Classifying Short Text in Social Media for Extracting Valuable Ideas

Chan-udom Apichai[1], Karman Chan[2], and Yoshimi Suzuki[1]

[1] Graduate Faculty of Interdisciplinary Research,
University of Yamanashi, Kofu, Japan 400–8511
{g15dm001,ysuzuki}@yamanashi.ac.jp
[2] IIJ Innovation Institute Fujimi, Chiyoda-ku, Tokyo, Japan
chan@iij-ii.co.jp

Abstract. Social media data such as Twitter becomes a huge number of data because of people around the world post frequently. Moreover, there are many kinds of tweets, even if these tweets mention the specific event. Some of them give very useful ideas and inspiration. How can we find such useful posts from a huge amount of social media data? In order to deal with this problem, we classify Thai tweets of the incident (Thai cave rescue, in June and July 2018) into four types. In this paper, we describe how to classify Thai tweet data of the specific event. The experimental data were collected from Twitter data with a specific hashtag. Even if tweets have the same hashtag, there are many types, such as suggesting solutions, emotional tweets, news reports, and others. We conducted experiments to classify tweets into four types using five machine learning algorithms. In addition, we compared tweets written in Thai language and tweets written in Japanese.

Keywords: Multi-label Classifying Short Text · Social Media Text Classification · Classification of Thai Tweets · Classification of Japanese Tweets

1 Introduction

Twitter is a social media, people can tweet or re-tweet or follow, what is happening at any moment a time, anywhere in the world that is an effective way to spread out information and express opinions or their feelings. Nowadays, the massive volume of tweets has become an interesting source for studying in various aspects. Text from social media provides a set of challenges, informal language, spelling errors, abbreviations, unknown word, and special characters and, emotional symbol (emoji, emoticon). Thai language written format is an agglutinative language. Therefore it is difficult to separate each word. Moreover, Twitter merely limited the number of character of a tweet, that is free writing formatted which users can tweet easily, there are many ill-formed sentences. Consequently, it is difficult to segment words. The different tokenizer obtains different words and also different meaning that effect to different experimental results like such the example shown as Fig 1.

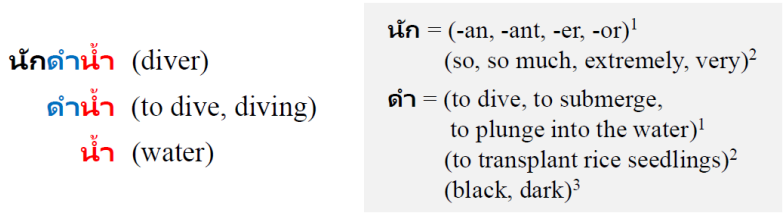

Fig. 1. Example of challenging in Thai words tokenizing

Some of them give very useful ideas and inspiration. How to extract valuable information? How can we find useful posts from a huge amount of social media data?

In this paper, we describe how to classify the multi-label Thai tweet data of the specific event. The experimental data were collected from Twitter data with a specific hashtag. Besides, there are many kinds of tweets, even if tweet data has the same hashtag (specific event), such as suggesting solutions, emotional tweets, news reports, and others.

At the incident: "Thai Cave Rescue" that happened in 2018 June-July, people from many countries tweeted and prayed for the safety of the boys trapped in the cave. We conducted experiments to classify tweets into four types using five machine learning algorithms.

We handle experiments to comparative classifiers to know which classifier is suitable for classifying short text in Thai. Additional, we compared the results using tweets written in Thai and the results using tweets written in Japanese. Our research contributions are the followings:

- Multi-label classification for the incident tweets in Thai language and Japanese,
- Comparison of the classification results differences due to the corpus size which is used for producing word embedding vector.

As a result of the experiment, when we used Linear SVM, we obtained the best results. Moreover, we compared classification results under considering the size of the corpus which was used for generating word embedding vector, and the number of vocabulary.

2 Related Work

There are many research papers for short text classification. For example, Alsmadi et al. proposed a term weighting scheme for short-text classification that called the supervised weight scheme [1] which obtained higher performance than traditional weighting i.e. term frequency (TF), binary-weight, term-frequency-inverse document frequency (TF-IDF). Dhingra et al. proposed character-based distributed representation for a tweet [4] which using a recurrent neural network (Bi-GRU) to predict hashtag of each tweet and also generate a vector of each

tweet at the same time. In terms of the Thai language, Nomponkrang et al. presented a comparison study of classification algorithms for Thai-sentence [10]. P. Sarakit et al. contributed the classifying emotion in Thai YouTube comments [14], that show the comparative results of SVM, Decision Tree and Multinomial Naive Bayes. To due with unbalanced data classifying, W. Wunnasri et al. proposed the approach for solving unbalanced data for Thai tweets sentiment classification [18].

