Krathu-500: Post-Comments Thai Corpus

This paper was downloaded from TechRxiv (<https://www.techrxiv.org>).

LICENSE

CC BY 4.0

SUBMISSION DATE / POSTED DATE

15-12-2021 / 17-12-2021

CITATION

Taveekitworachai, Pittawat; Chan, Jonathan H. (2021): Krathu-500: Post-Comments Thai Corpus. TechRxiv. Preprint. <https://doi.org/10.36227/techrxiv.17205755.v1>

DOI

[10.36227/techrxiv.17205755.v1](https://doi.org/10.36227/techrxiv.17205755.v1)

Krathu-500: Post-Comments Thai Corpus

Pittawat Taveekitworachai
School of Information Technology
King Mongkut's University of Technology Thonburi
Bangkok, Thailand
pittawat.ta@mail.kmutt.ac.th

Jonathan Hoyin Chan
School of Information Technology
King Mongkut's University of Technology Thonburi
Bangkok, Thailand
jonathan@sit.kmutt.ac.th

Abstract— The Krathu-500 contains 574 Pantip posts title, post body with all comments of each post. The number of total comments is at 63,293 comments. The corpus provide Thai language used in real life situation with various context and types in conversational form. The corpus serves as a good way to improve capability of machine learning techniques that dealing with Thai language. Sentiment labeled smaller version of the comments dataset also provided with 6,306 records. The labeled corpus is human-annotated dataset with three labels for negative, neutral, and positive comments. The project also consists of open-source repository that allow any people who interested to modify and built on top of the current source code and dataset.

Keywords— Natural language, Data collection

I. INTRODUCTION

Internet forum, or web board, is an online discussion service usually hosted on website. Web board may divide into categories and subcategories which posts in the same topic belong to. It typically allows user to comment on a post.

Post and comments normally support multi-media attachments in the modern Internet forum. Language used on the forum usually different from the style of the language from the used in chats, news, article, or modern social media. Pantip is a very famous web board community in Thailand.

The Krathu-500 aims to provide the high-quality corpus consists of 574 Pantip posts with all comments of each post. The number of total comments is at 63,293 comments. It contains information of the original posts with text-only version associated with comments that belong that that post.

The name of the dataset came from the fact that Pantip only has around 500 posts that pass the criteria at the time of collection. Krathu (กระทู้) is Thai word for posts. Thus, the name of the dataset is Krathu-500.

The dataset is a part of conversational dataset and can be used to help improve any conversational tasks. In fact, many tasks can be considered as conversational tasks, e.g., question answering, chatbot, virtual assistant, dialogue systems. It is the fact that current state of Thai dataset is lacking, especially, the conversational corpus.

The labeled dataset built for the neural language inference task to help any solutions incorporate the sentiment analysis, or crowd emotion on public web board.

It is a smaller version of the Krathu-500 with 6,306 records labeled with three possible values, NEG for negative comments, NEU for neutral comments, and POS for positive comments.

The project also hopes to support Thai natural language processing community and people who are interested in Thai language with aiming to create conversational agents with more robust and natural response.

All of the dataset, source code of the crawler and example, and post-processing script along with the instructions can be found at <https://github.com/Pittawat2542/krathu-500>.

In this paper, it also presents the example of use cases which is post body summarization for each post and baseline models for the labeled dataset.

II. RELATED WORKS

This work was focused primarily on building a Thai conversational dataset based on the community web board. The well-know and similar work to this is the conversational dataset that composed of Reddit, OpenSubtitles, and AmazonQA by Henderson et. Al, 2019 [1].

The conversational dataset has variety of uses, e.g., improving the dialogue evaluation [2], emotion recognition [3], or use as dataset for study about personality and demographic of people on the Internet [4].

Chatbot, or conversational agent, is another aspect that utilize the conversational dataset. Digital transformation is one of the factors that make chatbot popular for many of its use cases [5].

