LUPINE PUBLISHERS

Journal Of Anthropological And Archaeological Sciences

DOI: 10.32474/JAAS.2021.04.000189

ISSN: 2690-5752

Review Article

Thai Language Tweet Emotion Prediction Based on Use of Emojis

Anocha Rugchatjaroen*

National Science and Technology Development Agency, NECTEC, Thailand

***Corresponding author:** Anocha Rugchatjaroen, National Science and Technology Development Agency, NECTEC, Thailand

Received: June 24, 2021 **Published:** July 06, 2021

Abstract

Thai Language can be handled/considered in the same group of Chinese and Japanese where no explicit spaces exist between words. This article presents a work on the emotional identification of tweets based on the use of emojis which focusses on a Thai language context. The use of emojis in user tweets indicates the writer's emotions. The first phase of this study was to collect Thai tweets, clean them, and then to make a primary classification of the emojis into groups using K-nearest [1]. These group clusters are used as target outputs for the prediction of emoji classes. It was found that 22 is the appropriate K for considering 70 emojis for a collected set of tweets. The corpus includes any level of Thai language usage, which means that the processed data can consist of suffixes, slang, and unknown word from tokenization process. The vector representation advances the unknown accent. In sum, this research created a corpus of short messages collected from Twitter which were grouped into 22 emoji- classes. The corpus includes 7,825,857 messages prepared for classification based on emotions by applying 2 biLSTM layers. A table of emojis is proposed based on Ekman's six basic emotions: anger, disgust, fear, joy, sadness, and surprise were evaluated in both objective and subjective tests. The results show that word vectors work well for the classification of emotions through the use of emojis.

Introduction

The expression of emotions in a tweet can indicate social sentiments about a product, a person, or an organization. Twitter sentiment has been widely analyzed and predicted. One of the researches was proposed by A. Agarwal et al. in 2011. They presented their polarity results (positive, neutral and negative) of twitter data using POS analysis (Agarwal, et al. 2011). There are several algorithms which can be applied [2]. Applying such methods to Thai language needs fundamental word segmentation because Thai does not mark word boundaries (Haruechaiyasak and Kongthon 2013). C. Haruechaiyasak et al. worked on word segmentation and published an analysis of their results based on sentiment expressed in the domain of hotel reviews in 2010, and then twitter in 2013. In 2019, K. Pasupa et al. compared the results of a Thai sentiment analysis based on 1,115 sentences of Thai childrens' tales using deep learning techniques [3]. A set of pictograms/pictographs or ideographs called emojis are widely used to express the author's emotions or are inserted as inline objects. Some research has found that the emojis can represent a user's intentions by analyzing the sentiments and emotions of the surrounding message (Mohammad 2012, Felbo, et al. 2017, Jaouad, et al. 2019). This paper focuses on those which represent emotions and seeks to find the relationship between the emoji used and the emotion expressed in the context of a tweet in Thai. Since, an emoji is a single encoded-character which can be used to express the user's emotion, agreement, or sarcasm [4]. There has also been some research on emoji prediction, identification also translation.

The Corpus

The corpus consists of the emojis used in Thai tweets. It contains 7,825,857 messages based on the use of

a total of 70 emojis: 

Corpus characteristics

Thais use Twitter as one of the short messaging services for social networking in a maximum of 280 characters as a microblog. The research corpus contains 128,235,131 ML-segmented words using a hybrid algorithm (Haruechaiyasak and Kongthon 2013). After removing segmentation errors and stop words, a tweet contains 1 to 237 words with an average of 21.74 words excluding

hashtags. The 10 most common words is 'พี่" (elder sis/bro) found 1,002,813 times [5]. In the Thai context, face emojis can appear in any position in a message, but are mostly located at the end (Horsuwan, et al. 2020, Tangtreerat and Sinthupinyo 2020).

The five most used emojis in the corpus are:

= 3,873,281 times

("Loudly CryingFace"),

- = 1,112,450 times

("FacewithTearsofoy"),

- = 641,171 times

("CryingFace"),

- = 629,688 times

("Smiling Face with Heart-Eyes"),

- = 585,938 times

("Flushed Face"), and

the least used is "Neutral Face which appears 43,614 times in the corpus.

