Named Entity Recognition of South China Sea Conflicts

Nur Rafeeqkha Sulaiman[1], **Maheyzah Md Siraj**[2] **and Mazura Mat Din**[3]
[1,2,3]School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, 81310 Skudai, Malaysia.

E-mail: rafeeqkha@graduate.utm.my, maheyzah@utm.my, mazura@utm.my

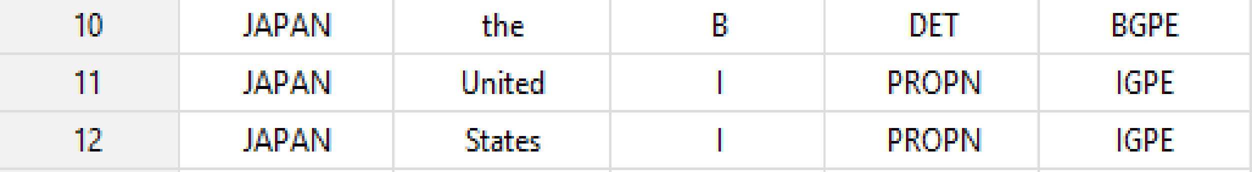
Abstract. Online news articles not only provide us with useful and reliable information and reports, it also eases information extraction and gathering for research purposes especially in Natural Language Processing (NLP) and machine learning (ML). The topics regarding the South China Sea have been popular lately due to the rise of conflicts between several countries claim on the islands in the sea. Gathering data through Internet and online sources proves to be easy, but to process a huge amount of data and to identify only useful information is no longer possible. Because of that, relevant information and the classification of news articles in relation to the conflicts need to be done. In this paper, a model is proposed to use NER that search for and classifies important information regarding to the conflicts. In order to do that, a combination of POS and NER are needed to extract meaningful information from the news. This study also aims to classify conflict related news by using Conditional Random Field (CRF) algorithm as classification method by training and testing the data.

1. Introduction

China's claim on South China Sea was deemed to be historical by China itself, however such claims is not taken seriously in international law, which from China's point of view downgrades China's ancestral heritage and is a source of anger [1]. The conflict in South China Sea includes not only China, but also other Asian countries like Malaysia, Brunei, Singapore, Vietnam, the Philippines, Indonesia and The United States of America. Newspapers are important repository for historical research. Due to the rise of Internet, newspapers have been written online and can be freely accessible by anyone with internet connections. The events occurred in connections with the conflicts in South China Sea have been documented directly. Besides that, online news articles make it easier for researchers to crawl and scrape data needed for research purposes or documentation.

Gathering data through Internet and online sources proves to be easy, but to process a huge amount of data and to identify only useful information is no longer possible. The advance in the world of computer software and applications nowadays eases the application of Natural Language Processing (NLP) on data. Some of NLP tasks are Part-of-Speech (POS) and Named Entity Recognition (NER). Natural Language Processing (NLP) are characterized by making complex interdependent decisions that require large amounts of knowledge for training prior to testing. The main goal of NLP is to convert human language into a specific language or representation that can be manipulated by the computers. One of NLP application is Named Entity Recognition (NER). Generally, NER is understood as the task of classifying information units like a person, countries, organizations and locations [2]. NER mainly focus on formal texts such as news articles due to the easier identification of texts or sentences compared to informal texts such as e-mail and tweets.

POS is known as assigning or marking each word of a text the proper morphosyntactic tag in its context of appearance [3]. There are many different libraries for POS taggers and each tagger has their own way of defining their text. we have always been taught in school that there are generally 9 POS in English. however, there are actually more categories and subcategories exists and POS tagger helps us in identifying them. The more specific POS tagger that has been developed in the world of NLP really helps in identifying the correct and more accurate text or word. POS and NER are two different taggers where NER tags a chunk of named entity while POS tags each word in a text. therefore, combining these two tags are impossible to imagine at first. IOB scheme helps in combining POS and NER so that though a chink is separated, IOB helps in identifying a separated chunk by applying a I, O or B in the beginning of a label. For example, Figure 1 shows a chunk "the United States" being tokenized and IOB scheme helps to identify the beginning and the inside of a chunk.



10	JAPAN	the	B	DET	BGPE
11	JAPAN	United	I	PROPN	IGPE
12	JAPAN	States	I	PROPN	IGPE

Figure 1. A screenshot of separated chunk and IOB label.

