Thai Social Media Trend Analysis using Big Data Feature on Cloud Computing.

Smith Simargool[1*], Pasd Putthapipat[2], Wutthikorn Threevithayanon[3], and Sudhiporn Patumtaewapibal[3]

[1]Vincent Mary School of Engineering, Assumption University, Samutprakarn, Thailand
[2]Digital Solutions – Analytic Business, True Digital Group, Bangkok, Thailand
[3]Faculty of Innovation and Technology, Panyapiwat Institute of Management, Bangkok, Thailand
*Corresponding author, E-mail: pasd.put@truedigital.com

Abstract

This work demonstrates a proof of concept framework for an automatic analysis solution for Thai Social Media using the collaborative framework and cloud solution. Different cloud services and frameworks have been used as the based component to build the framework with some additional development. This framework can improve the way in doing trend analysis or marketing research in term of manpower and time consumption. The use case of stock market news and discussion in the "PANTIP" is selected to demonstrate the work which is a collection of 33005 public comments related to "EARTH stock." The framework shows potential capability; however, it is still limited in terms of accuracy which could be improved.

Keywords: *Trend analysis, Big data, Cloud computing, Natural language, Marketing Research*

1. Introduction

In this era, people spend most of their time on the internet talking, discussing, and sharing opinions on social media and websites. So if this information can be used to understand what people are thinking and feeling, it will bring many benefits and is the reason for creating this project.

This project aims to extract specific data from the target website/web service and analyze it to determine whether people think about this data positively or negatively and the magnitude of that feeling. Previous works showed the strong capability of machine learning to both English (Sneha et al., 2017; Wang et al., 2016; Shahare, 2017) and Thai dataset (Sanguansat, 2016). However, an end-to-end system (Hashimoto, Aramvith, Chauksuvanit, & Shirota, 2014) which can automate the process in Thai is still quite limited and needed. This project shows the proof of concept framework which can be further developed as a solution platform that can measure the rate of success of the campaign and monitor the social media movement or event real-time trend analysis.

2. Objectives

The objective of this work is to design and demonstrate the Trend analysis framework for Thai social media context using the collaborative framework and Google APIs. The framework would scrape the data automatically from the source for certain topics the user picks. Then, the framework would analyze the sentiment and intention of the comments in those topics to measure the aggressiveness and trend direction. It would benefit the research in term of time and cost in data collection and data processing.

The end-to-end operation of the system is demonstrated and measured its performance using the case of stock market discussion, with this, "EARTH" stock topic from the Pantip website from 2015 to 2017 was selected.

3. Materials and Methods

3.1 Materials

This part is the explanation of the main components and the reason that they are adopted in this project.

3.1.1 Scrapy

Scrapy Plugins (Scrapy, 2019) is a free and open source web crawling framework, written in Python. Initially designed for web scraping, it can also be used to extract data using APIs or as a general purpose web crawler. Scrapy project architecture is built around 'spiders,' which are self-contained crawlers that are given a set of instructions.

In this project, Scrapy is used to command the spider to crawl the target website, which in this experiment case is Pantip, and extract the data according to the topic input.

3.1.2 Splash

Splash (Scrapy+Splash Plugins, 2019) is rendering service for JavaScript. It performs browser-like activity through an HTTP API. Also, it can perform a full asynchronous service.

In this project, Splash is used to help Scrapy extracting the data from JavaScript-based websites.

3.1.3 Google Cloud Translation API

Google Cloud Translation API (Google Cloud Translation API, 2019) can dynamically translate texts between thousands of language pairs. The Cloud Translation API lets websites and programs integrate with the translation service programmatically. The Google Translation API is a part of the larger Cloud Machine Learning API family.

In this project, Google Cloud Translation API is used to translate the data from Thai into English.

3.1.4 Google Cloud Natural Language API

Google Cloud Natural Language API (Google Cloud Natural Language, 2019) reveals the structure and meaning of the texts by offering powerful machine learning models in an easy-to-use REST API. Besides, it can be used to understand the sentiment about the products on social media or parse intent from customer conversations happening in a call center or a messaging app.