However, given the importance of conversational dataset, the Thai conversational dataset is still lacking. And this work aims to provide a first step for that.

III. DATA COLLECTION

The crawler is a program that can automatically collect data across sites. The uses of web crawler are variety, ranging from data mining [6], search engine [7], and more. It will start at the targeted webpage and looking for a specific document object model (DOM) [8] element to collect information, and/or follow the hyperlink to another webpage until the criteria met and end the program. In the following subsections, there are details about the collection of posts metadata, post content, and post comments.

The metadata about time is provided to ensure the time-sensitive information is being used with caution. However, most of the data are not time sensitive and more related to the context of the post.

A. Posts list collection

On Pantip, there are many tags and rooms, comparable to category. One post can belong to one room but can have many tags. “มนุษย์เงินเดือน”, or salaryman tag had been chosen as the tag of posts that will be included in the dataset.

This decision was made because this is one of a popular tag that people usually discussed in variety topics and form. For example, people tend to discuss about problem in workplace, how to plan personal financial, reviewing the experience aboard. This means posts in this tag has variety of types of content and is a good representative of Thai language used in everyday life.

Not all posts in this tag should include in the dataset, since it required to collect all comments in posts, the post should be a post that popular among people and not newly created posts, so it can gain more diversity of comments. So, applying a filter called “คลังกระทู้โปรด”, or favorite posts, is a good idea to guarantee that each post collected has good engagement. Moreover, comment count also used as a filter by only select the post with more than 50 comments.

Python is a primary programming language to write a web crawler with framework called Selenium [9]. Once the web crawler done collecting, the result will be an array of post list contain metadata and saved as a comma-separated values which is the main file format for dataset in data science community.

Post list was collected on 18 October 2021 at 8:28 PM GMT+7.

The following flowchart shows the steps of the post crawler.

Fig. 1. Flowchart of post crawler

B. Post body and comments collection

After having a list of posts, next step is to collect post body and comments of each post. The web crawler starts by reading a list of a post and visit each post via URL, collect main post body, then expand all comments and collect all comments

except removed comments due to the violation of terms of use on Pantip and best comment, since it is duplicate to already existed comment. Both main post body and comments are collected only the text and ignore all of the multimedia attachments, for example, pictures, videos.

Please note that the algorithm, does not including collect the sub-comments, comments to each post's comment. Since the sub-comments usually very unrelated to the main post title.

Post body and comments also saved as comma-separated file.

Comments data was collected on 19 October 2021 at 6:44 PM GMT+7. Each comment also has timestamp on collection time.

The following flowchart shows the steps of the comment crawler.

Fig. 2. Flowchart of comment crawler

IV. DATA PROCESSING

Collected data require a little cleanup for better use [10]. The processing steps including, URL removal, and emoji removal.

The decision to not remove other languages apart from Thai language from the dataset was because in many circumstances, Thai people usually communicate with mix of Thai and other languages, especially English.

A. URL removal

URLs are not a text. It is meaning in terms of human language. So, removing it make dataset represents Thai text in human language perspective better.

B. Emoji removal

Emoji can contain meaning inside, but it is widely known to be vague and ambiguous. Most of the time, comments with only emoji can be considered as spam. So, removing it could help improve the quality of dataset, since it is better to have only text-focused comment for better understanding of Thai language.

Figure 3 shows the algorithm of processing of the string before save to the dataset.

Fig. 3. Flow chart of processing function

V. CORPUS DESIGN

The design of the corpus is inspired by the Pushshift.io-based Reddit corpus's conversational-level information and corpus-level information. [11]

There are two files combined as a complete dataset. The first file is the posts.csv containing list of posts. Other file is comments.csv containing post body and comments for each post.

A. posts.csv

In this file, it contains the post list with total of 574 posts. Each data records contain six data points.

