UTM UNIVERSITI TEKNOLOGI MALAYSIA
INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

South China Sea Conflicts Classification Using Named Entity Recognition (NER) and Part-of-Speech (POS) Tagging

Nur Rafeeqkha Sulaiman & Maheyzah Md Siraj
School of Computing
Faculty of Engineering
81310 UTM Johor Bahru, Johor, Malaysia
Email: feeqkha@gmail.com; maheyzah@utm.my

Submitted: 20/01/2020. Revised edition: 30/04/2020. Accepted: 1/05/2020. Published online: 20/05/2020
DOI: https://doi.org/10.11113/ijic.v10n1.255

***Abstract*—Internet connects everyone to everything globally. The existence of Internet eases people in completing daily tasks. Thanks to Internet, information is being digitalized and spread openly to the public. 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 data and to identify only useful information manually takes a longer time to complete. Extracting important features from a text document can be done by using one or a combination of feature extraction methods. Relevant information and the classification of news articles in relation to the conflicts in South China Sea need to be done. In this paper, a model is proposed to use Named Entity Recognition (NER) that search for and classifies important information regarding to the conflicts. In order to do that, a combination of Part-of-Speech (POS) and NER are needed to extract type of conflicts from the news. This study also claims to classify news by using Conditional Random Field (CRF) algorithm and Multinomial Naïve Bayes (MNB) as classification methods by training and testing the data.**

***Keywords*—Named Entity Recognition, Part-of-Speech, Machine Learning, text classification**

I. INTRODUCTION

South China Sea is a part of the Pacific Ocean which covers the sea from the Karimata and Malacca straits to the Strait of Taiwan. The sea area of the sea is around 3,500,00 square kilometers and carries various strategic importance, thus become the main reason for dispute of territorial claims on South China Sea. Originally, the South China Sea was a calm area of sea since ancient time. Until recently, in the late twentieth century where it had become an abundant fishing ground for the fishermen from China and other countries, and one-third of the world's shipping passes through it, which carries around USD 3 trillion in trade per year. Moreover, a vast reserve of oil and gas are to be believed to lie beneath its seabed [1]. Due to the power of oil and natural gas, several Asian countries fought over the islands in South China Sea [2].

A method on extracting South China Sea conflicts information can be used to collect unstructured data and turn them into a structured or corpus, in order to analyze, correlate, and visualize the information into meaningful knowledge. One of the methods to extract and analyze data is Named Entity Recognition (NER).

The term "Named Entity" is now widely used in Natural Language Processing (NLP). Researchers realized that Information Extraction (IE) tasks where location names, and numeric expressions of structured information extracted from unstructured text such as news articles are very important. Identifying references to these entities in text was recognized as one of the important subtask of IE and is known as "Named Entity Recognition" [3]. Generally, NER is understood as the task of classifying information units like a person, countries, organizations and locations [3] and mainly focus on formal texts such as news articles due to the easier identification of texts or sentences compare to informal texts such as e-mail and tweets.

This study aims to classify types of conflicts based on data collected from news articles around the world. In order to achieve this, a model with a combination of NER and Part-of-Speech (POS) tagging is used to extract and identify the types of conflicts existed in South China Sea. POS is known as assigning or marking each word of a text the proper morphosyntactic tag in its context of appearance [4]. There are many different libraries for POS taggers and each tagger has their own way of defining text. In English, there are many categories and subcategories of POS and POS tagger helps in identifying them more accurately. Both POS and NER are two different taggers in NLP where NER tags a chunk of named entity while POS tags each word individually in a text. Therefore, combining these two taggers are impossible to imagine at first. This is where Inside-Outside-Beginning (IOB) scheme. IOB helps in combining POS and NER so that although a chunk is separated, IOB helps in identifying a separated chunk by applying a I, O, or B in the beginning of a label.

Text Classification is the act of classifying or allocating text document to the correct classes based on techniques in machine learning. It is generally done on the foundation of significant words or extracted feature from documents [5]. For this paper, an investigation on two classifiers are done in order to find out which classifiers are more suitable with NER and POS. The two classifiers used are Conditional Random Field (CRF) algorithm and Multinomial Naïve Bayes (MNB) algorithm. Lastly, the performance of accuracy of both classifiers are measured to find out which classifiers are more accurate when combined with NER and POS.