To expand and broaden the knowledge in classifying named entity recognition of South China Sea conflicts, a study is conducted as it can be used for further researches in NER classification for South China Sea conflicts. The data used are news articles concerning South China Sea conflicts and the data were crawled from various websites from 4 different countries: Japan, Malaysia, Singapore, and, Vietnam. In addition, CRF is used for classification of conflict related information.

2. Related Work

NER is being used extensively to study formal text such as the news and various methods and structures have been developed and studied. Though current studies of NER mainly focus on formal text, studies on informal texts such as tweet [4] and emails [5] has also been done in order to face the issues on detecting informal languages and texts through NER. In this sections, related works of NER discussed on two types of NER: NER on formal text and NER on informal text.

NER is defined as the task of detecting or categorizing a person name, organizations and other named entities depend on which library or corpus is used. Due to its popularity, NER has been developed and improved to other countries and is structured uniquely for different languages as no all languages have the same word, spelling and meaning in every language. In this section NER on formal text of three different languages are discussed. The three different languages are Telugu, Filipino or Tagalog, and English.

The increase in the number of crime information available on the web is crucial in the documentation process as it eases the retrieval and exploiting relevant information needed to provide the insight into criminal behaviour and networks to fight crime more efficiently and effectively. Crime NER and Crime type identification system based on ensemble framework was done in order to synthesize a more accurate classification procedure [6]. The text classification algorithm used were Naïve Bayes [7], Support Vector Machine [8], and K-Nearest Neighbour classifiers [9]. The data used in this study were crawled from the Malaysian National News Agency (BERNAMA). The named entities tagged were type of crime, weapons, location, and nationality involved. All these annotations were manually annotated and classified. Feature extraction was done to enhance the performance. Feature extraction converts each word to a vector of each feature values.

Before conducting the experiment, Vector Space Model (VSM) is used to convert a full text document to a document vector to make the document simpler and easier to deal with. Like most machine learning experiment, a test set and a training set is prepared. Once the data was tested by using NB, SVM and KNN, the results were analysed through standard evaluation namely: Precision, Recall, F-Measure, and Macro-average (F1). Based on the result, the highest result yielded by individual classifier was by SVM and the lowest result was yielded by the NB classifier. SVM also yielded the highest result according to the experiments of the crime named entity covered (weapon, nationality and crime locations).

A recent study done by [10] in 2018 focus on NER to detect or classify Filipino news articles related to disaster. The Philippines is an Asian country that is prone to natural disasters and is considered as the world's disaster 'hot spot'. Natural disasters that have hit the archipelago are earthquakes, volcanic eruptions, typhoons, floods and droughts. They have occurred so frequently that they have helped in shaping the Filipino society [11]. For this study, Pilipino Star NGAYON which is an online news portal for the Philippines were used as the data. Instead of English, the data was in Filipino and a total of 354 news articles were crawled from the web. Entities chosen for the study are <*TOD*> type of disaster, <*NOD*> name of disaster, <*MOS*> month, <*LOC*> location, and <*O*> for other.

The deep learning process is done by using NER and is built by using TensorFlow. An open source NER model using TensorFlow (LSTM + CRF + chars embeddings) is used to implement the data for Filipino. the words were first converted into vectors which represent the word by using bi-LSTM, after the word representation, contextual word representation was obtained through LSTM. The system then used a fully connected neural network to get a vector where each entry corresponds to a score for each tag and a linear-chain CRF to make the final prediction. The results were then measured by using Accuracy and F-measure. Table 1 shows the result of training set per epoch and table 2 shows the outcome for the test data.

Telugu is an entirely different language which uses different alphabet or characters. Therefore, NER on any Telugu words or text proved to be more challenging than languages using modern alphabets. In order to do NER on Telugu, morphological pre-processing has to be done on the dataset. A study on NER for Telugu news articles proposed a language dependent features like post-position feature, clue word feature and gazetteer feature to improve the performance of the model [12]. NER on Indian Language (IL) proves to be more challenging than other languages which uses the modern English Alphabet as capitalization feature play an important role as NEs are generally capitalized in English. the challenges specific to Telugu language are: a) it does not have capitalization feature b) two words in English can be mapped to one word in Telugu c) absence of part-of-speech tagger d) free word ordering. In this paper, Naïve Bayes classifier was used for NER task. The data used was crawled from Telugu Newspaper and was annotated with three NE namely Person, Location, Organization and not named entity class. Due to the different character or words used by Telugu, morphological pre-processing was done on the dataset.