The emoji usage in a tweet was used to be the label for the tweet text. The labelling was carried out after the results of the K-nearest classification, which will be explained in the next section. Thus, the corpus contains tweet texts, emoji multi-class labels, and the proposed corresponding emotions. The emotion labels will be explained in Section 4 [6]. In total, the corpus contains 7,825,857 messages with their generating labels. There are 37,348 unique ML-segmented words found. Each tweet contains 1 to 154 emojis with 2.54 on average fluctuating in a 0.44 range of standard deviation.

Emoji Recognition In Thai

Each emoji can be represented as a vector which is the same as the word2vec approach. K-nearest clustering is a process of clustering population into k groups. Figure 1 below shows the primary emoji clustering/group-labelling system. The k was varied from roughly sampled as 10, 20, 30 and 40 before narrowing down to a variation of [20, 21, ..., 30], after found the best results are 20 and 30. All results were manually evaluated by 3 Thai annotators. Finally, this research found that the number 22 was the most suitable for use with the collected data set. The flow in Figure 1 shows that the data was obtained by crawling ~7 million tweets using twitter API with requests for only those messages that contained at least one in the set of consideration emojis. This process then selects only distinct messages, then passed to the text pre- processing stage (Figure 1). This process then removes "RT", "URL", "@[mention]", "#[hashtag]", Escape character [\ ^$. | ? * + (), and word segments before converting them into a sequence of vectors using word2vec. Afterwards, the corresponding emojis for each sequence are also converted into emoji vectors to be used in the K-nearest classification, and got k=22 as described above. Then the tweets were multi-labelled after their appropriate cluster No [7-10]. A well-designed emoji predictor, called Deep Moji from MIT (Table 1) was used to predict relevant emojis from an input of tweets trained from the cleaned and vectorized 7.8m tweets. Two Keras bidirectional-LSTM layers work with an attention layer to predict classes. The results are shown in colours of cell in Table 2. The darker colours show the higher prediction accuracies. Italic numbers under each emoji set show the related number of tweets in the corpus.

Table 1: Seventy selected emojis divided into 22 clusters.

No.	Emoji	No.	Emoji	No.	Emoji
1		8		15	
	1,036,683		*132,972*		*46,233*
2		9		16	
	1,231,513		*195,819*		*1,280,253*
3		10		17	
	166,238		*1,065,657*		*74,823*
4		11		18	
	285,364		*132,842*		*2,391,063*
5		12		19	
	101,216		*6,593*		*324,260*
6		13		20	
	45,891		*25,474*		*68,049*
7		14		21	
	60,036		*359,236*		*26,218*
				22	
					31,712

Emotional Representations From The Clustered Emojis

The user's emotional expression in a short burst of inconsequential/consequential information of a tweet can be predicted by the content of the text and also the emoji used. Mohammad, Saif (2012) pointed out that emotions expressed in a text could be of benefit in a number of applications, such

as customer relations management or in determining the popularity of products (Mohammad 2012). Hence, he proposed a method to create an emotion- labelled tweet dataset from emotions expressed in hashtags - #anger, #disgust, #fear, #happy, #sadness, and #surprise [11,12]. This corpus is based on Ekman's six emotions according to the SemEval- 2007, which is a manually operated annotation of emotions in a news-paper headline corpus (Strapparava and Mihalcea 2007). Ekman's six emotions are also adopted and used in this research (Ekman 1992), Sadness, Happiness (Joy in this research), Fear, Disgust, Anger, Surprise, and an additional category of Neutral. All of the 22 emoji- classes were placed into 6 + 1 (neutral) emotions by 3 Thai annotators. As the previously trained DeepMoji system works with rule-based Ekman's clustered set, it can be used as an emotion prediction system for tweet messages.