- 1) **id** is a number represents identification number of Pantip post.
- 2) **title** is a string represent title of pantip post.
- 3) **url** is a string represent uniform resource locator link to the original post on Pantip.com.
- 4) **comment_count** is a number of comments excluding sub-comments, and removed comments.
- 5) **vote_count** is a number represents vote count for a post.
- 6) **published_at** is a date time that the post was published.

Table I provides the sample records from posts.csv.

TABLE I. SAMPLE RECORDS OF POSTS

<i>id</i>	<i>title</i>	<i>url</i>	<i>comment_count</i>	<i>vote_count</i>	<i>published_at</i>
38 59 73 54	ถึงคนที่ม่ทุก อย่างอย่างที ฝันไว้แล้ว.. ว่าจริงหรือ ที่ เงินมันซื้อ ความสุขได้	https://pan tip.com/ topic/ 38597354	126	3	2019-02-26 23:06:26 +06:42
41 04 15 62	เด็กจบใหม่ กับ ความเครียด จาก ครอบครัว และการหา งาน	https://pan tip.com/ topic/ 41041562	211	0	2021-10-15 01:21:22 +06:42
41 04 09 56	พนักงาน ออฟฟิศ สิ่ง ไหนที่ เจ้านายทำ แล้ว พนักงานรู้สึก แย่สุด	https://pan tip.com/ topic/ 41040956	275	0	2021-10-14 19:51:54 +06:42

Fig. 4. Sample records of the posts.csv

B. comments.csv

In this file, it contains the post body and comments associated with each post with total of 63,867 records separated into 574 post body and 63,293 comments. Each data records containing five data points.

- 1) **comment_id** is a number represents identification number of comment. The number before the post dash is the post's id and after the first dash is comment id. In case of post body the id will be post id follows by dash, and follows by 0.
- 2) **text** is a string represent either post body or comment.
- 3) **reply_to** is a number represent post id that a particular comment reply to. In case of post body, the value will not present (represents as None in Python.)

4) *published_at* is a date time that the comment was published.

5) *collected_at* is a date time that the comment was collected.

Table II provides the sample records from comments.csv.

TABLE II. SAMPLE RECORDS OF COMMENTS

comment_id	text	collected_at	published_at	reply_to
38597354-7817b4e7-6c30-414a-8145-c96864d38409	ข้อความรัก ไม่ได้ ถึงเบียร์ก็ได้ แต่คนไม่ จริงจัง	2021-10-19 16:42:42.03 4824	2019-02-26 23:09:18 +06:42	38597354
38597354-38e7901d-9402-445d-a235-99d2e772c641	ความสุข ของแต่ละ คนมันอยู่ที่ นิยามของ เจ้าตัวครับ	2021-10-19 16:42:42.12 5382	2019-02-27 04:34:28 +06:42	38597354
38597354-f2993436-6d00-420d-bfa1-ff0c6478f8e9	ความสุขอยู่ ที่ใจ	2021-10-19 16:42:42.19 3142	2019-02-27 07:10:51 +06:42	38597354

Fig. 5. Sample records of the comments.csv

By using *reply_to* in comments.csv, it could serve as a link to posts.csv, which in result is a full dataset.

VI. POST SUMMARIZATION

Natural language processing consists of many tasks, for example, question answering, language modeling, machine translation, text summarization. [12]

Fig. 6. Flow chart of demo program

To provide an example on the use of dataset, post summarization demo application was created. In this application, it provides a simple demonstration on how to load dataset, how to link two files together to form a complete dataset, and how to apply machine learning techniques for useful tasks.

The machine learning technique selected in the demo is natural language summarization provided by the *pythainlp* [13] package. It reduced the original post body to a summarized version that has length around 10% of original post body. The final output is a list of a pair of title of a post and summary on post body.

The following flowchart shows how the demo program works.

Table III provides the example of the result from the model.