For this study, two types of experiment were done: a) Contextual features and Naïve Bayes Classifier, b) Language dependent features and building comprehensive Naïve Bayes Classifier. In a, the contextual word and POS features are used to build the prediction model. In b, a Boolean feature was introduced by assigning '1' to a Proper noun and '0' to a non-Proper noun. Based on the result, the accuracies were improved after morphological process and language dependent features improved the prediction accuracies.

Twitter has become one of the centre source of information for gathering data for their datasets. NER on tweets is a challenge as it is a type of informal text. Most of the words used are short forms, slangs, mixed language, and inconsistent use of capitalization. A study done by [13] tackles the issues of tweets such as: insufficient information in a single tweet and noisy and short data. The proposed a method which controls redundancy in tweets by conducting a two-stage NER for multiple similar tweets. In the first stage, CRF-based labeller is used; and in the second stage, pre-labelled tweets were clustered and cluster level labelling using and enhanced CRF-based labeller that employs cross-tweet information was conducted. Just like tweets, emails include in informal texts categories. A study done by [5] proposed two methods for improving performance of person name recognizers for email: email-specific structural features and a recall-enhancing method which exploits name repetition across multiple documents. Their study featured POS tags and NP chunking of the email however POS is eliminated due to the amount of noise it created. CRF model was used in their study to classification results.

3. The Proposed Methodology

In this section we discuss on methods and processes used for the proposed method.

3.1. Data Collection and Data Pre-Processing

For South China Sea related news, articles from various websites from different countries are used. A total of 225 news articles were crawled from websites from 4 different countries namely: Malaysia, Singapore, Japan, and Vietnam. Table 1 shows the number of news articles from various websites of different countries. A total of 137297 texts are loaded as dataset however, only 17225 are trained and tested after words with <*O*> label is omitted.

Table 1. Number of data crawled from online news websites.

Country	Name of websites	Number of data
Malaysia	The Star Online	50
Singapore	The Straits Time	75
Japan	Kyodo News	50
Vietnam	VietNam News Vietnam News Agency (VNanet)	50

The articles were crawled by using Python and keywords related to South China Sea conflict were used for crawling. Each of the articles were manually checked to confirm its relation to South China Sea conflicts. News articles were chosen as the dataset for this study as they provide South China Sea conflict-related information and also free to be accessed. Besides that, news articles provide a widespread of what happened in the past and in the present, which proves to be very useful resource for researchers in this field. The collected data was cleaned by using Python which includes removing stop words, punctuations, and HTML parsing. Pre-processing is a crucial step during this phase in order to get the right keyword for the next process which is feature extraction. Pre-processing is the process where all raw data is cleaned in a way where the output will be a clean data. HTML parser removed all the HTML tags found in the raw data and is combined together with removing stop words and stemming in order to get the required keywords from data samples

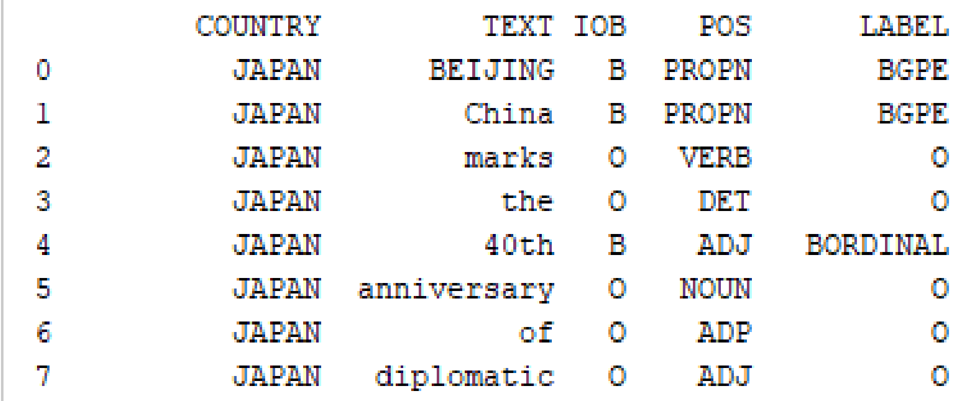
3.2. Named Entity Recognition and Part-of-Speech (POS) Tagging with spaCy

