for a location is a place entity.

TABLE 1
CONTEXTUAL COMPONENTS FOR PLACE ENTITY

Label	Description	Examples
LPRE	Location prefix	Kota (city), desa (village), wilayah (region), …
LLDR	Location leader	Gubernur (governor), walikota (mayor), …
GOAG	Government agency	Polda (police), pemda (state government), …
LOGA	Leader of a government agency	Kapolda (chief of a state police), kepala (head of a unit), …
PT	Popular town	Jakarta, Denpasar, Surabaya, Banda Aceh, ...
LOPP	Preposition followed by a location	Di (at), dari (from), …
SILO	Sign of location	Lokasi (location), kawasan (region), …
SIAD	Sign of address	Jl, jln, jalan (street), …
PEOP	Public place	Hotel, taman (park), gedung (building), …
RELO	Religious location	Mesjid (mosque), wihara (temple), …
DAY	Name of day	Senin (Monday), Selasa (Tuesday), …
MONTH	Name of month	Januari (January), Maret (March), …
OPRE	Organization prefix	Universitas (university), institut (institute), ...
OPOS	Position in an organization	Direktur (director), rektor (rector), …
APRO	Abbreviation of a province	Sumut (North Sumatera), Jatim (East Java), …

Gubernur Aceh memberikan bantuan kepada para korban deman berdarah

(Governor of Aceh provides assistance to the victims of dengue fever)

The word *Gubernur* in that sentence is a location leader, labeled as LLDR in Table 1. LLDR is a contextual component that gives us a hint that the next word, *Aceh*, should not be considered as a place entity even though morphologically the first letter of the word is in uppercase and the word is found in the database of places. The word *Aceh* after *Gubernur* in that sentence is actually the name of the province where the governor governs. The two examples discussed here illustrate the roles of contextual and morphological components in assisting the rule-based algorithm to identify the place entities in a sentence.

B. Removing Conflicting Words in Dictionaries

One of important steps in our proposed approach is to construct bag-of-words (*dictionaries*) for each class. We used three different datasets in this research, i.e. place of incidence dataset, date of occurrence dataset, and the number of victim dataset as listed in Table 2. The dictionaries consist of weighted one-gram, bi-gram, and three-gram words extracted from the sentences in class {+1} and {-1}. The dictionaries are used to construct numerical features of each sentence.

To avoid over fitting, which generally gives poor predictive performance, the dictionaries consist of n-gram words were only extracted from the sentences in the training sets. We discovered that many similar n-gram words (one-gram, bi-gram, or three-gram) are found in the dictionary of class {+1} and {-1}, for instance, the word *penyakit (disease)*. The weight of the word in the dictionary of class {+1} is 0.563 and weight of the same word in the dictionary of class {-1} is 0.538. The weights are normalized by dividing the frequency of the word over the maximum frequency of the word in the dictionary.

To remove the words with high or low weight in both dictionaries, the weight ratio is calculated by taking the larger weight as the denominator. If the ratio is greater than a given threshold, 0.5 for this work, then the word with a smaller weight is removed from the dictionary. However, if the ratio is smaller than a given threshold, then the words will be removed from both dictionaries. For the word *penyakit (disease)*, it was removed from both dictionaries, i.e. the dictionaries of class {+1} and {-1}, because the ratio is 0.955 (0.538/0.563), which is greater than a given threshold.

Another example is the bi-gram words *wabah lepra (leprosy outbreak)*. The weights in {+1} and {-1} class dictionaries are 0.023 and 0.003 respectively. Hence, by taking the larger weight as the denominator, the ratio is 0.13. Because the ratio is smaller than 0.5, then the bi-gram words *wabah lepra (leprosy outbreak)* will be removed from the dictionary of class {-1} only, i.e. the class in which the ratio is smaller, while for the dictionary of class {+1}, the bi-gram words are retained.

The process of removing conflicting words in the dictionaries is also done for the date of occurrence and number of victim datasets. The words in the dictionaries play an important role in constructing numerical features of a sentence and achieving good classification accuracy.

C. Constructing Numerical Features for SVM Model

The numerical features of a sentence are the ratio of 1-gram, 2-gram, and 3-gram words in that sentence. There are 3 features for each class, and therefore, a total

of 6 features will be constructed for each sentence. The dictionaries that have been created beforehand are used to construct the numerical features of all sentences.