In this project, Google Cloud Natural Language API is used to analyze the data to determine the magnitude of positivity and negativity of people opinions in that data.

3.1.5 Google VM Instance

Google Compute Engine delivers virtual machines running (Google Virtual Machine Instance, 2019) in Google's innovative data centers and worldwide fiber network. The Compute Engine's tooling and workflow support enable scaling from single instances to global load-balanced cloud computing.

In this project, Google VM Instance is used as a server to host the website which acts as an interface for the users to put the topic input, receive output report, and run all the intended programs.

3.2 Methods

The flowchart shown below in Figure 1 is to show the system or how the program works roughly.

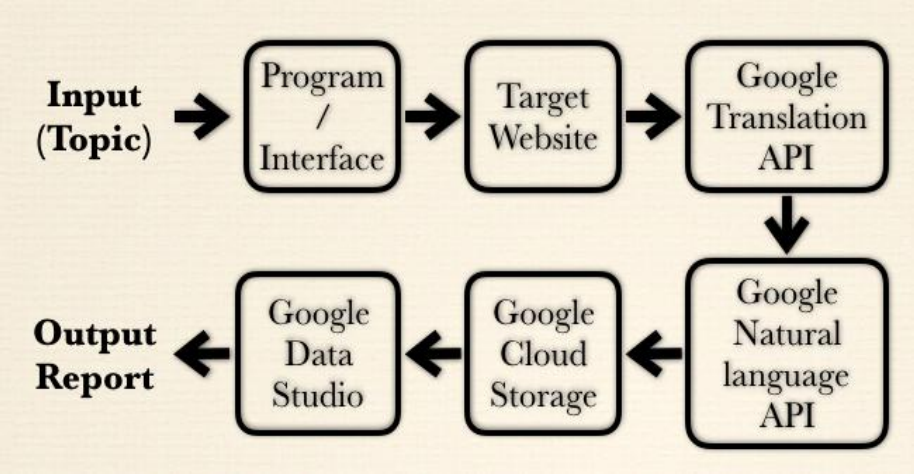


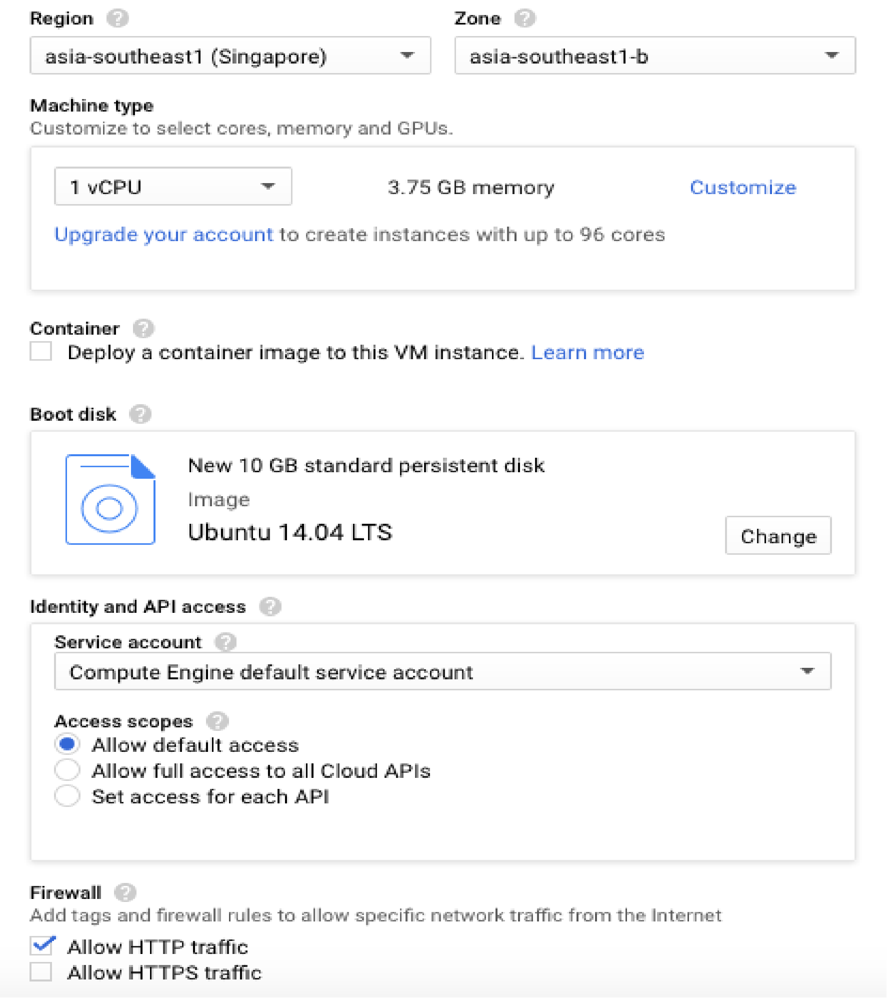
Figure 1 The flowchart describing workflow of the project from receiving inputs to creating output report