After the data is cleaned and pre-processed, NER tagging takes place. NER tagging or entity extraction is a popular technique used in information extraction to identify and segment the named entities and classify them under various predefined classes. For this research, Python is used with spaCy for named entity recognition. SpaCy library has been trained on the OntoNotes5 corpus. For this experiment, two types of NER annotation tools were used and spaCy was chosen as it is more accurate and detects more entity types than NLTK. During annotation, every articles are labelled with entity types and is saved into csv. files. When annotation was done, the datasets were tokenized in order to map the word token with the entity type. For this research token level entity is also done by using spaCy with IOB tagging scheme. Table 2 below shows the tag and its entity. Token with tag <O> will be omitted for this experiment to increase the rate of accuracy.

Table 2. IOB tags and its description.

Tag	Description
B	The first token of a multi-token entity
I	An inner token of a multi-token entity
O	A non-entity token

Part-of-speech (POS) tagger is used in tagging the news articles with English grammar and vocabulary to form the POS tag set to the datasets such as nouns, verbs, preposition, adverbs and etc. for this experiment, spaCY is used for POS tagging. Figure 2 shows a snippet of the output of POS, NER annotation, and IOB tagging implementation on our dataset. based on Figure below, COUNTRY stands for the origin of news articles, TEXT stands for word in the news article, IOB stands for IOB tags, POS is Part-of-speech, and LABEL is the combination of IOB and NER. For example, BGPE stands for B (the first token of a multi-token entity) and GPE (countries, cities, or states).



```
  COUNTRY         TEXT IOB    POS     LABEL
0   JAPAN      BEIJING   B  PROPN      BGPE
1   JAPAN        China   B  PROPN      BGPE
2   JAPAN        marks   O   VERB         O
3   JAPAN          the   O    DET         O
4   JAPAN         40th   B    ADJ  BORDINAL
5   JAPAN  anniversary   O   NOUN         O
6   JAPAN           of   O    ADP         O
7   JAPAN   diplomatic   O    ADJ         O
```

Figure 2. A snippet of the output of POS, NER annotations and IOB implementation.

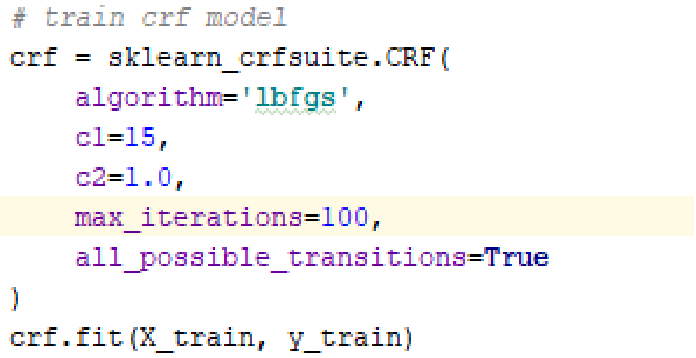
3.3. Training of the Model

In this phase, there are three parts which are: part of speech tagging, data training using classification algorithm, and result from the training data. For this project, Conditional Random Field (CRF) is used as classification algorithm in order to train and test the data with respective division sets of data. The data is split into two, one for training and another one for model testing. For this project, 60% of the datasets is used for training while the remaining 40% is used for testing data. CRFsuite wrapper is used in Python for this experiment. CRF is often used for labelling or parsing of sequential data such as NLP. Sklearn-crfsuite is used to train CRF model for NER on our dataset.

In this CRF model, the algorithm used was Limited-memory Broyden-Fletched-Goldfarb-Shanno (lbfgs). This method was chosen due to its parameter estimation in machine learning. c1 and c2 values are the regularization of the parameter. C1 is the coefficient for L1 regularization. The default value of c1can be zero in which it means no L1 regularization. C2 is the coefficient for L2 regularization. By default, c2 value is zero throughout the experiment.