Mathematically, the ratio of *k*-gram words in class *p*, denoted as $F_{k\text{-}gram,\ p}$, is equal to the number of *k*-gram words in the sentence that are found in the dictionary of class C_p, denoted as $Dic(C_p)$, divided by the total number of *k*-gram words in the sentence.

$$F_{k\text{-}gram,p} = \frac{\sum_{i=1}^{n} t_{k\text{-}gram,i} \in Dic(C_p)}{\sum_{i=1}^{n} t_{k\text{-}gram,i}} \quad (1)$$

where i=1,2,...,n; k=1,2,3; p=1,2 and n is the number of words in the sentence.

In this work, SVM is used as the classification method. SVM has shown high performance in solving classification problems [11], [17]. Its performance results usually outperform other classifiers [9]. SVM is basically a linear classifier, however, by using an appropriate kernel trick, such as linear, polynomial, or radial, SVM also works well on non-linear cases.

Let $x_i \in R^d$ be the data and $y_i \in \{-1,+1\}$ denotes the classes for $i = 1,2,...,l$ where l is the cardinality. The separation of class {-1} and {+1} in the d dimensions is defined as $w \cdot x + b = 0$. A new data x_i will be in the class {-1} if the inequality $w \cdot x + b \leq -1$ is true and x_i will be in the class {+1} if the inequality $w \cdot x + b \geq +1$ is true. The maximum hyper plane is achieved by optimizing the distance between the hyper plane and the support vectors from the two classes, i.e. |1|/w. The flow of SVM models construction is depicted in Figure 1.

D. Evaluation Metrics

The performance of SVM model is usually assessed using testing sets. This performance evaluation has been widely used to avoid potential bias of the result due to over fitting of the model to training set [17]. Precision, recall, and F-measure are used to measure the classification accuracy. Precision (*P*) is the number of correct assignments (true positives) divided by the number of all returned results (true positives + false positives), while recall (*R*) is the number of correct assignments (true positives) divided by the number of correct assignments that should have been returned (the actual number of sentences belong to that class). Recall is similar to the sensitivity (TPR) in ROC analysis [17]. F-measure extends the accuracy metric that just measures the ratio of the correct results and acts as the harmonic mean of the precision and recall. F-measure has a value in the range of 0 to 1, where 0 is the worst and 1 is the best

$$p = \frac{TP}{TP + FP} \qquad r = \frac{TP}{TP + FN} \qquad F = \frac{2pr}{p + r} \quad (2)$$

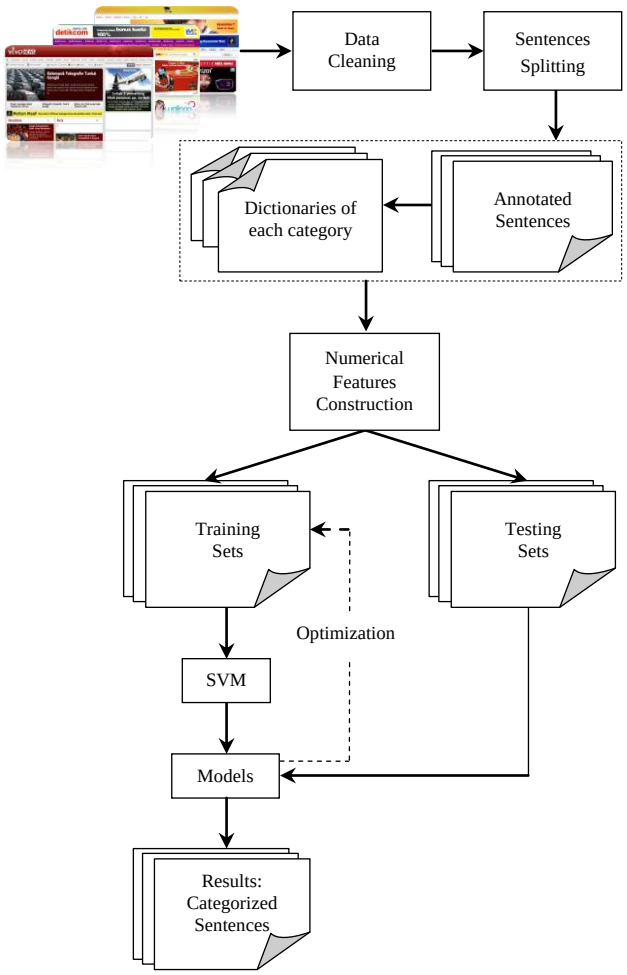


Figure 1. The flow of SVM models construction.