The program starts working when the users put a keyword or topic into the user interface. Then, the program scrapes related data from the Pantip website and saves it on the server. Next, the program processes to translate that information using Google Translation API and sends translated texts to Google Natural Language to analyze the magnitude of the texts which will be saved on Google Cloud Storage for Google Data Studio for the later process of generating the graph report.

The following part is the explanation of each working process of the system. However, for ease of understanding for readers, the explanation starts with the implementation of the project on Google VM instance along with the necessary programs and the purposes for using them. The rest of the explanation is the working process starting from receiving input from the users to showing the output graphs.

3.2.1 Implementation

This project is on cloud computing, so the server needed in this case uses Google VM instance with a specification as follows; 1 vCPU, 3.75 GB memory, Ubuntu 14.04, and HTTP traffic allowed as shown in Figure 2.



Region: asia-southeast1 (Singapore)
Zone: asia-southeast1-b

Machine type
Customize to select cores, memory and GPUs.
1 vCPU 3.75 GB memory Customize
Upgrade your account to create instances with up to 96 cores

Container
□ Deploy a container image to this VM instance. Learn more

Boot disk
New 10 GB standard persistent disk
Image
Ubuntu 14.04 LTS
Change

Identity and API access
Service account: Compute Engine default service account
Access scopes
● Allow default access
○ Allow full access to all Cloud APIs
○ Set access for each API

Firewall
Add tags and firewall rules to allow specific network traffic from the Internet
☑ Allow HTTP traffic
□ Allow HTTPS traffic

Figure 2 The detail configuration of the Google VM instance this project uses

The components needed to be installed or set are shown below.
- Apache2, access terminal using SSH and install apache in order to make its web server.
- Install PHP, Node.js, Python to run spiders and codes.

- Selenium, to get the code from Pantip search page after the program received the desired topic.
- Scrapy, to write spiders and extract needed data.
- Splash, to help Scrapy extracts JavaScript-based website data.
- Google service account, set service account on IAM in google console and select Role as Project Owner to access all APIs.
- Docker, to run Splash, the easiest way is to use Docker.

3.2.2 Program / Interface

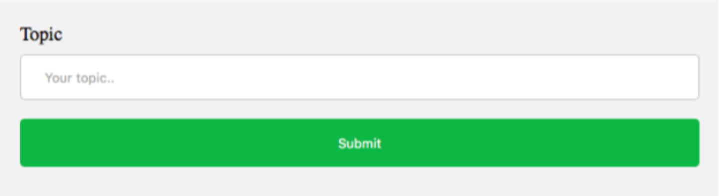


Figure 3 Text field on user interface used to receive a keyword or topic as an input

Users would enter the desired topic or keyword in the text field and press the submit button to begin an analysis as shown in Figure 3. By pressing submit, it will initiate the program to collect the input from the text field. Then, the program will transform the input into the target website search link. Following, the program will store the resulted link as a text file in the server.