3.4. Testing of the Model

The model was tested to evaluate the F1 score. At every named entity in the classification report, F1 score was generated together with Recall and Precision value. The report ends with micro average, macro average and weighted average. Only weighted values for F1-score was taken. Weighted average was used in order to find the accuracy of CRF model as it finds the average weighted by the support number from each labels. For testing, 3 tests were done by applying 3 different c1 values. The values are 1.0, 10, and 15. Figure 3 below shows the example of CRF model code snippet in Python.



```
# train crf model
crf = sklearn_crfsuite.CRF(
    algorithm='lbfgs',
    c1=15,
    c2=1.0,
    max_iterations=100,
    all_possible_transitions=True
)
crf.fit(X_train, y_train)
```

Figure 3. Example of CRF model code in Python.

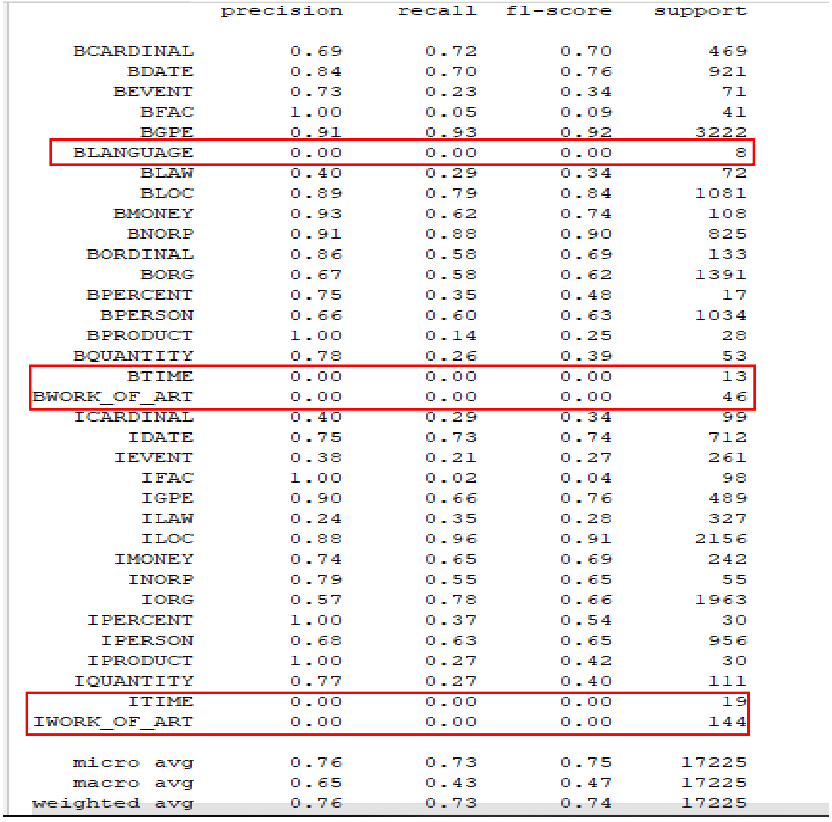
4. Results and Discussion

This section explains the result obtained based on the implementation of NER technique and use of CRF model in training and testing the data. Table below shows the accuracy of CRF model based on our datasets. Figure below shows the example of classification accuracy output in python. Weighted overage of F1-score was taken for accuracy values. Table below shows the result of training set with both parameters set to 0.1. The reason both parameters are set to 0.1 is to see the performance of the model to the datasets. F1-score is used to evaluate the performance as it is interpreted as a weighted average pf the precision and recall. F1-score reaches its best value at 1 and worst at 0. Compared to Accuracy, F1-score is a better measure to use if there is an uneven class distribution (large number of Actual Negatives). In this project weighted average of F1-score is used to measure the accuracy of CRF model. It is stated that the higher the F1-score, the higher the accuracy of the CR model. Weighted average takes note of the class imbalance by computing the average of binary metrics in which each class's score is weighted by its presence in the true data sample.

Table 4. The result of training data.

Parameters	Precision	*Recall*	F1-Score
Micro Average	0.80	0.76	0.78
Macro Average	0.65	0.50	0.55
Weighted Average	0.80	0.76	0.77