TABLE III. DEMO RESULT EXAMPLE

Post Title	Post Body (Summarized)
ตอนนี้คุณมีเงินเก็บในชีวิตเท่าไรครับ	ไม่ได้มีเจตนาแง่ลบหรือความคิดไม่ดีนะครับ ถ้าบอกอาชีพการทำงานด้วยจะดีมากเลยนะครับ อยากทราบอายุและเงินเก็บของแต่ละคนครับ ผมอยากทราบว่าแต่ละคนส่วนมากแล้วมีเงินกันเท่าไร ขอไม่ดราม่า ไม่โกหกกันนะครับ ขอบคุณครับ
เงินเดือนเท่าไร ถึงจะมีลูกได้แบบไม่ต้องลำบากเรื่องเงินครับ	เงินเดือนของทั้งพ่อและแม่ที่เพียงพอจะดูแล คนๆหนึ่งให้มีชีวิตแบบกลางๆ เงินเดือนในที่นี้ หมายถึง ไม่ต้องหุรหุรหุรนะครับ
อายุ30กว่า ยังไม่มีเงินเก็บ ใช้ชีวิตแบบเดือนชนเดือน มีใครเป็นแบบนี้บ้างไหมคะ	มีใครเป็นแบบนี้บ้างไหมคะ?? อายุ30กว่าแล้ว ใช้ชีวิตแบบเดือนชนเดือน ยังไม่มีเงินเก็บ หรือไม่เคยอยู่ถึงสิ้นเดือน

Fig. 7. Result example of example application

VII. LABELED CORPUS

Labeled corpus is a human-annotated dataset based on sentiment of the comments or post body. The labels are POS for positive comment, NEU for neutral comment, and NEG for negative comment.

Positive comment is any comments that falls into the category of gratitude, or supportive message. While neutral comment is a comment that describe, explain, or contain both positive and negative message. And for the negative comment, it is a comment that express the sadness, anger, or hate.

The corpus that being annotated is a smaller version of full corpus (comments-small.csv). It contains the post body, and the first ten comments of each post in the dataset.

The structure of the dataset is like comments.csv with one additional field called *class label*. The added field has a type of string with three possible values (POS, NEU, NEG) to indicate the sentiment of each comment.

The dataset has 6,306 records divided into 5,521 records being a neutral comment. 400 records are negative comments. And 385 records are positive comments

Table IV provides the sample records from comments-small-labeled.csv.

TABLE IV. SAMPLE RECORDS OF LABELED COMMENTS

<i>comment_id</i>	<i>text</i>	<i>collected_at</i>	<i>published_at</i>	<i>reply_to</i>	<i>class_label</i>
38597354-38e7901d-9402-445d-a235-99d2e772c641	ความสุข ของแต่ละ คนมันอยู่ ที่นิยาม ของเจ้า ตัวครับ	2021-10-19 16:42:42. 125382	2019-02-27 04:34:28 +06:42	3859 7354	NEU
32247592-1e885bd1-266e-4eac-a30e-f9d363fd2ebf	ขอแต่ละ อย่าง .. ปวด กบาลได้ โล !!	2021-10-19 16:44:38. 641224	2014-06-27 10:55:27 +06:42	3224 7592	NEG
37341871-d7dd8246-651b-4e47-ad7b-7e8dba91a390	ขอบคุณ สำหรับ ประสบ การณ์ที่ มาแบ่ง ปันกัน ครับ	2021-10-19 17:04:02. 849067	2018-02-03 10:16:36 +06:42	3734 1871	POS

Fig. 8. Sample records of the comments-small-labeled.csv

Fig. 9. Layer architecture of LSTM model

VIII. BASELINE MODELS FOR LABELED CORPUS

There are three simple text classification models serves as a baseline. Each model trained on the 85% of labeled corpus

(training set) and evaluated on the 15% of the labeled corpus (testing set).

Three models are simple LSTM model, simple CNN model, and simple BERT model.

The architecture of LSTM model, CNN model, and BERT model is displayed in the figure 9, 10, and 11 respectively.

Fig. 10. Layer architecture of CNN model

Fig. 11. Layer architecture of BERT model

The performance of the baseline model shows in the Table V with accuracy and performance of each model on the testing set.