IV. Experimental Results

A. Datasets

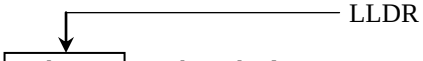
The datasets are sentences collected from 1,863 manually annotated web pages categorized as tropical disease [9]. The sentences were separated into two categories: (a) the sentences that contain the location of the incidence, the date, or the number of victims, labeled as {+1}; and (b) the sentences that contain no incidence information, labeled as {-1}. Table 2 shows the distribution of the sentences in each dataset. We divided the datasets into training and testing sets, and randomly selected 20-40% of the datasets for testing sets. The training set was used to construct SVM models while the testing set was used to evaluate their performance.

B. Results in Identifying Place Entities

We conducted an analysis to find all possible patterns to identify the location entities in the sentences. The patterns can be grouped into 4 cases:

Case 1: The words are possible to be the place entities, however, prior to them, the determinant contextual components are found and negate them as place entities.

The determinant contextual components are LLDR (location leader), GOAG (government agency), LOGA (leader of a government agency), POPL (public place), RELO (religious location), OPRE (organization prefix), SIAD (sign of address), and OPOS (position in an organization). We checked that the rule can handle this case very well as long as the determinant components are complete defined. Let's discuss an example of this case:



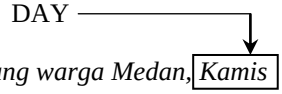
Governors of Aceh, Bali, and West Papua meet to discuss about malaria outbreak

The words *Aceh*, *Bali*, and *Papua Barat* are initially recognized as location entities. However, because prior to them an LLDR component is found, then all of them are canceled out.

TABLE 2
DATASET DISTRIBUTION BY CLASS LABELS

Dataset	Class	Number of Sentences	
		Training Set	**Testing Set**
Place of Incidence	+1	340	147
	-1	441	190
Total		**781**	**337**
Date of Occurance	+1	100	71
	-1	200	76
Total		**300**	**147**
Number of Victim	+1	300	72
	-1	100	38
Total		**400**	**110**

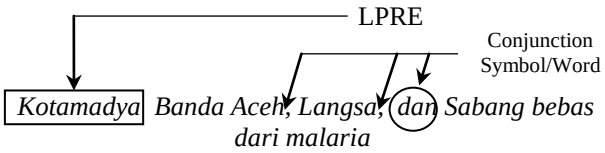
Case 2: The words satisfy the morphology rules, i.e. they are written in title case or uppercase, however if the words are labeled as DAY or MONTH, then they will not be considered as place entities. Let's see an example:



Dengue fever attacks the residents of Medan, Thursday

Kamis (Thursday) is the name of day of the week. Although the word is written in title case and satisfies the morphology rule, however, because it is a DAY, then the word will not be considered as a place entity. In the above example, only *Medan* is identified as a location entity. The same rule is applied if a word is a MONTH.

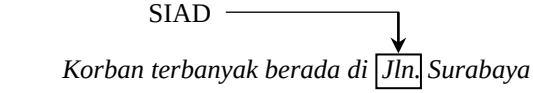
Case 3: The word satisfy the morphology rules, i.e. they are written in title case or uppercase, and prior to them, an LPRE (location prefix) or conjunction symbol is found. If that is the case, then the words will be considered as place entities. Let's discuss this example:



The city of Banda Aceh, Langsa, and Sabang are free from malaria

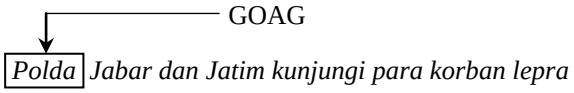
The two-word term *Banda Aceh* satisfies the morphology rule, exist in the database of places, and prior to it, a location prefix *Kotamadya* is found. Thus, *Banda Aceh* is recognized as a place entity. After the words *Banda Aceh*, a comma is found, and the next word after the comma, *Langsa*, satisfies the morphology rule and is found in the database of places, and the word itself is not labeled as LPRE, then the word *Langsa* is also identified as a location entity. The same rule is also true for the word *Sabang*. Hence, for the above example, the words *Banda Aceh*, *Langsa*, and *Sabang* are identified as place entities.

Case 4: The words satisfy the morphology rules, i.e. they are written in title case or uppercase, and they are tagged as APRO (abbreviation of province) or PT (popular town), but prior to them, the words are not tagged as one of the determinant contextual components mentioned in case 1. If this is the case, then the words will not be considered as place entities. Here are a few examples:



Most victims are found in Surabaya Street

Surabaya is listed in a popular town (PT). However, because prior to it a word *Jln* is a SIAD, then the word *Surabaya* will not be considered as a place entity.