There are many researchers have proposed extraction knowledge or semantic information from twitter data like as; A. Panasyuk et al. proposed extraction of semantic activities from twitter data that describes some results utilize tweets to determine events [11], A. Sogaard et al. presented using frame semantics for knowledge extraction from Twitter, based on syntactic and semantic parsing[15], P. Teufl and S. Kraxberger proposed extracting semantic knowledge from Twitter that using semantic patterns to highlight the semantic knowledge extraction for tweets related to a specific topic[16], A. Bifet and E. Frank presented sentiment knowledge discovery in Twitter streaming data using classifiers for opinion mining and sentiment analysis, and deal unbalanced classes with the sliding window Kappa statistic [2].

3 Tweet Classification

We found that specific incident tweets (Thai cave rescue) can be classified into the following four types.

- Suggesting solutions
- Emotional tweets
- News Reports
- Other tweets

In this study, We do experimental short text classification, the tweets written in Thai language, and also Japanese tweets on the same incident for comparative.

3.1 Word Embedding

We used two different word embedding models for comparison study. The first model is from the site: "Pre-trained word vectors of 30+ languages" [12]. The second model that we trained word2vec tool with Wikipedia data. These models are provided by word2vec [8], [9]. The features of the models are as shown in Table 1.

3.2 Classification Methods

The classifying short-texts(tweets) in our experiments, which compare the performance of several classification algorithms, for classifying our experimental twitter data (Thai cave rescue incident) into four classes (i.e. S,E,R and O). We

Table 1. The features of the two Thai word vector models

Features	First model ("Pre-trained word vectors of 30+ languages")	Second model ("Wikipedia data")
Corpus size	696MB	557MB
Vocabulary size	30,225	148,650
Vector size	300	300

employed the Python Scikit-learn package[3] [13] that based on most literature recently, to consider the Gaussian Naive Bayes classifier, Bernoulli naive Bayes classifier[17], Linear Support Vector Machine classifier [5], linear models with stochastic gradient descent (SGD) learning [19], and Passive Aggressive Classifier [3]. Our short-text classification approach shows in Fig 2 we employ a word embedding (word vector) to represent a tweet vector and then feed to the input of classifier.

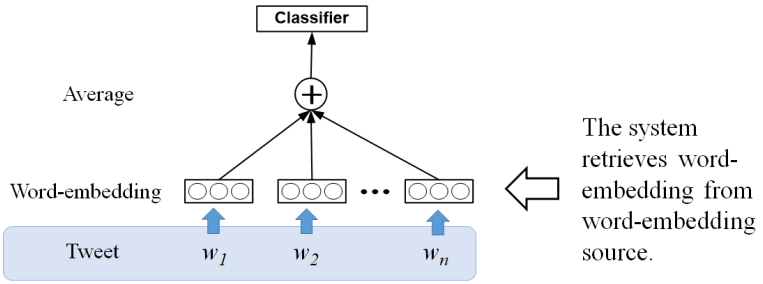

Fig. 2. Our approach for short-texts (tweets) classification

4 Experiments

We conducted experiments about short-texts (tweets) classification on specific incident twitter data, written in Thai language and Japanese. The objective of this study is for extracting valuable ideas or semantic information such as How to rescue the boys and their coach from deep inside the cave, How many days the miners trapped underground in Peru can stay without food and water? or How can the boys get clean-water inside the cave? , etc, and the incident timeline.

[3] http://scikit-learn.org/

4.1 Experimental Setup

As the experimental data, we selected tweets about "Thai cave rescue (Jun-July 2018)" in Thai language and Japanese. Besides, we have separated follow by dates and removed duplicate tweets. Table 2 illustrates the quantities of tweets written in Thai language and Japanese.

Table 2. Quantities of tweets about "Thai Cave Rescue"

Date	Thai	Japanese
July 2	2	3
July 3	10	21
July 4	20	20
July 5	24	17
July 6	38	6
July 7	69	5
July 8	8,512	1,065
July 9	7,820	865
July 10	9,823	915
July 11	5,706	909
July 12	1,311	416
# of tweets	33,335	4,242

Although Twitter users in Thailand are fewer than Japanese users (Thai: 12M users, Japanese: 40M users), the number of Thai tweets about "Thai cave rescue" is larger than the number of Japanese tweets. For tweet classification, we decide to classify the tweets into four classes that show in Table 3 as below.