3.2.3 Target Website

In this study, the target website is the Pantip website. The program will run the Selenium code which will follow the link in the text file, created from input UI. Next, the program will open the Pantip search page and later collect the web page source code in an HTML file. The program will run Scrapy code which will create a spider to craw the HTML file and collect data from every link within the page and save all those data into text files, sorted by month and year.

3.2.4 Google Translation API

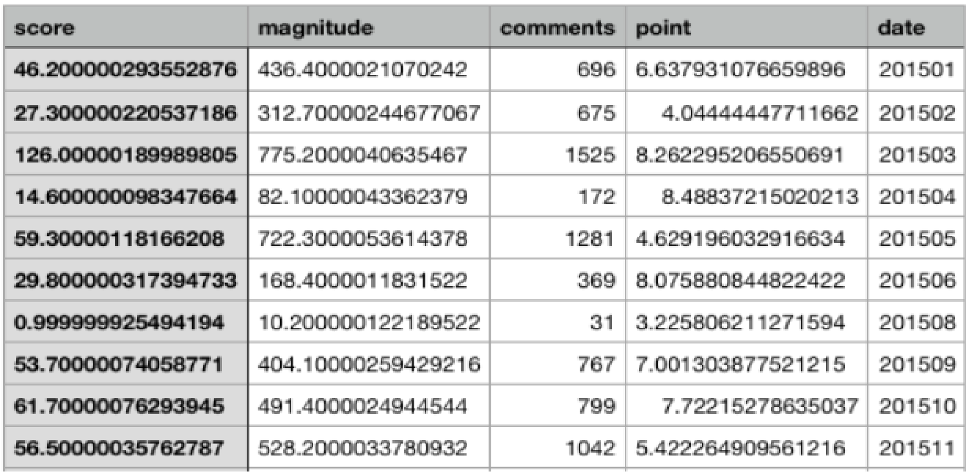
The program will read scraped text files. The timeframe can be configured according to the users' need. In this study, the timeframe was set between 2015 to 2017. The program then uses the Google Translation API to translate those files from Thai to English. The translated text is usually understandable, but it is not well organized which creates some minor errors. The program will save translated files into the server separately from original files.

3.2.5 Google Natural Language API

The program will analyze the translated text files using Google Natural Language API. The program uses sentimental analysis which inspects the given text and identifies the prevailing emotional opinion within the text, especially to determine a writer's attitude as positive, negative, or neutral. The program will collect point, magnitude, numbers of comments and save them as a CSV file as shown below in Figure 4.

3.2.6 Google Cloud Storage

- The program will upload the CSV file into Google Cloud Storage.
- The easiest method to provide inputs for Google Data Studio is to use Google Cloud Storage.



score	magnitude	comments	point	date
46.200000293552876	436.4000021070242	696	6.637931076659896	201501
27.300000220537186	312.70000244677067	675	4.04444447711662	201502
126.00000189989805	775.2000040635467	1525	8.262295206550691	201503
14.600000098347664	82.10000043362379	172	8.48837215020213	201504
59.30000118166208	722.3000053614378	1281	4.629196032916634	201505
29.800000317394733	168.4000011831522	369	8.075880844822422	201506
0.999999925494194	10.200000122189522	31	3.225806211271594	201508
53.70000074058771	404.10000259429216	767	7.001303877521215	201509
61.70000076293945	491.4000024944544	799	7.72215278635037	201510
56.50000035762787	528.2000033780932	1042	5.422264909561216	201511

Figure 4 Example of CSV file used to collect outputs from Google Natural Language and use them as inputs for Google Data Studio