West and East Java Police visit the victims of leprosy

Both *Jabar* and *Jatim* are the abbreviation of province (APRO). However, because prior to those words a word *Polda*, tagged as GOAG, is found then both *Jabar* and *Jatim* will not be considered as location entities.

Our empirical results show that from 1,328 location entities, 1,322 location entities were correctly identified by our algorithm and only 6 entities were incorrectly identified. The errors are due to misspelling and the use of capital letters for all the words in the sentences. In other words, the sentences are written in capital letters. The accuracy of our rule-based algorithm to identify the location entities, scored by F-measure, is 99.8%.

The sentences, which have been identified by the rule-based algorithm contain at least one place entity, are further classified by SVM classifier to determine which of those sentences contain the location of the tropical disease incidence. The values of F-measure of all SVM kernels, evaluated on training set, are shown in Table 3.

Empirically, polynomial is the best kernel of SVM for this purpose, i.e. 95.73%. The classification accuracy, evaluated on testing set consists of 337 sentences, is 82%. Table 4 summarizes the results on testing set.

TABLE 3
F-MEASURES OF ALL KERNELS EVALUATED ON TRAINING SET

SVM Kernel	F-Measure (%)
Linear	95.25
Polynomial	95.73
Radial	95.43

TABLE 4
CLASSIFICATION RESULTS TO DETERMINE WHETHER THE SENTENCES CONTAIN THE LOCATION OF TROPICAL DISEASE INCIDENCE

Dataset	Class	Sentences Classified as	
		+1	-1
Place of Incidence	+1	91	30
	-1	10	206
Total		101	236

$$\text{Precision} = \frac{91}{91+10} = 0.90, \text{Recall} = \frac{91}{91+30} = 0.75$$

$$\text{F-measure} = \frac{2 \cdot 0.90 \cdot 0.75}{0.90 + 0.75} = 0.82$$

The results are very conclusive. The accuracy of SVM classifier reaches 82%, and the SVM classifier yields recall and precision up to 75% and 90%, respectively.

C. Experimental Results in Identifying the Sentences that Contain the Occurrence Date

We also built an SVM classifier to determine whether a sentence contains information about the occurrence date of the tropical disease incidence. The occurrence date or time is usually written in a specific format as described in Section III. The numerical features of each sentence, used for SVM classifier, are also constructed using formula (1), i.e. estimating the ratio of 1-gram, 2-gram, and 3-gram words in the sentence and the dictionaries. If the sentences contain continuous time series information, they will be converted into ratio values based on n-grams. For the training set, the numbers of sentence in class {+1} and {-1} are 100 and 200, respectively. For the testing set, the numbers of sentence in class {+1} and {-1} are 71 and 76, respectively. Table 5 shows the values of F-measure for all SVM kernels, evaluated on testing set. The experimental results show that polynomial is also the best SVM kernel for this purpose. The evaluation on testing set demonstrates that the accuracy to classify the sentences that contain the occurrence date of tropical disease incidence is very convincing, i.e. up to 96.41%.

TABLE 5
F-MEASURES OF ALL KERNELS EVALUATED ON TESTING SET

SVM Kernel	F-Measure (%)
Linear	96.25
Polynomial	96.41
Radial	96.25

D. Experimental Results in Identifying the Sentences that Contain the Number of Victims

An SVM classifier is also trained and learned to effectively classify the sentences that have the number of victims of tropical disease in them. In order to complete this task, the numerical features of each sentence are constructed using formula (1). For the training set, the numbers of sentences in class {+1} and {-1} are 300 and 100, respectively. For the testing set, the numbers of sentences in class {+1} and {-1} are 72 and 38, respectively. Table 6 lists the values of F-measure for all SVM kernels, evaluated on testing set. Similar to the previous SVM, polynomial is also the best SVM kernel for this purpose. The evaluation results on testing set show that the accuracy to classify the sentences that contain the number of victims in them is also very promising, i.e. up to 93.38%.