Table 3. Four classes of tweets

Class	Description
Solution	tweets which include how to escape from the cave
Emotion	tweets which include user's emotion
Report	tweets which include news report
Others	other tweets

The training and testing data for our experiments, we labeled 3,060 tweets written in Thai language. Table 4 shows the distribution of labeled tweets.

Most of the tweets consist of more than one part. For example, many people tweet their own emotion after quoting a part of news reports. Therefore we conducted multi-label classification against Thai language tweets like as the example shows in Fig 3.

Table 4. Distribution of labeled Thai tweets

Class	Number of tweets
Solution	276
Emotion	1,682
Report	1,285
Others	559
Total	3,060

Label	Tweet
S, E	Elon musk สร้างแคปซูลจากชิ้นส่วนจรวด Falcon เพื่อช่วยทีมหมูป่าออกจากถ้ำ น่ายกย่องคนพูดจริงทำจริงขอให้ใช้จริงแล้วเด็กออกมาอ ... Elon musk created a capsule from the Falcon rocket piece to help the wild pig team out of the cave, praising the real person.
R, E	ศอร. แถลงช่วยทีมหมูป่าออกจากถ้ำเพิ่ม 4 คน - นายกฯ ให้กำลังใจเจ้าหน้าที่ทีมช่วย หมูป่า CRES, informed that helping the boar team out of the cave for another 4 people - the Prime Minister encouraged the team to help the boar.
S	นี่คือสิ่งที่อีลอนมาสค์เสนอ Inflatable Tube ให้นักประดาน้ำลากท่อไนล่อนขนาด เส้น ผ่าศูนย์กลาง 1 เมตร ผ่านจุดที่อันตรายที่สุด จ ... This is what Elon Musk offers Inflatable Tube to the diver to drag the nylon tube in diameter 1 meter through the most dangerous spots.

Fig. 3. The example of labeled tweets

4.2 Experimental Results

Each experiment was conducted with 10 fold cross-validation. For classifying short-texts(tweets), we employed five kinds of machine learning algorithms: Linear SVM, two kinds of Naive Bayes (Gaussian and Bernoulli), Stochastic Gradient Descent and Passive Aggressive Classifier.

Table 5 and Table 6 illustrates the results of the experiment using the word embedding model from the large corpus and from the small corpus, respectively. In each table, SVM, NB-G, NB-B, SGD, and PAC means Linear SVM, Naive Bayes (Gaussian), Naive Bayes (Bernoulli), Stochastic Gradient Descent, and Passive Aggressive Classifier, respectively. P, R, F1, and AC means Precision, Recall, F-score, and Accuracy, respectively.

Table 5. Results (corpus size: 696MB, vocab: 30,225)

Method	avg P	avg R	avg F1	AC
SVM	0.675	0.637	0.655	0.831
NB-G	0.485	0.722	0.580	0.702
NB-B	0.559	0.689	0.618	0.776
SGD	0.658	0.587	0.620	0.804
PAC	0.660	0.590	0.623	0.798

Table 6. Results (corpus size: 557 MB, vocab: 148,650)

Method	avg P	avg R	avg F1	AC
SVM	0.766	0.591	0.688	0.846
NB-G	0.466	0.763	0.579	0.669
NB-B	0.592	0.716	0.648	0.799
SGD	0.754	0.595	0.665	0.843
PAC	0.707	0.580	0.637	0.804

In both experiments, we obtained the best results when we used Linear SVM. The results in Table 6 are better than the results in Table 5. It is because the results in Table 6 used the word embedding model which bigger vocabulary than the results in Table 5.

5 Discussion

5.1 Error Analysis

The experimental results of Linear SVM classifier indicate on Table 7 and Table 8. We found that it is difficult to classify 'Emotion'. That's because most of the emotional tweets are too short to analyze. Recall of 'Solution' is not so high. It seems that even if there are some new solutions in the test data, it is difficult to classify the tweets into 'Solution'.