TABLE V. BASELINE MODEL PERFORMANCE

<i>Model Name</i>	<i>Accuracy</i>	<i>Loss</i>
LSTM Model	0.854	0.519
CNN Model	0.855	0.533
BERT Model	0.853	0.523

Fig. 12. Baseline model performance (accuracy and loss)

From the table, the performance of each model does not drastically difference. Every models has similar accuracy and loss. And the accuracy of each model is quite good at 85% with LSTM and CNN model trained at 10 epochs, and BERT model fine tuned at 16 epochs.

IX. CONCLUSION

In this paper, it provides the Thai dataset that data is very variety in topics and context. The posts collected are from many types even it is in the same tags. For example, question post, review post, discussion post. The paper also provide example on how to use the dataset.

Data is always the one of the important factors that could determine the success or failure of the project. Good and more variety of the dataset helps take the model to another step forward. With the new Thai dataset, it opens new possibilities and become another steppingstone for better models and datasets in the future.

This could serve as another dataset to improve Thai machine learning model, especially, in the field of natural language processing.

ACKNOWLEDGMENT

I would like to express my gratitude to IEEE for document template.

REFERENCES

- [1] Henderson, Matthew, Paweł Budzianowski, Inigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrkšić et al. "A repository of conversational datasets." arXiv preprint arXiv:1904.06472 (2019).
- [2] Sai, Ananya B., Akash Kumar Mohankumar, Siddhartha Arora, and Mitesh M. Khapra. "Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining." Transactions of the Association for Computational Linguistics 8 (2020): 810-827.
- [3] Zygadło, Artur, Marek Kozłowski, and Artur Janicki. "Text-Based Emotion Recognition in English and Polish for Therapeutic Chatbot." *Applied Sciences* 11, no. 21 (2021): 10146.
- [4] Vukojevic, Matej Gjurkovic Mladen Karan Iva, and Mihaela Bošnjak Jan Šnajder. "PANDORA Talks: Personality and Demographics on Reddit."
- [5] Völkle, Christiane, and Patrick Planing. "Digital automation of customer contact processes—an empirical research on customer acceptance of different Chatbot use-cases." In *Digitalen Wandel gestalten*, pp. 217-229. Springer Gabler, Wiesbaden, 2019.
- [6] M. Thelwall, "A web crawler design for data mining," *Journal of Information Science*, vol. 27, no. 5, pp. 319–325, 2001.
- [7] Peshave, Monica, and Kamyar Dezhgosha. "How search engines work: And a web crawler application." PhD diss., University of Illinois Springfield, 2005.
- [8] A. Ayes, "Document object model (DOM)," *Essential Dynamic HTML fast*, pp. 102–110, 2000.
- [9] Selenium-python.readthedocs.io. 2021. *Selenium with Python — Selenium Python Bindings 2 documentation*. [online] Available at: <<https://selenium-python.readthedocs.io>> [Accessed 20 October 2021].
- [10] M. Saravanan, P. C. Raj, and S. Raman, "Summarization and categorization of text data in high-level data cleaning for information retrieval," *Applied Artificial Intelligence*, vol. 17, no. 5-6, pp. 461–474, 2003.
- [11] "Reddit corpus (by subreddit)," *Reddit Corpus (by subreddit) - convokit 2.5.1 documentation*. [Online]. Available: <https://convokit.cornell.edu/documentation/subreddit.html>. [Accessed: 04-Nov-2021].
- [12] Wannaphong Phatthiyaphaibun, Korakot Chaovavanich, Charin Polpanumas, Arthit Suriyawongkul, Lalita Lowphansirikul, & Pattarawat Chormai. (2016, Jun 27). PyThaiNLP: Thai Natural Language Processing in Python. Zenodo. <http://doi.org/10.5281/zenodo.3519354>
- [13] Liddy, Elizabeth D. "Natural language processing." (2001).