The remainder of this paper is organized as follows. In Section II, the related work on NER and POS is summarized. Section III explains in detail of the proposed methodology of this study. Every phases in this research is discussed alongside the data and their outputs. In section IV, results and analysis are discussed and explained. In the final section, this paper is concluded with outlines on some potential research directions in the future.

II. 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 [6] and emails [7] 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, the overview of the related work done by past researchers in extracting information from different domains and is extracted into named entities which come from unstructured news or data.

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 behavior 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 [8]. The text classification algorithm used were Naïve Bayes [9], Support Vector Machine [10], and K-Nearest Neighbor classifiers [11]. 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 analyzed through standard evaluation namely: Precision, Recall, F-Measure, and Macro-average (F1).

A recent study done by [12] 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 [13]. 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 embedding) 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.

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 [14]. 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 center 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 [15] 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 labeler is used; and in the second stage, pre-labelled tweets were clustered and cluster level labelling using and enhanced CRF-based labeler that employs cross-tweet information was conducted. Just like tweets, emails include in informal texts categories. A study done by [7] 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.

III. The Proposed Methodology

A. Phase 1: Crawl and Pre-Process News Articles on South China Sea Conflicts

Phase 1 consists of data collection, crawling, and data pre-processing. The data collection is the process where datasets from online English resources are collected and built the datasets were collected from: The Star Online, The Straits Time, Kyodo News, VietNam News, and Vietnam News Agency (VNanet). The dataset collected will be used in Phase 2 for information extraction and classification by using NER and POS tagging techniques.

In order to download online news resources, a crawler is used. The crawler is created in python by using python modules urlib and beautifulsoup4. The data is collected based on South China Sea conflicts keywords. However, for the non-conflicts dataset, the articles were randomly crawled based on Sports and Entertainment categories. Urlib4 module fetches URLs using a variety of protocols in a very simple interface. Beautifulsoup4 is a python library for pulling data out of HTML and XML files. Beautifulsoup4 takes a little time to collect data and save hours or days of work. Unnecessary data object like url links, images and other non-text subjects were excluded from crawling and only text was taken out for pre-processing.

Pre-processing phase is the process of converting textual data into a data-mining ready structure, where the most important text-features that help to distinguish between text-categories are identified. The main goal of pre-processing is to represent each document as a feature vector to distinct the text into specific words. In this stage, websites contents are extracted into text document. The importance of this stage is to ease the process of NER and POS tagging in Phase 2. To do so, converting the web page into text document is a must and HTML parsing and word stopping are done.

One of the process in pre-processing is stopping process which eliminates stop-words such as ‘and’ and ‘or’. Before classifying the dataset, the data need to be represented in a form that is understandable by classifiers. A correct representation of the dataset helps in increasing the accuracy and speed up the process. Text representation is a process where grouping of the related and relevant keywords regarding types of conflicts.

B. Phase 2: To Select and Classify Data Based on Named Entity Recognition (NER) and Part-of-Speech (POS) Tagging

In second phase, the building of South China Sea conflicts corpus using NER technique after the pre-processing is done. NER tagging is done by using Python. NER tagging or entity extraction is a popular technique used in information to identify and segment the named entities and classify them under various predefined classes. For this research, annotation using WebAnno annotation tool was used. WebAnno is an online tool that allows users to annotate text documents according to feature attributes related to types of conflicts. Once annotation is done, the data will be converted in CoNNL-2002 format.

POS tagging technique is added as another technique in order to get conflicts related attributes. This will then lead to build South China Sea conflicts attributes corpus building using the enhanced NER with POS technique. For this project, Python and spaCy is again used for POS tagging. SpaCy uses POS tags based on Penn Treebank. The code loaded a document from a directory and the output is written in .csv files for easy viewing in tables and columns. IOB tagging is done in order to tag tokens in a chunking task. During annotation with WebAnno, the tagged sentences or words are automatically listed according to IOB schemes.

