Journal of Fundamental and Applied Sciences

Research Article

ISSN 1112-9867

Special Issue

Available online at http://www.jfas.info

COMPARATIVE STUDY ON CORPUS DEVELOPMENT FOR MALAY INVESTMENT FRAUD DETECTION IN WEBSITE

M. M. Din*, N. H. H. Hashim and M. M. Siraj

Department of Computer Science, Faculty of Computing, Universiti Teknologi Malaysia,81310 Johor Bahru, Johor, Malaysia

Published online: 10 November 2017

ABSTRACT

In the online world, fraudsterscaneasily manipulate people to gain something and usually for monetary gain. Corpus development research can be use identify keywords used by fraudsters online to prevent the crime. The aim of this research is to develop a corpus for Malay investment fraud so that it can be used in detection and classification of investment fraud in Malay website and compare the most suitable technique. In this research, Part-of-Speech tagger (POS) and Named Entity Recognition (NER) tagger are selected. Proposed methodology that are used in this research is corpus development, training and development of dataset using Naïve Bayes and performance evaluation. The dataset used in this research is online news archive and discussion forums. This research able to help the law enforcements agencies in collecting and notifying the keyword used by fraudsters so that they can take any legal actions.

Keywords: corpus development; information extraction; part-of-speech; named entity recognition; fraud detection.

Author Correspondence, e-mail: mazura@utm.my

doi: http://dx.doi.org/10.4314/jfas.v9i6s.62

1. INTRODUCTION

In recent times, cyber fraud has been identified as a threat to organizations and individuals. It can lead to financial loss for individual and a lot of problem including loss of shareholder's and public confidence to company. If it is not prevented, it may give a very negative impact for the victims. With the rapidly growth of the World Wide Web (WWW) makes the volume of cyber fraud cases grows parallel. It can occur in chatroom, social media, message board and blogs. Some victims never report this because of privacy, shy or event think it is not worth it.

Although there exist organizations to penetrate cyber fraud, there are no automated system to handle this problem which expresses that some improvement still needed to prevent the crime earlier effectively. Current prevention works reactively where the public only report the incident to the law enforcement after the fraudster victimize them. After that, the data are entered to a system and the law enforcement agencies and legal actions will then take actions of the reported fraud cases. Extraction of information from web pages are advantages, there must be approaches available to efficiently identify and classify for this purpose. Text mining is the most suited approaches to be used when it comes to extracting huge amount of semi structured text as in web pages. A lot of works has been done to extract information in website in English while none of the research are done in Malay language fraud detection. Thus, using text mining can extract the meaningful text from the web and able to differentiate between fraud keywords and normal keywords to solve the problem so that the keyword can be extracted and displayed to the public.

This paper describes the process of developing a corpus for Malay investment fraud so that it can be used in detection and classification of fraud in Malay website.

1.2. Related Works

In this section, we review on the existing studies about corpus building in fraud detection.

1.2.1. Named Entity Recognition

Generic NER systems tend to focus on finding the names, places and organization that are mentioned in text news. As studied by0, they proposed a method for creating rules and gazetters for Iban language using text processing modules from A Nearly New Information Extraction (ANNIE) system. It uses Java Annotation Pattern Engine (JAPE) language to write

several rules to identify entities such as Person, Organization, Location and other entities such as monetary, date, percentage and time.

Linear-Chain CRF machine learning proposed by[1] is developed to train NER model in their study. NER model in this study is independently labelled, but the NER label for neighboring words are dependent. The advantage of this proposed technique is the ability to detect an organization based on the short form. For example, this technique can detect an organization with suffix like “Sdn” for Sendirian and “Bhd” for Berhad. This study has the 75.12% accuracy for identifying *Person* which is the highest among other entities.