Table 7. The results by Linear SVM (corpus size: 696MB, vocab: 30,225)

Class	P	R	F1	AC
Solution	0.96	0.95	0.95	0.91
Emotion	0.78	0.80	0.79	0.75
Report	0.82	0.84	0.83	0.79
Others	0.53	0.48	0.50	0.84
average	0.77	0.77	0.77	0.82

Table 8. The results by Linear SVM (corpus size: 557MB, vocab: 148,650)

Class	P	R	F1	AC
Solution	0.75	0.41	0.52	0.93
Emotion	0.80	0.81	0.80	0.78
Report	0.79	0.76	0.77	0.81
Others	0.73	0.39	0.50	0.86
average	0.766	0.591	0.668	0.846

5.2 Comparison between Thai and Japanese Tweets

We compared the classification results of Thai tweets and that of Japanese tweets. Table 9 illustrates the average number of words per tweet written in Thai language and Japanese. In the pre-process about word tokenization for Thai we utilize Deepcut tokenizer [6] and Mecab tokenizer for Japanese.

Table 9. The average number of words per a tweet written in Thai language and Japanese

Statistics of dataset	Thai	Japanese
# of tweets	3,060	3,000
# of vocabulary	6,565	6,720
# of words(token)	66,152	110,782
Average # of words(token) per tweet	21.62	36.93

The data show Japanese tweets are longer than Thai tweets. For classification experiments, we labeled 3,000 tweets written in Japanese (multi-label). From our observation, we found Japanese tend to use quotation in their tweets. Table 10 shows the distribution of labeled tweets.

Table 10. Distribution of labeled Japanese tweets

Class	Number of tweets
Solution	185
Emotion	911
Report	2,460
Others	294
Total	3,000

Table 12 illustrates the results against Japanese tweets by the same methods. We used two different word embedding vectors. The first model is from the site: "Pre-trained word vectors of 30+ languages" [12]. The second model is made with Wikipedia data. These models are provided by word2vec [8], [9]. For morphological analysis of Japanese tweets, we used MeCab [7]. The features of the models are as shown in Table 11. The second model is made with larger corpus than the first mode.

Table 12 and Table 13 illustrates the results of the experiment using the word embedding vector from the small corpus and from the large corpus, respectively. In each table, SVM, NB-G, NB-B, SGD, and PAC means Linear SVM, Naive Bayes (Gaussian), Naive Bayes (Bernoulli), Stochastic Gradient Descent, and Passive Aggressive Classifier, respectively. P, R, F1, and AC means Precision, Recall, F-measure, and Accuracy, respectively.

In both experiments, we obtained good results when we used Linear SVM, Stochastic Gradient Descent and Passive Aggressive Classifier. Moreover, we

Table 11. The features of the two Japanese models

Features	First model ("Pre-trained word vectors of 30+ languages")	Second model ("Wikipedia data")
Corpus size	1 GB	3.3 GB
Vocabulary size	50,108	519,275
Vector size	300	300

Table 12. Results (corpus: 1GB, vocab: 50,108)

Method	avg P	avg R	avg F1	AC
SVM	0.730	0.705	0.717	0.886
NB-G	0.551	0.796	0.651	0.788
NB-B	0.574	0.756	0.653	0.809
SGD	0.737	0.685	0.701	0.877
PAC	0.776	0.682	0.726	0.872

Table 13. Results (corpus: 3.3GB, vocab: 519,275)

Method	avg P	avg R	avg F1	AC
SVM	0.866	0.688	0.767	0.909
NB-G	0.569	0.782	0.659	0.818
NB-B	0.614	0.752	0.676	0.828
SGD	0.860	0.665	0.750	0.900
PAC	0.829	0.689	0.752	0.881

found the results of Japanese tweets are better than the results of Thai tweets. It is because in the experiments of Japanese tweets we used word embedding vectors from the big corpus.

Table 15, shows the results by linear SVM using the model made from the big corpus.

Table 14. Results by Linear SVM (corpus: 1GB, vocab: 50,108)

Class	P	R	F1	AC
Solution	0.65	0.59	0.61	0.96
Emotion	0.72	0.70	0.71	0.83
Report	0.61	0.56	0.56	0.85
Others	0.53	0.52	0.51	0.91
average	0.63	0.59	0.60	0.89

Table 15. Results by Linear SVM (corpus: 3.3GB, vocab: 519,275)

Class	P	R	F1	AC
Solution	0.90	0.66	0.74	0.97
Emotion	0.79	0.69	0.73	0.85
Report	0.91	0.95	0.93	0.88
Others	0.86	0.46	0.57	0.94
average	0.866	0.688	0.767	0.909

We found that it is difficult to classify 'Emotion' from tweet data. Some of 'Emotion' tweets are classified as 'News Report'. It is because many Japanese users quote news reports even if they would like to mention their emotion.