The result obtained in Phase 1 will be used in this phase as a continuation to get the desirable output for classification. As a result, the Named Entity containing all conflicts attributes was obtained with the help of combination of both techniques. Once the attributes have been extracted, the data

will be used for testing, training and classification.

The main goal of text classification is to decide on a label for a text based on how alike it is to other text that is already labelled. The dataset will be transformed into feature vectors in pre-processing process and a model is built by using machine learning in classification process. In this phase, CRF model and Naïve Bayes will be used as classifiers for the dataset. Implementing CRF and Naïve Bayes to the datasets with the results obtained from tagging process. This includes NER, POS and IOB. From the data that have been extracted, testing and training data will be divided. In this project, 67% of the data will be used as training data while 33% of the data will be used as testing data.

In classification phase, CRF and Multinomial Naïve Bayes will be applied separately to classify type of conflicts from the online news. Both classifiers will use training data and once classifiers have learnt from the trained data, testing data is passed to the classifiers to classify conflicts based on the trained data. The data will then be categorized into its own classes. Once classification is done, the performance of classification is evaluated in order to observe the performance of the algorithm. This is the last sub-phase in phase 2. The results from classification process will be evaluated in phase 3.

C. Phase 3: Performance Evaluation

Performance evaluation is used to evaluate or measure the performance of the classification. This phase will find the accuracy of classifying conflicts. The results will be presented in the form of graphs, tables and figures. The results will be evaluated for accuracy, precision, recall and F1-score. Table I shows the terminologies used in performance evaluation.

TABLE I. TERMINOLOGIES USED IN PERFORMANCE EVALUATION

Terminology	Definition
True Positive (TP)	Collection of the data that are correctly identified as the number of particular class assigned.
True Negative (TN)	Collection of data that are correctly identified as negative class.
False Positive (FP)	Collection of the data that are incorrectly identified and being rejected by the class assigned and is a member of some other class.
False Negative (FN)	Collection of the data that are incorrectly identified and being rejected by the class assigned and is a member of some other class.

Accuracy is the fraction of the classification result that are correct.

$$accuracy = \frac{TP + TN}{(TP + TN + FP + FN)}$$

Precision is the fraction of the predicted documents in a class that are correct.

$$precision = \frac{TP}{(TP + FN)}$$

Recall is the fraction of documents in a class that correctly predicted.

$$recall = \frac{TP}{(TP + FN)}$$

F1-score is a weighted harmonic mean of precision and recall.

$$F1\ score = \frac{2 \times precision \times recall}{precission + recall}$$

D. Dataset

This study uses datasets from 5 English online news websites. The Star Online is an online news website based in Malaysia. The Straits Time is an English language daily broadsheet newspaper based in Singapore currently owned by Singapore Press Holdings. Kyodo News is a non-profit cooperative news agency based in Japan. Their online news site is in multiple languages and for this research, English articles were downloaded from this website. VietNam News and Vietnam News Agency (VNanet) are both online news websites based in Vietnam. Table II shows a description of the dataset labels. Overall, 225 news articles were crawled from these websites. From these articles, five type of conflicts is extracted: <ECONOMICAL>, <LAW>, <RESOURCE>, <MILITARY>, and <TERRITORIAL>.

TABLE II. DESCRIPTION OF THE DATASET LABELS

Attributes (Type of conflicts)	Details	Example
Military	Conflicts that involves military activities, weapons, military vehicles.	Warships, bombing activities, military build-up
Resource	Conflicts that involves resources located in South China Sea.	Oil and gas development, reefs, banks, fishing.
Law	Conflicts that is related to law or agreements involving South China Sea.	Code of Conduct, United Nations Convention for the Law of the Sea
Territorial	Territories, islands, and location claimed by claimant countries.	Senkaku island, Palawan island, Paracel island, Spratly island
Economical	Conflicts involving financial and economy produced through South China Sea.	Loans, tradewar, shipment, trade route

In this datasets containing 11 named entities excluding <O> token, there will be a sum of 130517 which also equals to 130517 annotated words from 225 news articles. The separation of the total of named entities based on the dataset is shown in Table III.