Evaluation And Results

The evaluation was conducted both objectively and subjectively. The objective test was designed to evaluate the ML prediction results of the proposed emoji sets in terms of their relative emotions. The subjective test was to verify human perceptions of the emoji sets. Table 3 shows the objective test results. These results are derived from the training bidirectional LSTM models based on the architecture of Deep Moji using 100 embedding dimensions, 512 for each biLSTM layer, an Attention Weighted Average layer, and finally 22 for a multiclass classifier, called eval#1 [13]. The details of the proposed system are shown in Figures 1 & 2. The test was conducted in the bottom part of Figure 2. The output for each input of the Thai tweet texts is a predicted emotion from predicted emoji class. The proposed emoji members of each emotion and their number of appearances in the corpus are shown in italics in Table 2. The table also shows the emotion prediction accuracies in a histogram which uses light to dark and low to high colours. The percentages of true positive are 17.18% for Anger, 86.79% for Disgust, 7.79% for Fear, 76.39% for Joy, 13.67% for Neutral, 69.63% for Sadness, and 17.56% for Surprise. Another quick objective test was an emotion model using the 7 emotions as the targets for a multi-class classification model, which is called eval#2. This uses a general 1D Convolutional Neural Network (CNN) with kernel size=5, and a layer of Global Max Pooling1D [14] to learn the embedded sequences of tweet words, then to predict the corresponding emotion, which is the truth in the previous test. The results are shown in Table 4. These test results indicate that tweet words themselves, without emojis, could improve the objective accuracy. However, an emoji can have different meanings. It can emphasize or twist the tone of a tweet. Therefore, a subjective test was conducted to evaluate the human perceptions of a set of emoji tweets. The subjective test was a questionnaire, called eval#3, which asked 11 Thais aged 22-55 years to identify the emotions of the tweets. There are 2 identical lists of Ekman's emotions provided for two most closely selections. The results shown in Table 4. The ML prediction results derive from eval#1, which uses only sequences of trained word vectors as input with no emojis involved, while the human results are from using Thai texts with emojis [15]. The results in the table show high values, dark colours in diagonal, which could indicate a possible direct variation between ML and humans. Interestingly, these results could support the idea of establishing an emotion identification system by using word vectors in terms of raw text with corresponding emojis.

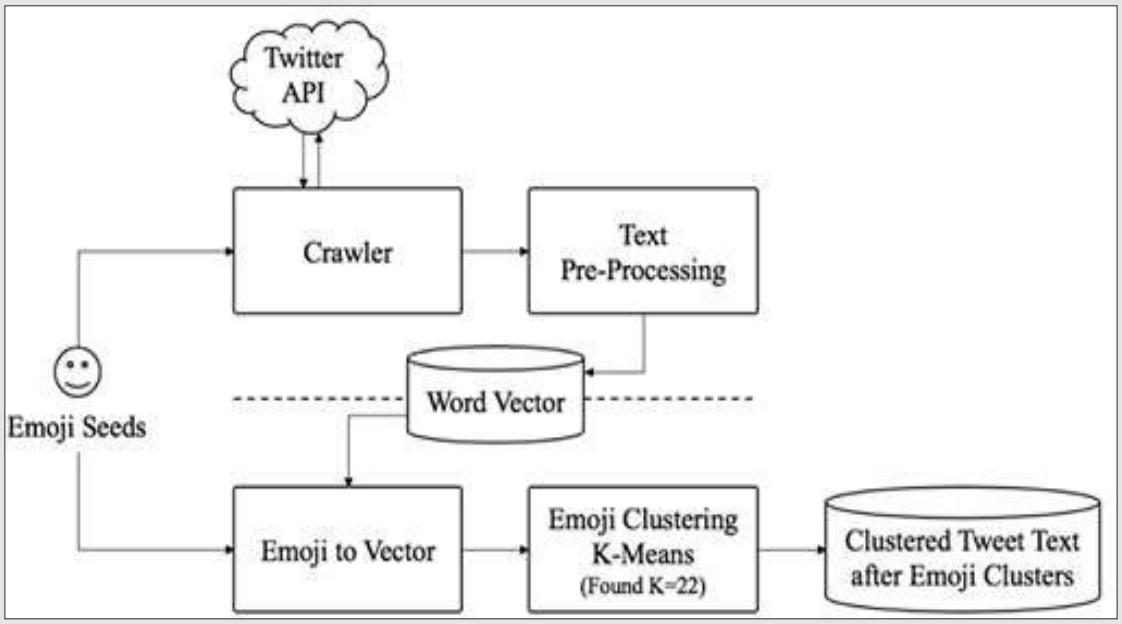

Figure 1: Primary emoji classification system for twitter data.