3.2.7 Google Data Studio

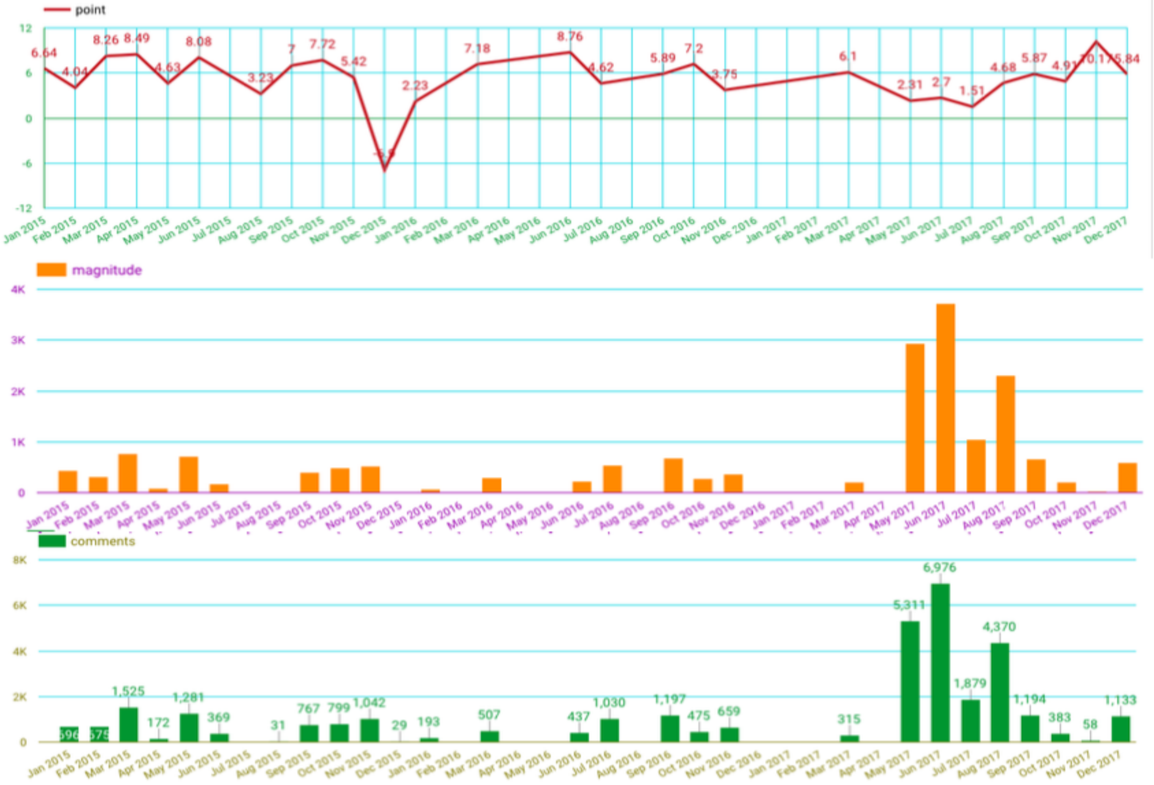


Figure 5 Example of output graphs ordered month by month from 2015 to 2017, point shows the overall emotional leaning of the text, magnitude indicates the overall strength of emotion

- The program will open the result graphs link from Google data studio.

- Google Data Studio uses CSV file in Google cloud storage to make graphs as shown in Figure 5.
- Point (first graph), the lower the point value, the more negative people opinions are in that month.
- Magnitude (second graph) is the aggressiveness of people towards the topic; it is related to a number of comments directly.
- Number of comments (last graph) shows how many comments that are related to the topic in that month, if the number of comments is too low, it can make the point swings significantly.

4. Results and Discussion

The framework performs Trend analysis from the Thai language decently by using keywords to determine the score instead of whole sentences or files. The main error is in grammar, so it does not affect much, but there are still some mistranslated words which cause the accuracy to drop.

The demonstrated used case “หุ้น EARTH” or “EARTH stock” as a topic and compared to the price of EARTH from 2015 to 2017 to figure (SiamChart, 2019) out the program efficiency. From Pantip website, the data was sorted by month and the program extracted 27 months of data from 36 (no post about EARTH in those 9 months) and the price of EARTH stopped in June 2017, so the data after June 2017 could not be used to test. The total extracted comments are 33005 comments. The output graphs are shown in Figure 6.