TABLE III. TOTAL COUNT OF NAMED ENTITIES

Named Entity	Count
O	127238
BMILITARY	1055
IMILITARY	100
BRESOURCE	127
IRESOURCE	5
BLAW	310

Named Entity	Count
ILAW	657
BTERRITORIAL	223
ITERRITORIAL	144
BECONOMICAL	467
IECONOMICAL	191

During the annotation process, reading through each news articles is crucial in order to determine and differentiate between conflicts and non-conflicts news. In the initial phase of data collections, different news websites are accessed in order to retrieve South China Sea conflicts related news. As the websites are from different countries, some news articles seem to be reported differently based on their country. A country tends to be bias when reporting conflicts and tend to take sides in the news. Some articles were explained in detail while some articles give only a short explanation on the topic.

Once annotation is done, the datasets are tokenized in order to map the word token with the entity type. A total of 11 entity types been generated from IOB encoding after being annotated and export to the conll text format. The 11 entity types are:

- O: Outside/Other
- BMILITARY: Beginning of Military entity
- IMILITARY: Inside of Military entity
- BLAW: Beginning of Law entity
- ILAW: Inside of Law entity
- BECONOMICAL: Beginning of Economical entity
- IECONOMICAL: Inside of Economical entity
- BRESOURCE: Beginning or Resource entity
- IRESOURCE: Inside of Resource entity
- BTERRITORIAL: Beginning of Territorial entity
- ITERRITORIAL: Inside of Territorial entity

Part-of-Speech (POS) tagging is used in tagging the news articles with English grammar and vocabulary in order to form the POS tag encoded to the dataset. Example of POS tags are nouns, verbs, conjunction, adverbs, and etc. POS tagger is one of famous tagger that many researchers used nowadays to tag texts in their documents. POS tagger also have been built and developed by many organizations and education institutes to aid in their researchers' projects. For this project, Python and spaCy is again used for POS tagging. SpaCy uses POS tags based on Penn Treebank.

The right source code need to be used for the token in order to get the right annotated articles with NER and POS tagging. The code loaded a document from a directory and the output is written in .csv files for easy viewing in tables and columns.

E. Training of the Model: CRF

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, 67% of the datasets is used for training while the remaining 33% 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 C1 can 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.

F. Training of the Model: Multinomial Naive Bayes

For this classification, a Python is used as a tool for Multinomial Naïve Bayes classifier. Just like CRF model, 33% of the dataset will be used as testing data while 67% of the dataset will be used as training data for Multinomial Naïve Bayes. 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 Multinomial Naïve Bayes model as it finds the average weighted by the support number from each labels. The experiment was done by using three different alpha values namely: 0.25, 0.5, and 1.0.

IV. Result and Analysis

A. Train Model

This section explains the result obtained based on the implementation of NER technique and use of CRF and Multinomial Naïve Bayes models in training and testing the data. Weighted overage of F1-score was taken for accuracy values. The reason both of CRF parameters are set to 0.1 is to see the performance of the model to the datasets, same goes with Multinomial Naïve Bayes.

F1-score is used to evaluate the performance as it is interpreted as a weighted average of 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 and Multinomial Naïve Bayes model. It is stated that the higher the F1-score, the higher the accuracy of the 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.

There is no "O" tagged tokens. This is because "O" tokens do not belong in any type of conflicts, therefore "O" tagged tokens were omitted to eliminate noise data. Based on the classification report, there are three different classification

values which are:

Micro average: Calculation of the metrics globally by counting the total true positives, false negatives and false positives.

Macro average: Calculation of the metrics for each label, and find their unweighted mean. In this case, it does not take label imbalance into account.

Weighted average: Calculation of the metrics for each label and find their average weighted by support which means the number of true instances for each label. In this case, it alters the 'macro' to account for label imbalance, and resulting in an F1-score that is not between the precision and recall value.

After training process is done, the performance of classification is evaluated in order to observe the performance of the algorithm. Weighted average is used in this algorithm to show the accuracy of the model to the datasets and it has been initialized in the source code.