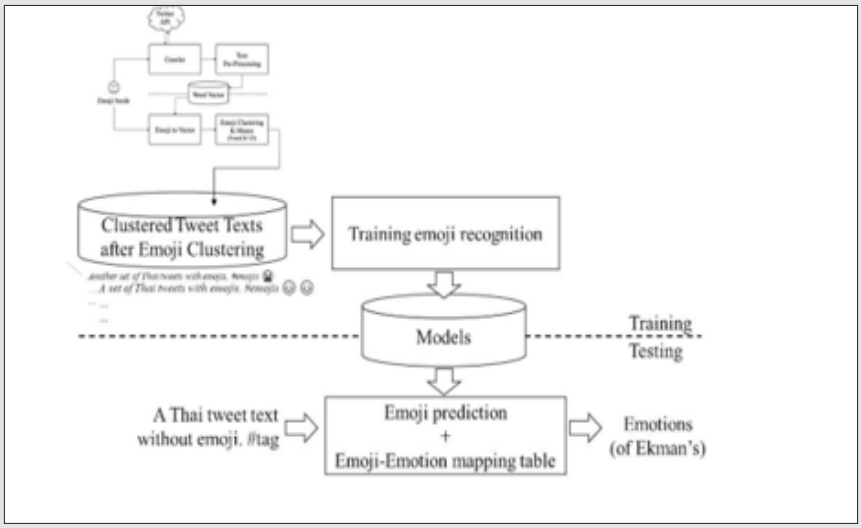

Figure 2: Emotion prediction training and testing process.

Table 2: Proposing emojis clustered into Ekman's basic emotions with number of emojis found in the corpus in italic.

Emotion	Emojis
Anger 0.16m	
Disgust 2.67m	
Fear 0.10m	
Joy *3.88m*	
Neutral *0.46m*	
Sadness *1.35m*	
Surprise *0.44m*	

Table 3: Truth table of normalized subjective test results. Columns indicate emotion answers from human perceptions can have 1 or 2 emotions per tweet and the rows indicate emotions from ML.

ML	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise
Anger	0.78	0.146	0.195	0.024	0.122	0.341	0.073
Disgust	0.073	0.439	0.098	0.22	0.293	0.488	0.098
Fear	0.024	0.146	0.585	0.195	0.171	0.146	0.463
Joy	0.049	0.098	0	1	0.317	0	0.122
Neutral	0.049	0.171	0.049	0.22	0.707	0.195	0.122
Sadness	0.146	0.171	0.171	0	0.244	0.927	0.049
Surprise	0.098	0.049	0	0.585	0.341	0.098	0.585

Table 4: Emotion prediction accuracy comparing the biLSTM+Emotion table and the CNN prediction.

	Accuracy
Emoji prediction using biLSTM+ Emotion table	74.49
Emotion prediction using CNN	99.86

Discussion and Conclusion

The highest subjective test results are Joy and Sadness. They represent the greatest variation between human perceptions and ML. These could be further used in sentiment polarity, which could support mass sensing in any market sensing product. The scores in Human Neutral column spread over ML emotion rows. This could mean that some levels of emotion could be identified as neutral. In this preliminary study, a Thai short expressive message corpus was created from Twitter. The emoji usage in them indicates emotions. A set of 22 emotional emoji groups used in a Thai context were formed by using K-nearest [16,17]. An analysis of these groups suggests that the emotions portrayed could be related to Deep Moji's architecture which could provide a possible list of emojis relating to a short text message input. Clustering the multi-label emoji groups according to Ekman's 6 basic emotions can be used to interpret the social emotional meaning of the message. The emojis in Table 2 are derived from the groups in Table 1. They are used in a final automatic social emotion detection system. It is a scheduled crawler for up-to-date Thai tweets and passes them to an emoji classification which leads to a group of emotions. Thus, a demo prototype called Emo Sense can be established. http://pop.ssense.in.th/EmoSense/.