Table 4 shows the outcome of tuning of c1 parameter done in testing the CRF model and its classification result. From the table, the difference between the values of precision and recall were only slightly different from each other. It can be seen that as the value of c1 increases, the value of Precision, Recall, and F1-score decreases. Although c1 value was increased in a large number, the values of Precision, Recall, and F1-score did not decrease dramatically and only differ by less than -+0.2 in value. However, in all three classification results, there are some entity with no value or 0.00. this is an example of ill-defined entity due to the high value of c1 which might affect the value of these entities. As the number of c1 value increases, the number of named entities with 0.00 values increases. Figure 4 shows 5 entities with 0.00 value when c1 is set to 1.0 namely: BLANGUAGE, BTIME, BWORK_OF_ART, ITIME, and IWORK_OF_ART.



```
                precision    recall  f1-score   support

     BCARDINAL       0.69      0.72      0.70       469
         BDATE       0.84      0.70      0.76       921
        BEVENT       0.73      0.23      0.34        71
          BFAC       1.00      0.05      0.09        41
          BGPE       0.91      0.93      0.92      3222
     BLANGUAGE       0.00      0.00      0.00         8
          BLAW       0.40      0.29      0.34        72
          BLOC       0.89      0.79      0.84      1081
        BMONEY       0.93      0.62      0.74       108
         BNORP       0.91      0.88      0.90       825
      BORDINAL       0.86      0.58      0.69       133
          BORG       0.67      0.58      0.62      1391
      BPERCENT       0.75      0.35      0.48        17
       BPERSON       0.66      0.60      0.63      1034
      BPRODUCT       1.00      0.14      0.25        28
     BQUANTITY       0.78      0.26      0.39        53
         BTIME       0.00      0.00      0.00        13
  BWORK_OF_ART       0.00      0.00      0.00        46
     ICARDINAL       0.40      0.29      0.34        99
         IDATE       0.75      0.73      0.74       712
        IEVENT       0.38      0.21      0.27       261
          IFAC       1.00      0.02      0.04        98
          IGPE       0.90      0.66      0.76       489
          ILAW       0.24      0.35      0.28       327
          ILOC       0.88      0.96      0.91      2156
        IMONEY       0.74      0.65      0.69       242
         INORP       0.79      0.55      0.65        55
          IORG       0.57      0.78      0.66      1963
      IPERCENT       1.00      0.37      0.54        30
       IPERSON       0.68      0.63      0.65       956
      IPRODUCT       1.00      0.27      0.42        30
     IQUANTITY       0.77      0.27      0.40       111
         ITIME       0.00      0.00      0.00        19
  IWORK_OF_ART       0.00      0.00      0.00       144

     micro avg       0.76      0.73      0.75     17225
     macro avg       0.65      0.43      0.47     17225
  weighted avg       0.76      0.73      0.74     17225
```

Figure 4. A snippet of the whole result and entities with 0.00 value.

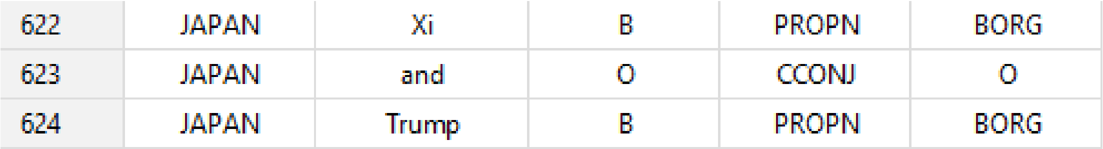
Table 5. Results of testing data with different c1 values.

Parameters	**Micro Average**			**Macro Average**			**Weighted Average**		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
C1: 1.0	0.76	0.73	0.75	0.65	0.43	0.47	0.76	0.73	0.74
C1: 10	0.66	0.63	0.64	0.34	0.28	0.30	0.64	0.63	0.62
C1: 15	0.62	0.59	0.61	0.32	0.26	0.27	0.61	0.59	0.59

Based on the result from different parameter values, the result shows a slight difference for micro average and weighted average for the three metrics of Precision, Recall, and F1-score. However, in macro average result the difference were -+0.5 which is a big difference compared to weighted average and micro average values. In conclusion, the lower the c1 value, the more accurate the CRF model.

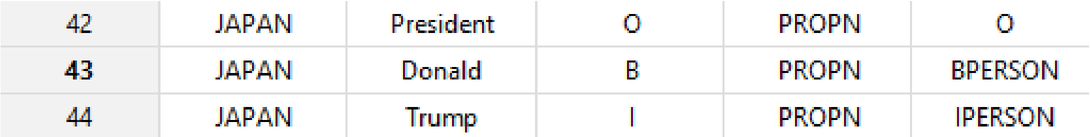
As for the system, the distribution of classes in training and test sets is unknown. This is because *train_test_split* function only take accounts of percentage of dataset to be split without taking in consideration of number of classes. Therefore, the values of named entities in both training and test datasets might differ a lot. However, F1-score is known as the mean of precision and recall. F1-score is very useful in an uneven class distribution as it takes both false positive and false negative into accounts.