In the research done by[2], he uses Rule-Based Part of Speech (RPOS) tagger to identify named entities in Malay dataset. The tagger is incorporated with a POS tagset and affixing rule so that it can be used for identification of word. The rules in this study are designed to detect three entities that is Person, Organization and Location. This study use affixing and word relation rule to categorize the correct word in correct category which enables Malay words to be formed with prefixes, suffixes, circum fixes and/or infixes. POS dictionary used in this study is manually constructed using Thesaurus Bahasa Melayu to assign all possible tags. In this study, infixes are considered not important and ineffective for tagging task. The dataset used within this research is from news article and Malay bio-medic article and the results indicated that it achieves higher performance with 89% accuracy compared to statistical POS tagger. It is revealed that this research enable to predict unknown words at a reasonable accuracy, but unable to tag borrowed word from English and word with no affixation.

A similar research done by [3]proposed a free indexing method to recognize person names in Malay text. The research manually analyzed the structures of Malaysians’ names commonly presented in Malay news articles. They extracted common titles and names and build the name indexer. The experiment conducted on 117 news articles produced 68% accuracy rate in identifying person’s names. The algorithm worked poorly on business news articles but showed an increased performance rate in political news articles. This was due to the nature of political news that involved name of authorities, VIPs or honorable persons.

1.2.2. Part-of-Speech Tagger

Even though Malay POS tagger are limited, there are some research published and one of the

research of POS Malay tagger is developed by[4]using trigram Hidden Markov Model (HMM). The purpose of this study is to detect tagset in Malay sentences and discusses about the effect of using prefix and suffix to predict POS tag correctly. The study test the effect of using both prefix and suffix individually as well as the combination of them using a corpus consisted of 1835 tagged token and 21 tagset used in Dewan Bahasa Pustaka (DBP) tagset. Experimental results disclosed that only using prefixed information which length is three letters made the best predictions with 67.9% accuracy. Other results revealed that using combination of first and last three letter give 66.7% accuracy. When using only suffixes information using five letters suffix length, 60% accuracy achieved. The result from this study shows that HMMs is applicable for predicting Malay word's POS tag.

Apart from rule based, statistical machine learning method is also used to develop POS tagger. This study was done by[5],which is called "Lazy Man's Way". The reason behind unique name is because of the annotation process. This method does not require heavy process of annotating the dataset for training like other method as it annotates dataset using a Malay-English lexicon. The annotation process of this method are as follows.First, the Malay words are translated to English using Google translate. Next, the dataset is tagged using Brill' stagger and then the result are mapped using Malay-English lexicon. This method has 86.87% precision. Even though this method has somewhat high percentage, it may produce imprecise result since the grammatical structure in English and Malay are different and this method relies on the Malay-English lexicon.

Another study of POS tagger is proposed by[6]developed for Bahasa Indonesia. This study used combination of two machine learning for comparison; Conditional Random Fields (CRF) and Maximum Entropy (MaxEnt). CRF enables several feature functions weighted with the result from training corpus. Compared to HMM, CRF are more general and not restricted. The second method used is MaxEnt so that the POS developed is flexible and can maximize the use of context information. They used two corpora for the evaluation where the dataset used 37 and 25 tagset. The first corpus is built manually with 14165 token and the second corpus is part of Pan Localization project which has around 500000 tokens. Results reveals that the model that has the highest accuracy for both corpora is MaxEnt at 85.02% and for CRF model the highest accuracy is when using the second corpus and 25 tags with 91.15%.

Besides that, a research completed by[7]for Malay POS tagger called Mi-POS revealed to have higher accuracy at 95.16%. Mi-POS is a statistical POS tagger approaches that use a large dataset. For creating dataset, they manually built three corpora; one for training purposes with 64354 tokens and another two for testing purposes with 359 and 550 tokens respectively. The author had stated that they used dataset from BERNAMA article and non- articles. This research is comparing five existing Malay corpora in terms of strengths and weaknesses. The results show that dataset in BERNAMA articles achieve higher accuracy as the training corpus is composed from BERNAMA articles.