Comparison of two classifiers on training data. For CRF, weighted average value of Precision is 0.96, Recall at 0.84, and 0.89 for F1-score. Table IV shows the value of precision, recall, and F1-score for CRF. Micro average values for three metrics differs slightly but are still in high value, but it would not be used for this project as it only takes counts of TP, TN and FN. Macro average value is produced from the unweighted weight value from each label of NE and it does not care if any imbalance in the dataset occurs. This is why weighted average is being used to find accuracy of CRF and Multinomial Naïve Bayes model as it gives the average weighted by the support number from each labels.

TABLE IV. CRF: VALUE OF PRECISION, RECALL AND F1-SCORE

Evaluation Metrics	Precision	Recall	F1-score
Weighted Average	0.96	0.84	0.89
Micro Average	0.97	0.84	0.90
Macro Average	0.82	0.72	0.75

For Multinomial Naïve Bayes, weighted average value of Precision is 0.54, Recall at 0.72, and 0.57 for F1-score. Table V shows the values of precision, recall, and F1-score for Multinomial Naïve Bayes. Table V shows the value of precision, recall, and F1-score for Multinomial Naïve Bayes. Micro average values for three metrics differs slightly except for Recall and are still not much difference with weighted average, but it would not be used for this project as it only takes counts of TP, TN and FN. Macro average value is produced from the unweighted weight value from each label of NE and it does not care if any imbalance in the dataset occurs.

TABLE V. MULTINOMIAL NAÏVE BAYES: VALUE OF PRECISION, RECALL AND F1-SCORE

Evaluation Metrics	Precision	Recall	F1-score
Weighted Average	0.54	0.72	0.57
Micro Average	0.40	0.72	0.51
Macro Average	0.48	0.60	0.48

For both model, "IRESOURCE" tag gives 0.00 values to all three metrics. This is because there is lack of data or tokens with "IRESOURCE" tags, therefore, there is not enough "IRESOURCE" tagged data for testing and training performance evaluation to be produced.

B. Test Model

Once the model is trained, it will be tested to evaluate the F1-score. As mentioned, 33% of the dataset will be used as testing data while 67% of the dataset will be used as training data. 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 CRF model testing, 3 tests were done by applying 3 different C1 values. The values are 1.0, 10, and 15.

Once the model is trained, it will be tested to evaluate the F1-score. Just like CRF model, 33% of the dataset will be used as testing data while 67% of the dataset will be used as training data for Multinomial Naïve Bayes. 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 Multinomial Naïve Bayes model as it finds the average weighted by the support number from each labels. For Multinomial Naïve Bayes model testing, 3 tests were done by applying 3 different Alpha values: 0.25, 0.5, and 1.0.

C. Performance Evaluation

In this section, the performance evaluation of classifications on NER and POS by using two classifications techniques are discussed. Both classifications were done in Phase 2. Once the reports are acquired, the performances are evaluated in this phase. The performance was evaluated by finding their accuracy, precision, recall, and F1-score.

1) CRF

Based on Fig. 1, when C1 value was increased in a large number, the values of Recall and F1-score decreased dramatically and differs by -+0.4 in value. Precision values decreased only by -+0.2. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. High precision relates to the low false positive rate. However, in all three classification results, there are some entity with no value or 0.00. As C1 value increases, number of classes with 0.00 values increases. C1 value shrinks the less important feature's coefficient to zero. Thus, removing some features altogether. Higher C1 values is more suitable for large dataset as it removes less important features to increase accuracy of model. This is an example of ill-defined entity due to the high value of C1 which might affect the value of these entities. In these reports, the classes removed are

"IRESOURCES", "IMILITARY", and "BRESOURCES". Table VI shows the accuracy or weighted average of F1-score CRF classification based on different C1 value.

TABLE VI. CRF: VALUE OF ACCURACY

Value	Accuracy
C1: 1.0	0.80
C1: 10	0.47
C1: 15	0.39

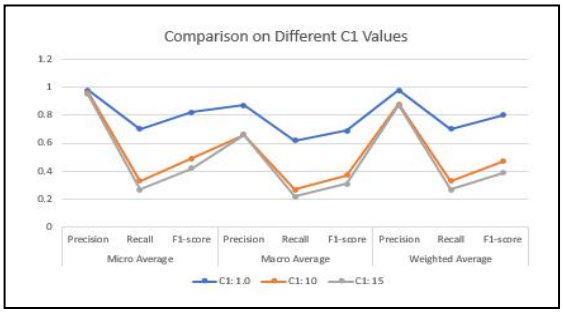


Fig. 1. Comparison of different C1 values

2) Multinomial Naïve Bayes (MNB)