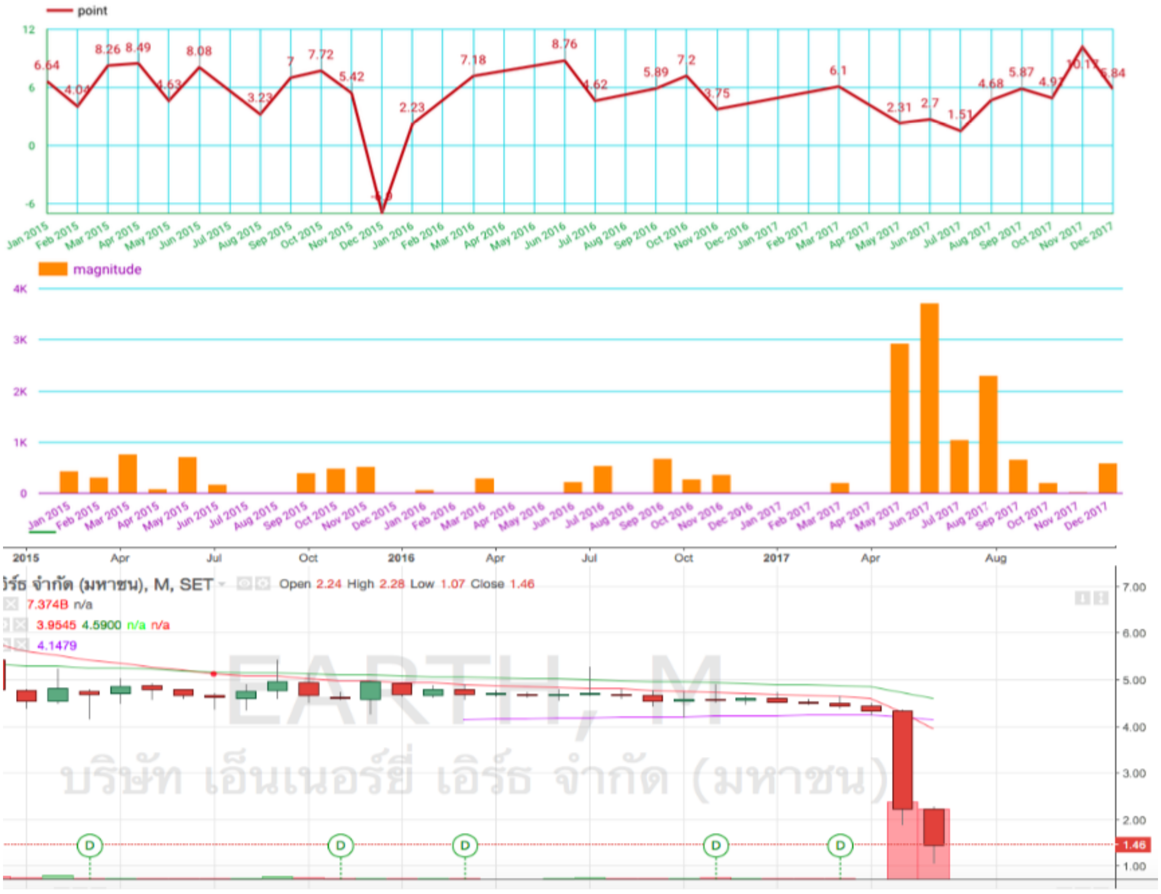


Figure 6 EARTH stock output and price graphs used to determine the accuracy of the program by comparing the point and magnitude to the drop and rise of EARTH price

There were two parts that the program was tested. The first part was in the apparent situation, or when the output should be exactly either positive or negative, the program was tested whether it can determine that. According to the graph, when the price dropped significantly at the end, it was when the EARTH price broke down. When the crisis happened, the period was compared to the magnitude graph, and it was found that the magnitude rose greatly as compared to all previous months because people talked about this topic frequently and intensively, and the point dropped from 6.1 in March 2017 to 2.31 in May 2017 meaning that most of the opinions were negative as shown in Figure 7.