2. METHODOLOGY

Developing a corpus is a laborious task so machine learning technology is adopted in this study which is Naïve Bayes. The flowchart for this study is represented in Fig. 1.

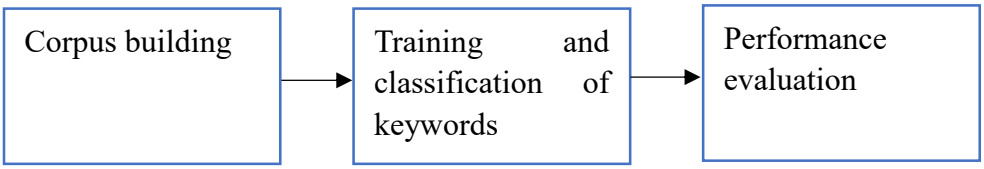

Fig.1. Flowchart for developing a Malay investment fraud detection in website

2.1. Corpus Building

Raw data are collected from the web and needs to be cleaned using HTML parser as the resulted plaintext are in the form of HTML. The data from the website are then added with the data from Dewan Bahasa and Pustaka (DBP) corpus. After that, the data is going to be processed using MaxEnt POS tagger and the process are continued with training and classification. POS tagger will reveal the highest occurrence of word used in the dataset. Then, the process is repeated with Illinois NER tagger. NER can allocate *person* (PER), *location* (LOC), *organization* (ORG) and *miscellaneous* (MISC) in the corpus while Part-of-Speech (POS) tagging capable to annotate each word in the articles with a unique tag representing its function.

Corpus building consist of three steps: i) collecting raw data ii) HTML parser and iii) building the corpus using POS and NER tagging.

2.1.1. Raw Data

The data used in this experiment is crime news article from Berita Harian and Utusan Online's news and forum in CariGold.com. These articles are manually downloaded from respective website and compiled in a software called SketchEngine to build the corpus. From there, the data then are collected, compiled and stored like a database. The articles extracted are only limited to crimes involving investment fraud since year 2014 until 2017.

2.1.2. HTML Parser

HTML parser is a tool that allows user to manipulate and analyze HTML documents and often used to remove all the HTML tag in the data. We use Tika Apache pakage and Eclipse software.

2.1.3. Building the Corpus

To complete the experiment, the cleaned data are then going through a POS and NER tagger.

2.1.3.1. POS Tagger

We use the POS tagger called MaxEnt tagger which is openly available and its tagset are listed in Table 1 which consists of 31 tagsets. Most of the tagset are similar with other tagset but in this tagset, there exist symbol and foreign word tagset. Besides that, what makes it different is that MaxEnt POS tagger is developed for Malay language.

Table 1. Tagset for MaxEnt POS Tagger

SYM:	Symbol	CDP:	Primary cardinal numeral
NNC:	Countable common noun	CDO:	Ordinal cardinal numeral
NNU:	Uncountable common noun	CDI:	Irregular cardinal numeral
NNG:	Genitive Common noun	CDC:	Collective cardinal numeral
NNP:	Proper noun	CD:	Cardinal numeral
NN:	Common noun	NEG:	Negation
PRP:	Personal pronoun	IN:	Preposition
PRN:	Number pronoun	CC:	Coordinate conjunction
PRL:	Locative pronoun	RB:	Adverb
PR:	Common pronoun	UH:	Interjection
WP:	WH-pronoun	DT:	Determiner
VBT:	Transitive Verb	WDT:	WH-determiner
VBI:	Intransitive Verb	RP:	Particle
VB:	Verb	FW:	Foreign word
MD:	Modal	JJ:	Adjective
SC:	Subordinate conjunction		

2.1.3.2. NER Tagger

We use Illinois NER which is a java program that uses gazetteers extracted from Wikipedia. In NER, there are different types of serialized model to choose for labelling. This model is limited to the amount of tag that user want to use. There are three types of serialized model which is: 3 classifiers, 4 classifiers and 7 classifiers. The most basic classifiers with serialized model included is 3 class NER tagger that can label: PERSON, ORGANIZATION and LOCATION entities. For Illinois NER, it can tag the sentences in news article into four types of entities; (Organization, Location, People and Miscellaneous).