One issue to be taken into consideration on the dataset is the inconsistency of POS and NER tagging. Due to the tokenization of word, the named entity is being separated and the probability of incorrect tags of named entity is high. Although, IOB tagging helps in identifying a named entity chunked, there are still some entity in the chunked being left out thus affecting the class of the entity. Besides that, incorrect label of named entity is also a concern in this study. Based on Figure below, Xi and Trump is labelled as <*ORG*> which means they are labelled as organizations rather than <*PERSON*> which is an entity for a person. However, when Trump is used together with Donald, the system recognized them as one entity which is a <*PERSON*>. The inconsistency labelling in this dataset is considered as one of the main factor in the accuracy of the CRF model.



622	JAPAN	Xi	B	PROPN	BORG
623	JAPAN	and	O	CCONJ	O
624	JAPAN	Trump	B	PROPN	BORG

Figure 5. Incorrect label of entities.



42	JAPAN	President	O	PROPN	O
43	JAPAN	Donald	B	PROPN	BPERSON
44	JAPAN	Trump	I	PROPN	IPERSON

Figure 6. Example of correct labelling on entities.

5. Conclusion

In this paper, a model for named entity recognition classification for South China Sea conflicts is proposed by using a CRF classifier. This model helps in identifying which news articles are connected or related to the conflict in South China Sea and to extract any relevant information in the immense amount of data. The proposed method may also be a way of improving conflict detection in a wide form of text. Besides that, this method take advantage of each named entity and handle them separately which results in better performance.

Python is one of the best language for machine learning. It is a great object-oriented, interpreted, and interactive programming language. The existence of modules such as spaCy and nltk, classes, exceptions, very high level dynamic data types, and dynamic testing makes it preferable as a tool for machine learning compared to other languages. Although the processing is slower than other language, its data handling capacity outdone other languages.

Acknowledgement

We would like to thank Ministry of Higher Education (MoHE) and Universiti Teknologi Malaysia for funding this work under vot number (05G73).

References

[1] Buszynski, L., The South China Sea: Oil, Maritime Claims, and U.S.–China Strategic Rivalry. *The Washington Quarterly*, 2012. 35(2): p. 139-156.

[2] Nadeau, D. and S. Sekine, A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 2007. 30(1): p. 3-26.

[3] Màrquez, L. and H. Rodríguez., Part-of-speech tagging using decision trees. *European Conference on Machine Learning*. 1998. Springer.

[4] Liu, X., Recognizing named entities in tweets. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: *Human Language Technologies*-Vol. 1. 2011. Association for Computational Linguistics.

[5] Minkov, E., R.C. Wang, and W.W. Cohen., Extracting personal names from email: Applying named entity recognition to informal text. *Proceedings of human language technology conference and conference on empirical methods in natural language processing*. 2005.

[6] Shabat, H.A. and N. Omar., Named entity recognition in crime news documents using classifiers combination. *Middle-East Journal of Scientific Research*, 2015. 23(6): p. 1215-1221.

[7] Rish, I., An empirical study of the naive Bayes classifier. *IJCAI 2001 workshop on empirical methods in artificial intelligence*. 2001.

[8] Scholkopf, B. and A.J. Smola, Learning with kernels: support vector machines, regularization, optimization, and beyond. 2001: *MIT press*.

[9] Cover, T. M. and P. Hart, Nearest neighbor pattern classification. *IEEE transactions on information theory*, 1967. 13(1): p. 21-27.

[10] Cruz, B. M. D., Named-Entity Recognition for Disaster Related Filipino News Articles. *IEEE Conference TENCON 2018-2018*. 2018.

[11] Bankoff, G., Cultures of disaster: Society and natural hazard in the Philippines. 2003: *Routledge*.

[12] Gorla, S., et al. Named Entity Recognition for Telugu News Articles using Naïve Bayes Classifier. *NewsIR@ ECIR*. 2018.

[13] Liu, X. and M. Zhou, Two-stage NER for tweets with clustering. *Information Processing & Management*, 2013. 49(1): p. 264-273.