References

1. Agarwal Apoorv, Ilia Vovsha Boyi Xie, Owen Rambow, Rebecca Passonneau (2011) Sentiment Analysis of Twitter Data. Proceedings of the Workshop on Language in social media (LSM 2011). Portland, Oregon Association for Computational Linguistics p: 30-38.
2. Chamlertwat Wilas, Pattarasinee Bhattarakosol, Tippakorn Rungkasiri, Choochart Haruechaiyasak (2012) Discovering Consumer Insight from Twitter via Sentiment Analysis. Journal of Universal Computer Science 18: 973-992.
3. Choochart Haruechaiyasak, Alisa Kongthon, Pornpimon Palingoon, Kanokorn Trakultaweekoon (2013) S-Sense: A sentiment analysis framework for social media sensing. Nagoya, Japan Asian Federation of Natural Language Processing.
4. Ekman Paul (1992) An Argument for Basic Emotions. Cognition and Emotion 6 (3/4) pp: 169-200.
5. Felbo Bjarke, Alan Mislove, Anders Søgaard, Iyad Rahwan, Sune Lehmann (2017) Using millions of emoji occurrences to learn any- domain representations for detecting sentiment, emotion and sarcasm. Conference on Empirical Methods in Natural Language Processing (EMNLP).
6. Haruechaiyasak Choochart, Alisa Kongthon (2013) LexToPlus: A thai lexeme tokenization and normalization tool. Nagoya, Japan Asian Federation of Natural Language Processing.
7. Horsuwan Thanapapas, Kasidis Kanwatchara, Peerapon Vateekul, Boonserm Kijsirikul (2020) A Comparative Study of Pretrained Language Models on Thai Social Text Categorization. the 12th Adian Conference on Intelligent Information and Database Systems 2020. Springer Nature Switzerland AG 2020 p. 63-75.
8. Jaouad, Muftisada, Mustapha Bassiri, Malika Tridane, Said Belaaouad (2019) Engineering Emotional Intelligence in Moroccan University. International Journal of Advanced Trends in Computer Science and Engineering 8(1.4): 288-293.
9. Kitsuchart Pasupa, Thititorn Seneewong Na Ayutthaya (2019) Thai sentiment analysis with deep learning techniques: A comparative study based on word embedding, POS-tag, and sentic features. Sustainable Cities and Society 50: 101615.
10. Kongthon Alisa, Sarawoot Kongyoung, Chatchawal Sangkeettrakarn, Choochart Haruechaiyasak (2010) Thailand's tourism information service based on semantic search and opinion mining. Pattaya: ITC-CSCC2010.
11. Medhat Walaa, Ahmed Hassan, Hoda Korashy (2014) Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal pp. 1093-1113.
12. Mohammad Saif M (2012) Emotional Tweets. the First Joint Conference on Lexical and Computational Semantics. Montreal: Association for Computational Linguistics pp. 246-255.

13. Netisopakul Ponrudee, Kitsuchart Pasupa, Rathawut Lertsuksakda (2017) Hypothesis Testing Based on Observation from Thai Sentiment Classification. Artificial Life and Robotics 22: 184-190.

14. Strapparava, Carlo, Rada Mihalcea (2007) Semeval- 2007 task 14: Affective text. SemEval-2007. Prague, Czech Republic p. 70-74.

15. Tangtreerat, Suppachai, Sukree Sinthupinyo (2020) Classification of Generation of Thai Facebook Users Using Deep Learning with Probability of Words. Recent Advances in Information and Communication Technology 2020. IC2IT 2020 Springer p. 49-59.

DOI: 10.32474/JAAS.2021.04.000189