2.2. Training and Classification of Keywords

In this experiment, the feature will be trained and tested using Naïve Bayes. The dataset of fraudulent keyword will be annotated with features and labels where 70% will be choose as a training data and another 30% will be used for testing.

2.2.1. POS Tagger

Firstly, the training process used the unigram tagger which is the process of grouping one tagsets into a group. So,the keywords are arranged according to the highest occurrence in the dataset. After that, 70% of the data which is the frequently used word are used for training. The tagset used are the ones related to the keywords only.

Then, the Naive Bayes algorithm will learn the weight between features from training dataset and will create a model. Finally, the model will assign label for each of the sentences in the article. The process of training for POS tag are using Naïve Bayes and the keywords are assigned as keywords in respective tagset with numerical number.

2.2.2. NER Tagger

For training, 50% of the dataset is trained using machine learning with two classes, FOS and NEG labelled assignment. FOS means that the sentences contain organization name, while NEG indicates that there is no organization name within the sentences. These labelled feature is created to identify the fraud organization tagged by NER. After that, 30% of the data is tested to identify the accuracy of each labelled sentences.

After that, we continue with classification process where the keywords are arranged into their group. The classification process is using TF-IDF method by ignoring the less frequent keyword as it does not have any function in text classification. Besides that, removing the low frequency keywords resulting in better performance.

2.3. Performance Evaluation

In evaluation, we chose the best methods from the previous stage. After that, the effectiveness in terms of accuracy between both taggers are tested to highlight the best taggers for Malay investment fraud corpus.

$$accuracy = \frac{number\ of\ correct\ POS\ tagged\ data}{number\ of\ correct\ POS\ tags\ in\ gold\ data} \quad (1)$$

3. RESULTS AND DISCUSSION

Dataset used in this experiment is 19623 words with 3392 token. Tagset that are used is from MaxEnt tagset, which contains 31 tagsets. From the data, both NER and POS tagger contains missed tagged token which contributed to lower accuracy. From the experiment above, the

accuracy is measured by confusion matric which and the result reveals that using POS tagger will give higher accuracy than using NER in Malay corpus. As illustrated in Table 2 and 3, it is shown that the accuracy for POS is 10% with 340 correctly tagged tagset. While for NER, the accuracy is 50% with 1696 tagset that are correctly tagged.

Table 2. Tagging Accuracy for MaxEnt POS tagger

Tokens (A)	Correctly Tagged (B)	Accuracy (B/A ×100) %
3392	340	15.21

Table 3. Tagging Accuracy for Illinois NER tagger

Tokens (A)	Correctly Tagged (B)	Accuracy (B/A ×100) %
3392	1696	50

In this experiment, the Naïve Bayes algorithm that are used are different for POS and NER as we use the mono classification for NER and multiclass classification for POS. The results are revealed in Fig. 2.