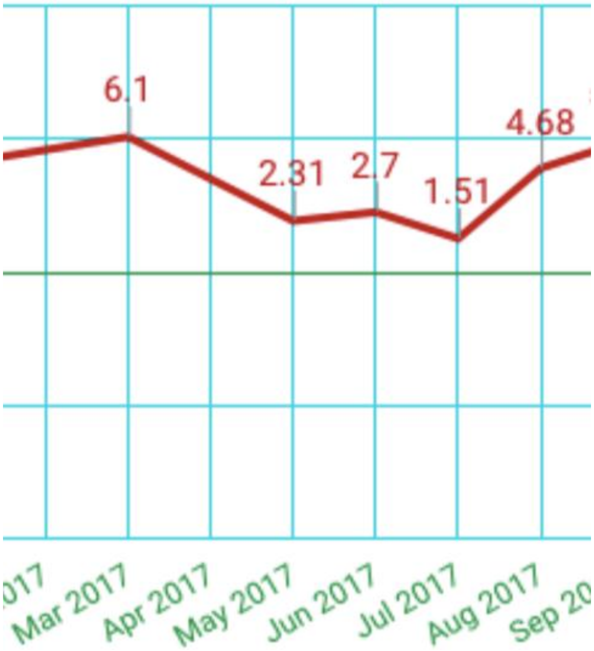


Figure 7 Point graph showing emotion towards EARTH before, during, and after massive price drop in May 2017

For the second part, the program was tested to see whether it can determine the price direction of the next month using the data from the previous month. Because the EARTH price was being manipulated, so the previous month data represented when news or manipulations were released, and the next month was when the price had been affected. In this part, the data used was only before the crisis happened.

The months that had numbers of comments lower than 385 comments were not taken into account because the average number of the comments in a normal case is 642 comments. If 642 comments are 100%, then 385 comments are only 60%. The lack of source data can cause errors and make the results biased, so this test only took into account the results from those that reached 385 comments or more. The average point was 5.38, any point above that would be considered positive and vice versa.

However, it should be reminded that the data in the Previous Month column were from the program and the data in the Next Month column were the actual price of the EARTH in that month in percentage counted from the starting price of the month to the closing price of the month.

As seen in Table 1, the program got 9 results correctly and 4 results incorrectly or in other words, 69.23 % accuracy.

In the apparent situation, the program can analyze easily and get an accurate result because in these cases there is a massive amount of information to work on and most data goes in the same direction. For example, when the EARTH crisis happened, many people complained and criticized heavily on Pantip, but for normal situations, when people think differently, the program needs to weight the opinion of people and determine whether it is positive or negative.

Table 1 Comparison table of the previous month point to the next month change of price in percentage

Previous Month	Point	Positive or Negative	Next Month	Change of price in percentage	Positive or Negative	Result after comparison
Jan 2015	6.64	Positive	Feb 2015	6.14%	Positive	Correct
Feb 2015	4.04	Negative	Mar 2015	-1.26%	Negative	Correct
Mar 2015	8.26	Positive	Apr 2015	3.39%	Positive	Correct
May 2015	4.63	Negative	Jun 2015	-2.5%	Negative	Correct
Sep 2015	7	Positive	Oct 2015	-5.65%	Negative	Incorrect
Oct 2015	7.72	Positive	Nov 2015	-0.43%	Negative	Incorrect
Nov 2015	5.42	Positive	Dec 2015	8.26%	Positive	Correct
Mar 2016	7.18	Positive	Apr 2016	0.43%	Positive	Correct
Jun 2016	8.76	Positive	Jul 2016	0.43%	Positive	Correct
Jul 2016	4.62	Negative	Aug 2016	-0.43%	Negative	Correct
Sep 2016	5.89	Positive	Oct 2016	0.88%	Positive	Correct
Oct 2016	7.2	Positive	Nov 2016	-0.43%	Negative	Incorrect
Nov 2016	3.75	Negative	Dec 2016	0.87%	Positive	Incorrect

The error comes to affect these cases due to the translation part because when API translates Thai to English, the data is understandable, but it hardly forms a complete sentence. Moreover, there are some words that API cannot understand its true meaning at all. In this project, the program uses keywords to determine if it is positive or negative; nonetheless, the error still remains.