TABLE 6
F-MEASURES OF ALL KERNELS EVALUATED ON TESTING SET

SVM Kernel	F-Measure (%)
Linear	92.73
Polynomial	93.38
Radial	93.21

E. Organizing Extracted Entities into KML to Integrate with Google Earth Application

After the place entities and the sentences that contain the occurrence date and the number of victims are extracted, they are organized into a standard KML file so that the locations of the tropical disease incidence can be viewed geographically in Google Earth application. The information that can be viewed, besides the locations, are the occurrence date of the event and the number of victims. KML is a scheme to describe and define geographic information on Google Earth software. It is a standard XML (eXtensible Markup Language) format containing specific elements and attributes [17].

A *placemark* tag is used to define a location on earth based on longitude and latitude coordinate values. The tag is symbolized by a yellow push pins in Google Earth application. A *point tag* is used to define the coordinates of an object, while a description tag is used to show additional information in a popup window. Figure 2

depicts the integration result in Google Earth.

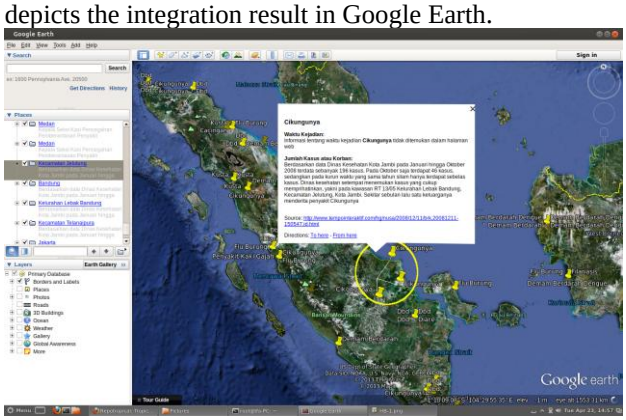


Figure 2. The extracted entities viewed in Google Earth.

V. Conclusion

Many tropical disease incidences in Indonesia are reported online in numerous news portals. News portals are valuable online data sources for creating a tropical disease repository if the locations of the tropical disease incidence, the date of occurrence, and the number of victims can be automatically extracted. In this paper, a rule-based algorithm to automatically identify the locations of tropical disease incidence from the web is proposed. The rule-based algorithm incorporates the database of places and the contextual and morphology components. The accuracy to identify the location entities is very conclusive, i.e. 99.8%. The accuracy of SVM classifier to determine the sentences that contain one or more locations of tropical disease incidence is 82%. The accuracy of SVM classifiers to classify the sentences that contain the date of occurrence and the number of victims are 96.41% and 93.38%, respectively. We believe that if the automatic extraction of location entities and sentences containing the date of occurrence and the number of victims can be scheduled, a tropical disease repository with a frequently updated data can be created.

VI. Acknowledgment

This work was supported by the Directorate General of Higher Education, Ministry of Education and Culture, Indonesia through Hibah Bersaing Grant 2012, 141/UN11/A.01/APBN-P2T/2012. We would to thank Teuku Ardiansyah and Rahmad Dimyati, the members of Data Mining and IR Research Group, Department of Informatics, for their valuable insights and help.

References

[1] Indonesian Ministry of Health, "Neglected Tropical Diseases in Indonesia: An Integrated Plan of Action", 2011.

[2] Behrens, et al., "The Incidence of Malaria in Travellers to South-East Asia: Is Local Malaria Transmission a Useful Risk Indicator?" Malaria Journal, vol. 9, no. 1, p. 266, 2010.

[3] Kompas Cetak, "Penyakit Tropis Tidak Teratasi", cetak.kompas.com/read/xml/2008/08/11/00563886/ penyakit.tropis.tidak.teratasi), cited on May 1, 2011.

[4] Worldwidewebsize.com, http://worldwideweb.com.

[5] H. Yu, J. Han, and K. Chang, "PEBL: Web Page Classification without Negative Examples", Journal of IEEE TKDE, vol. 16, no. 1, pp. 70–81, 2004.

[6] R. Bie, Z. Fu, Q. Sun, and C. Chen, "A Comparison Study of Bayesian Classifiers on Web Pages Classification", New Generation Computing, Ohmsha and Springer, vol. 28, pp. 161–168, 2010.

[7] X. Qi and B. Davison, "Web Page Classification: Features and Algorithms", ACM Computing Surveys Journal, vol. 41, no. 2, 2009.

[8] T. Abidin, A. Misbullah, and M. Subianto, "Determining Features of Web Documents and Building a Web Classifier using SVM", AISS (Advance in Information Sciences and Service Sciences: An International Journal of Research and Innovation), vol. 3, no. 10, pp. 401–408, 2011.