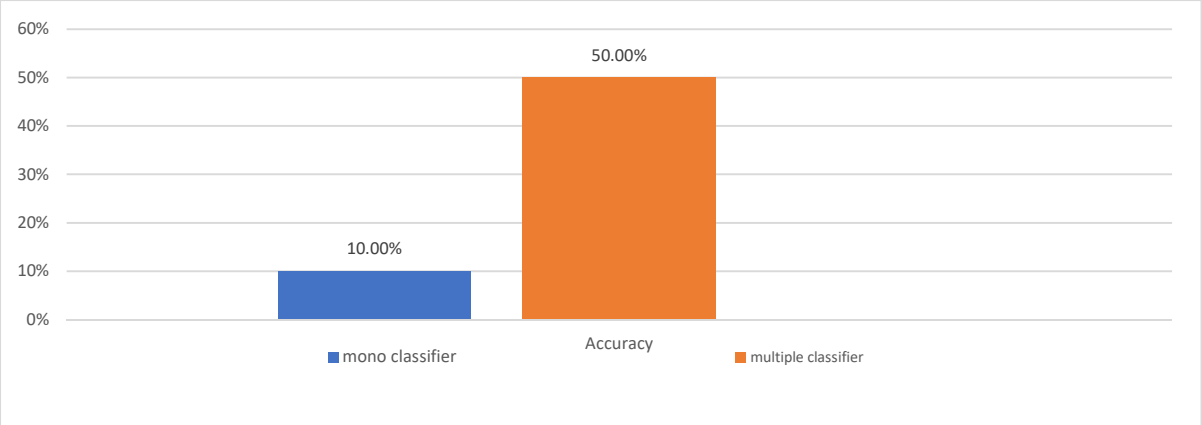

Fig.2. Accuracy for mono and multiple classifiers for POS and NER using Naïve Bayes

4. CONCLUSION

In recent times, cyber fraud has been identified as a threat to organizations and individuals. Hence, a predictive algorithm which able to analyze the web and identify the fraudulent website needs to be developed. Based on conducted experiment, it can be concluded that there are several factors that need to be considered to get a better result for accuracy. One example is to starts with a good tagset and training dataset. Results for this experiment reveals that the most suitable technique for Malay investment fraud corpus is NER technique in terms of

accuracy. Even though the proposed method claimed to be satisfactory, there are some limitations that needs to be improvised and it is quite difficult to develop a Malay corpus as it requires a lot of time and effort. Besides that, the dataset used in this project are from forums where language used in forum are not as standard language which is less formal. Future work may be conducted in multilingual to represent a broader range of language pattern and using other machine learning for testing and classification.

5. REFERENCES

[1] Fong Y S, Ranaivo-Malançon B, Wee A Y. NERSIL-The Named-Entity Recognition System for Iban language. In 5th Pacific Asia Conference on Language, Information and Computation, 2011, pp. 549–558

[1] Ulanganathan T, Ebrahim A, Xian B C M, Bouzekri K, Mahmud R, Hoe O H. Benchmarking Mi-NER: Malay entity recognition engine. In 9th International Conference on Information, Process, and Knowledge Management, 2017, pp. 52–58

[2] Alfred R, Leong L C, On C K, Anthony P. Malay named entity recognition based on rule-based approach. International Journal of Machine Learning and Computing, 2014, 4(3):300–306

[3] Sharum M Y, Abdullah M T, Sulaiman M N, Murad M A A, Hamzah Z A Z. Name extraction for unstructured Malay text. InIEEE Symposium on Computers and Informatics, 2011, pp. 787–791

[4] Mohamed H, Omar N, Ab Aziz M J. Statistical Malay part-of-speech (POS) tagger using Hidden Markov approach. In International Conference on Semantic Technology and Information Retrieval, 2011, pp. 231–236

[5] Zamin N, Oxley A, Abu B Z, Farhan S A. A lazy man's way to part-of-speech tagging. In D. Richards,& B. H. Kang (Eds.), Knowledge management and acquisition for intelligent systems: 12th Pacific Rim knowledge acquisition workshop.Cham: Springer International Publishing,2012, pp. 106–117

[6] Pisceldo F, Adriani M, Manurung R. Probabilistic part of speech tagging for Bahasa Indonesia. In 3rd International MALINDO Workshop. 2009, pp. 1-6

[7] Xian B C M, Lubani M, Ping L K, Bouzekri K, Mahmud R, Lukose D. Benchmarking

Mi-POS: Malay part-of-speech tagger. International Journal of Knowledge Engineering, 2016, 2(3):115–121

How to cite this article:
Din M M, Hashim N H H, Siraj M M. Comparative study on corpus development for malay investment fraud detection in website. J. Fundam. Appl. Sci., 2017, *9(6S), 828-838.*