Based on Fig. 2, when Alpha value was increased, the values of Recall and F1-score decreased slightly and differs by -+0.1 in value. Precision values however increases when Alpha is 0.5 and decreases when Alpha value is 1.0. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. High precision relates to the low false positive rate. However, in all three classification results, there are many entities with no value or 0.00. As Alpha value increases, number of classes with 0.00 values increases. Alpha or Laplace smoothing is a technique for smoothing data. It is a type of shrinkage estimator where the frequency-based probability might introduce zeros when multiplying the probabilities, leading to a failure in preserving the information contributed by the non-zero probabilities. Therefore, a smoothing approach, for example, the Laplace smoothing, must be adopted to counter this problem. However, from this study, we have learnt that the higher the Alpha value, the more features it shrinks. Higher Alpha values is more suitable for large dataset as it removes less important features to increase accuracy of model. Table VII shows the accuracy of Multinomial Naïve Bayes based on different Alpha values.

TABLE VII. CRF: VALUE OF ACCURACY

Value	Accuracy
Alpha: 0.25	0.59
Alpha: 0.5	0.57
Alpha: 1.0	0.46

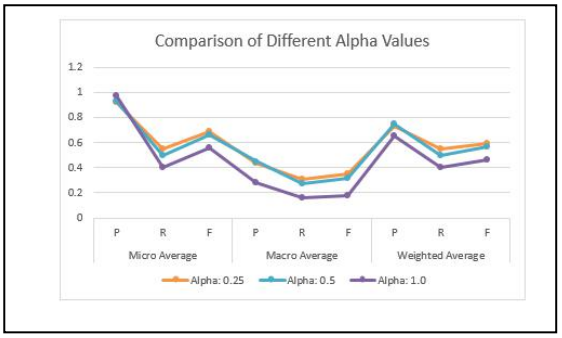


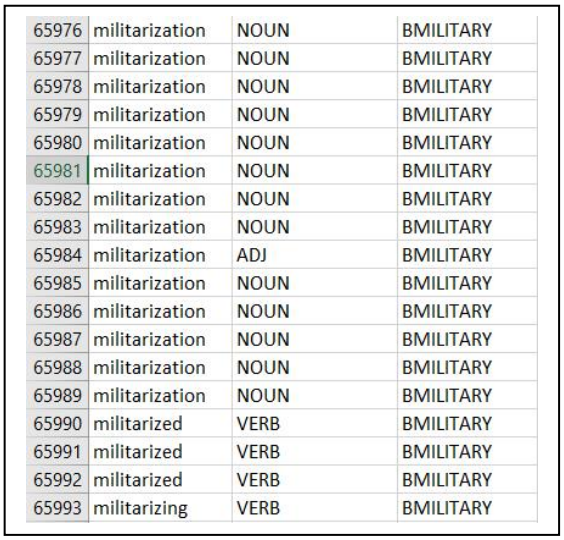
Fig. 2. Comparison of different alpha values

D. Discussion and Analysis on Experimental Results

You must submit the IJIC Electronic Copyright Form as in his study, the classification report is analysed for the ability of classifiers to classify type of conflicts. Weighted average of F1-score was used in order to find the accuracy the higher the value of accuracy, the better the performance. Based on the results, observations have been made and are listed below:

- CRF achieves the highest accuracy measurement than Multinomial Naïve Bayes.
- Accuracy of CRF decreases when C1 value increases; accuracy of Multinomial Naïve Bayes decreases when Alpha value decreases.
- When C1 value and Alpha value increase, the number of classes with no values increases. This proves that both C1 and Alpha shrinks the number of unimportant features in order to increase accuracy. However, features shrinking does not work for experiments with small data set.

Based on this analysis, NER and POS works better with CRF. This is due to CRF taking into account contextual information and state of the neighbors affect the predictions. Multinomial Naïve Bayes however seems to not work well with multi label classifier, NER and POS. As for the both systems, 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.