[9] T. Abidin, R. Ferdhiana, and H. Kamil, "Learning to Classify Tropical Disease Web Pages from Large Indonesian Web Documents", In Proc. of the 4th International Conference on Computer and Electrical Engineering, pp. 14–15, 2011.

[10] Pos Indonesia, "List of Places and Zip Codes in Indonesia", http://kodepos.nomor.net, cited on January, 2012.

[11] Joachims, "Making Large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning", B. Scholkopf, C. Burges and A. Smola (ed.), MIT Press, 1999.

[12] J. Zhao and F. Liu, "Product Named Entity Recognition in Chinese Text", Journal of Language Resources and Evaluation, vol. 42, no. 2, pp. 197–217, 2008.

[13] Y. Sari, M. Hassan, and N. Zamin, "Rule-based Pattern Extractor and Named Entity Recognition: A Hybrid Approach", Information Technology, IEEE, vol. 2, pp. 563–568, 2010.

[14] H. Chanlekha and N. Collier, "Analysis of Syntactic and Semantic Features for Fine-Grained Event-Spatial Understanding in Outbreak News Reports", Journal of Biomedical Semantics, vol. 1, no. 3, pp. 1–11, 2010.

[15] A. Spencer, Morphological Theory: an Introduction to Word Structure in Generative Grammar. Oxford & Cambridge, 1991, pp. xviii + 512.

[16] I. Budi, S. Bressan, G. Wahyudi, and Z. Hasibuan, "Named Entity Recognition for the Indonesian Language: Combining Contextual, Morphological, and Part-of-Speech Features into a Knowledge", pp. 57–69, 2005.

[17] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. Khoury, "Application of SVM Modeling for Prediction of Common Diseases: the Case of Diabetes & Pre-diabetes", Journal of BMC Medical Informatics and Decision Making, vol. 10, no. 16, pp. 1–7, 2010.

[18] Google Developers, "KML Documentation Intro.", code.google.com/apis/kml/documentation/, cited on April 2012.

Taufik F. Abidin is a faculty member at the Department of Informatics, College of Science, Syiah Kuala University, Banda Aceh, Indonesia. He received his B.Sc. from Sepuluh Nopember Institute of Technology, Indonesia in 1993 with predicate Cum Laude. He received his Master Degree in Computing from RMIT University, Melbourne, Australia in 2000, and completed his Ph.D. in Computer Science at North Dakota State University (NDSU), USA in 2006. He received the ND EPSCoR Doctoral Dissertation Award from NDSU in 2005 and has been a Senior Software Engineer at Ask.com in New Jersey, USA to develop algorithms and implement efficient production-level programs to improve web search results. His research interests include Data Mining, Text and Web Mining, Database Systems, Information Retrieval, and ICT for Development.

Dr. Abidin is a member of APTIKOM, International Association of Engineers and Computer Scientist (IAENG), and International Association of Computer Science and Information Technology (IACSIT). He is a Program Committee for the International Conference on Software Engineering and Data Engineering (SEDE) for many years. He has US Patent (7,836,090 B2) on the Method and System for Data Mining of Very Large Spatial Datasets using Vertical Set Inner Products.

Ridha Ferdhiana is a faculty member at the Department of Statistics, College of Science, Syiah Kuala University, Banda Aceh, Indonesia. She completed her Bachelor Degree in Mathematics from Sepuluh Nopember Institute of Technology (ITS), Indonesia in 1997 and completed her M.Sc. in Applied Statistics from North Dakota State University (NDSU), USA in 2006. Her research interests include data mining, nonparametric modelling, statistical computing with R, and analyzing key factors that impacted undergraduate students' GPA.

Mrs. Ferdhiana is an active member of FORSTAT, a statistics forum for statistician and people who interested in statistics. She is also an active member of Indonesian Mathematical Society (IndoMS).

Hajjul Kamil was born in East Aceh, Indonesia and he is a faculty member at the Departement of Nursing, College of Medical, Syiah Kuala University, Banda Aceh, Indonesia. He completed his Bachelor of Nursing from the University of Indonesia, Jakarta in 1999 and received his Master of Nursing from the same university in 2001. He is currently pursuing his doctoral degree program at the University of Gadjah Mada, Yogyakarta, Indonesia. His research interests include health, quality of health service, patients' safety, management of nursing, and public health.

Mr. Kamil is a member of The Association of Indonesian Nurse Education Center (AINEC), The Indonesian National Nurse Association (INNA), and The International Nurses of Council (ICN).