Another factor is that to get a better result, the program should analyze weekly or daily instead of monthly, so the flow of data can be seen more precise and it will be easier to determine the result.

5. Conclusion

As shown above, the program used EARTH as a study case to determine the accuracy and reliability. It is safe to say that the program can analyze obvious cases easily and have 69.23% accuracy in normal cases.

In conclusion, the project translated and analyzed a total of collected 33005 comments of “EARTH stock” topic from the Pantip website to create the graphs showing people’s emotions, represented by point and magnitude, towards EARTH month by month from 2015 to 2017 with the efficiency of 69%. The accuracy in this study is quite limited since the dataset from Pantip was also limited and biased by the group of the people. Extending the platform into multiple data source system could also improve the accuracy.

This project can be further developed to create Real-time trend analysis to monitor the change of trend or interest of people or use to create quantity research in a short time. For example, the company can use this program to obtain the people opinions towards its campaign launched last week in only 1-2 hours.

Suggestion for further study, Google Cloud Natural Language cannot be used with the Thai language directly so the way to improve the accuracy is to either use Natural Language service which can be used with Thai language or change Translation platform as using the Google Cloud Translation to translate the Thai language is only decent but not efficient. Automation process for analyzing daily or weekly should also help as well. Also, the extra component to combine multiple sources process can greatly improve accuracy.

6. Acknowledgments

We are thankful to our colleagues Chonlatee Yuttasoontorn and Dollachat Kitsawat who helped and provided expertise that greatly assisted us.

7. References

Google Cloud Natural Language (2019). *Cloud Natural Language.* Retrieved February 15, 2019, from https://cloud.google.com/natural-language/

Google Cloud Translation API (2019). *Cloud Translation API documentation* Retrieved February 15, 2019, from https://cloud.google.com/translate/docs/

Google Cloud Virtual Machine Instances (2019). *Virtual Machine Instances.* Retrieved February 15, 2019, from https://cloud.google.com/compute/docs/instances/

Hashimoto, T., Aramvith, S., Chauksuvanit, T., & Shirota, Y. (2014, October). Framework for language independent social media analysis platform to detect reactions on global topics. In *TENCON 2014-2014 IEEE Region 10 Conference* (pp. 1-6). IEEE.

Sneha, C. V., Sarangh, E. R., Vindhya, V., Hegde, R. R., & Lakshmi, V. A. (2017, August). Comprehensive analysis of CSR data using interactive reports. In *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)* (pp. 1460-1463). IEEE.

Shahare, F. F. (2017, June). Sentiment analysis for the news data based on the social media. In *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 1365-1370). IEEE.

Sanguansat, P. (2016, February). Paragraph2Vec-based sentiment analysis on social media for business in Thailand. In *2016 8th International Conference on Knowledge and Smart Technology (KST)* (pp. 175-178). IEEE.

Scrapy (2019). *Scrapy a fast and powerful scraping and web crawling framework*. Retrieved February 15, 2019, from https://scrapy.org

Scrapy+Splash Plugins (2019). *Scrapy+Splash for JavaScript integration*. Retrieved February 15, 2019, from https://github.com/scrapy-plugins/scrapy-splash

SiamChart (2019). *Earth stock price graph.* Retrieved February 15, 2019, from http://siamchart.com/stock-chart/EARTH/

Wang, Z., Chong, C. S., Lan, L., Yang, Y., Ho, S. B., & Tong, J. C. (2016, December). Fine-grained sentiment analysis of social media with emotion sensing. In *2016 Future Technologies Conference (FTC)* (pp. 1361-1364). IEEE.