65976	militarization	NOUN	BMILITARY
65977	militarization	NOUN	BMILITARY
65978	militarization	NOUN	BMILITARY
65979	militarization	NOUN	BMILITARY
65980	militarization	NOUN	BMILITARY
65981	militarization	NOUN	BMILITARY
65982	militarization	NOUN	BMILITARY
65983	militarization	NOUN	BMILITARY
65984	militarization	ADJ	BMILITARY
65985	militarization	NOUN	BMILITARY
65986	militarization	NOUN	BMILITARY
65987	militarization	NOUN	BMILITARY
65988	militarization	NOUN	BMILITARY
65989	militarization	NOUN	BMILITARY
65990	militarized	VERB	BMILITARY
65991	militarized	VERB	BMILITARY
65992	militarized	VERB	BMILITARY
65993	militarizing	VERB	BMILITARY

Fig. 3. Inconsistency of POS tagging

One issue to be taken into consideration on the dataset is the inconsistency of POS and NER tagging. Due to the varieties of conflicts terms used as keywords, the named entity is being tagged inconsistently probability of inconsistency in NER tagging 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, inconsistency in POS tagging in this study. Based on Fig. 3, "military" is labelled as "ADJ", "NOUN" and "VERB". Since CRF takes POS into account while doing classification, the difference in POS tagging might decrease the accuracy of classification. The inconsistency labelling in this dataset is considered as one of the main factor in the accuracy of the CRF model.

V. Conclusion

There are various previous works done for NER and POS tagging classifications. Most of the researchers focused on the classification methods and feature extraction techniques. Therefore, the main goal of this project is to use NER and POS tagging techniques on two different classification techniques: CRF and MNB. The initial result of the first phase until classification with CRF is done. Each of the classifiers were evaluated using accuracy, precision, recall, and F1 score.

The feature extraction method in this study are NER and POS. NER and POS work great together as NER locate and classify Named Entities (NE) mentions in unstructured text into pre-defined categories based on our preferences. For this study, five named entities were chosen: "ECONOMICAL", "LAW", "RESOURCE", "TERRITORIAL", and "MILITARY". POS classifying words or texts into their part of speech or word classes. For classification, two classifiers were compared namely CRF and MNB. Based on Discussion analysis, CRF gives better result than as CRF takes into account the value of its neighbours during classification.

Acknowledgment

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

References

[1] Gao, Z. and B. B. Jia. (2013). The Nine-dash Line in the South China Sea: History, Status, and Implications. *American Journal of International Law*, 107(1): 98-123.

[2] Rowan, J. P. (2005). The US-Japan Security Alliance, ASEAN, and the South China Sea Dispute. *Asian Survey*, 45(3): 414-436.

[3] Nadeau, D. and S. Sekine. (2007). A Survey of Named Entity Recognition and Classification. *Lingvisticae Investigationes*, 30(1): 3-26.

[4] Màrquez, L. and H. Rodríguez. (1998). Part-of-speech Tagging Using Decision Trees. *European Conference on Machine Learning*. Springer.

[5] Dalal, M. K. and M. A. Zaveri. (2011). Automatic Text Classification: A Technical Review. *International Journal of Computer Applications,* 28(2): 37-40.

[6] Liu, X., *et al.* (2011). Recognizing Named Entities in Tweets. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics.

[7] Minkov, E., R. C. Wang, and W. W. Cohen. (2005). 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*.

[8] Shabat, H. A. and N. Omar. (2015). Named Entity Recognition in Crime News Documents Using Classifiers Combination. *Middle-East Journal of Scientific Research,* 23(6): 1215-1221.

[9] Rish, I. 2001. An Empirical Study of the Naive Bayes Classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*.

[10] Scholkopf, B. and A. J. Smola. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.

[11] Cover, T. M. and P. Hart. 1967. Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory,* 13(1): 21-27.

[12] Cruz, B. M. D., *et al*. (2018). Named-entity Recognition for Disaster Related Filipino News Articles. *TENCON 2018-2018 IEEE Region 10 Conference*. IEEE.

[13] Bankoff, G. (2003). *Cultures of Disaster: Society and Natural Hazard in the Philippines*. Routledge.

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

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