6 Conclusion

We conducted classification experiments for specific incident twitter data using five kinds of machine learning algorithms. When users often tweet their feelings, and their ideas after quoting news articles. Therefore, it is difficult to improve accuracy with the method using only the frequency of words that was conventionally done. Moreover, we compared the classifying results of Thai tweets and Japanese tweets against on the same incident; "Thai cave rescue" which occurred from June to July 2018. In the experiments of Japanese tweets classification, we obtained better results than Thai tweets classification results. The possible reasons are as follows:

- The accuracy of morphological analysis,
- amount of corpus for word embedding creating,
- the number of words per tweet.

From experiments, we extracted many tweets which mentioned some ideas for rescuing children from inside the deep cave: e.g., drainage by air pumps, mini-submarine, air pumps & tubes and so on.

References

1. Alsmadi, I., Hoon, G.K.: Term weighting scheme for short-text classification: Twitter corpuses. In: Neural Computing and Applications. pp. 1–13 (2018)
2. Bifet, A., Frank, E.: Sentiment knowledge discovery in twitter streaming data. In: Pfahringer B., Holmes G., Hoffmann A. (eds) Discovery Science. DS 2010. Lecture Notes in Computer Science, vol 6332. Springer, Berlin, Heidelberg. pp. 1–15 (2010)
3. Crammer, K., Dekel, O., Keshat, J., Shalev-Shwartz, S., Singer, Y.: Online passive-aggressive algorithms. Journal of Machine Learning Research **7**, 551–585 (2006)
4. Dhingra, B., Zhou, Z., Fitzpatrick, D., Muehl, M., Cohen, W.: Tweet2vec: Character-based distributed representations for social media. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 269–274. Association for Computational Linguistics, Berlin, Germany (August 2016), http://anthology.aclweb.org/P16-2044

5. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: A library for large linear classification. Journal of Machine Learning Research **9**, 1871–1874 (2008)
6. Kittinaradorn, R.: Deepcut: A thai word tokenization library using deep neural network (2017), https://github.com/rkcosmos/deepcut
7. Kudo, T., Yamamoto, K., Matsumoto, Y.: Applying conditional random fields to japanese morphological analysis. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. pp. 230–237 (2004)
8. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Proceedings of Workshop at ICLR (2013)
9. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems **26**, 3111–3119 (2013)
10. Nomponkrang, T., Sanrach, C.: The comparison of algorithms for thai-sentence classification. International Journal of Information and Education Technology **6**, 801–808 (2016)
11. Panasyuk, A., Blasch, E., Kase, S.E., Bowman, L.: Extraction of semantic activities from twitter data. In: Proceedings of STIDS. pp. 79–86 (2013)
12. Park, K.: Pre-trained word vectors of 30+ languages (2017), https://github.com/Kyubyong/wordvector
13. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in python. Journal of Machine Learning Research **12**, 28252830 (2011)
14. Sarakit, P., Theeramunkong, T., Haruechaiyasak, C., Okumura, M.: Classifying emotion in thai youtube comments. In: Proceedings of the 6th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES) (2015)
15. Sogaard, A., Plank, B., Alonso, H.M.: Using frame semantics for knowledge extraction from twitter. In: Proceedings of the Twenty-Ninth AAAI conference on Artificial Intelligence. pp. 2447–2452 (2015)
16. Teufl, P., Kraxberger, S.: Extracting semantic knowledge from twitter. In: Tambouris E., Macintosh A., de Bruijn H. (eds) Electronic Participation. ePart 2011. Lecture Notes in Computer Science, vol 6847. Springer, Berlin, Heidelberg. pp. 48–59 (2011)
17. V. Metsis, I.A., Paliouras, G.: Spam filtering with naive bayes which naive bayes? In: 3rd Conf. on Email and Anti-Spam (CEAS) (2006)
18. Wunnasri, W., Theeramunkong, T., Haruechaiyasak, C.: Solving unbalanced data for thai sentiment. In: Proceedings of the 10th International Joint Conference on Computer Science and Software Engineering (JCSSE). pp. 200–205 (2013)
19. Zhang, T.: Solving large scale linear prediction problems using stochastic gradient descent algorithms. In: In Proceedings of ICML (2004)