Identification of sentiment polarity in Thai texts using SO-PMI-IR and TF-IDF based algorithms: A case study of crypto valuation

Sasiphan Nitayaprapha

Faculty of Liberal Arts, Thammasat University

DOI: 10.29322/IJSRP.13.04.2023.p13609
http://dx.doi.org/10.29322/IJSRP.13.04.2023.p13609

Paper Received Date: 15th February 2023
Paper Acceptance Date: 27th March 2023
Paper Publication Date: 6th April 2023

Abstract- Cryptos have emerged and played an important role in financial market. The researcher draws from the theory of the relationship between social media and crypto equity value. This research hypothesizes that social media affect crypto valuation. The research attempts to construct an appropriate Thai sentiment dictionary for cryptos to improve a textual sentiment analysis, and to understand the relationship between social media sentiment variables and crypto valuation. The designed system architecture composes of six modules: data acquisition, data preprocessing, sentiment dictionary construction, sentiment estimation, sentiment categorization, and time series analysis. The outcomes of the research would be an appropriate Thai crypto sentiment dictionary, a method useful to estimate sentiment of posts on social media, and the understanding of the predictive relationship between crypto valuation and social media.

Index Terms- sentiment analysis, text mining, data mining, machine learning, crypto valuation.

I. Introduction

This paper outlines the ongoing research which proposes a method for sentiment analysis in Thai textual context. The research aims to construct an appropriate opinion dictionary for crypto valuation, and to improve a textual sentiment analysis.

Cryptos have emerged and played an important role in financial market. The researcher draws from the theory of the relationship between social media and crypto equity value [1]. This research hypothesizes that social media affect crypto valuation. The examination of relationship between social media and crypto value an opportunity for better understanding of the economic value of social media.

It is evidenced that social media usage has become one of the most popular online activities. As of 2021, the world population size is 7.9 billion [2], of which 4.26 billion are active social media users. The number of social media users is projected to reach six billion in 2027. The commonly used social media are Facebook, Instagram, Line, and Twitter. These media have become the important sources of news and information. Facebook has 2.895 billion monthly active users. In fact, Facebook is the largest social media platform [3].

Facebook is the place where people update by reading news, posts, and watching streaming videos. Additionally, many of Facebook users provide feedbacks by providing comments, sharing, and clicking 'LIKE' button. The important question regarding posts and comments is how do people think about the issue? Sentiment analysis have become a critical method to evaluate user comments and posts. The ability to analyze the underlying sentiment enables an in-depth understanding of the issues. This makes sentiment analysis become a critical task in data analysis.

In general, analysis of textual posts is carried out by using text mining techniques which is based on NLP (Natural Language Processing techniques), and machine learning methods. NLP and machine learning techniques are deployed to estimate textual post sentiment, whether it is positive, neutral, or negative. or others. The estimation is complicate, because firstly, an appropriate opinion corpus is not available, and secondly a precise analysis requires a large size of textual data.

The research context of Thai textual posts is targeted at the crypto Facebook sites. The data are collected from posts on three main Thai crypto Facebook pages. The collected posts are the domain used to construct an opinion corpus. The expected research outcomes would be 1) An appropriate Thai crypto opinion dictionary, 2) A method to evaluate the sentiment of posts on social network sites, and 3) The relationship between crypto valuation and social media.

1.1 Research Problems

1) There is no Thai crypto opinion corpus available.
2) There is no software tool to perform crypto sentiment analysis on Thai social network sites.
3) There is no research on relationship between crypto value and social media in the context of Thai textual data.

1.2 Research Objectives

1) To propose a method to construct a Thai crypto opinion corpus.

2) To propose a method to estimate the sentiment of posts on social network sites.
3) To study a relationship between crypto valuation and social media.

1.3 Research Outcomes

1) A Thai crypto sentiment dictionary.
2) A method to evaluate sentiment of posts on social network sites.
3) The relationship between crypto valuation and social media.

II. Literature review

Sentiment analysis involves Natural Language Processing (NLP). NLP involves with text classification and text segmentation, thus is an important technique to interpret the sentiment of content.

This research deploys TLex+; a Thai word NLP tool, to do word segmentations and reduce noises in words. TLex+ uses machine learning method reduce noises in texts. TLex+ is constructed based on Conditions random fields. Hirankan, et al. (2013) analyzes Thai words by using TLex, the previous version of Tlex+[4]. The authors aim to detect repetitions consonantal character and vowel characters. Some research use conditions random fields algorithm to detect spelling errors in posts extracted from social network [5].

The research methods are as follows: 1) perform Thai word segmentation using TLex+, 2) construct sentiment dictionary based on PMI-IR and TF-IDF algorithms, 3) carry out sentiment analysis using decision tree, Naïve-Bay, and Support vector machine, and 4) study predictive relationship between sentiment variables and crypto valuation using VECM. The following section details some theories used in this research.

2.1 Thai Language

Thai language has a unique writing system. The language composes of three parts, consonant, vowel and tone [6]. Thai language has 87 characters, 44 consonant characters, 18 vowel symbols, 5 diacritics, 4 tone marks, 10 numerals, and 6 other symbols. The language is written from left to right. The vowel symbols can be placed in front, back, above, or below a consonant character. Tone marks are placed above a consonant. Diacritics are written either above or below a consonant. There is no word separation between words, and no rule of the use of spaces. There might or might not be a space between phrases and sentences. There are no capital letters. Additionally, verbs do not change forms according to concord or tense. Tense is stated with auxiliary verbs or time adverbials [7].

2.2 Social Media and Natural Language

Social media enables users to make textual comments and posts, share contents, and express feeling through LIKE buttons. There have been many research conducted to interpret textual comments and posts to understand users' feelings or sentiments. This research aims to analyze sentiment on Thai textual posts collected from Facebook sites.

Social network provides virtual communities which allow users to communicate, collaborate, and share opinions. The example of widely used social network providers are Facebook, Google+, Instagram, Line, Twitter, and YouTube. The social network sites become one of the main information sources for data mining and warehousing. There have been many studies on text mining [8]. These studies involve corpus construction, information extraction, information evaluation, information perception, text classification, text clustering, and sentiment analysis. Carrying out text mining needs preprocessing tasks related to natural language processing. The NLP tasks include identification of letters, words, sentences, as well as their structures and meanings.

Because, the texts retrieved from social networks are not well-structured as those maintained in database system. The posts are usually tying in informal way without careful proofreading. They often contain non-standard spellings, slangs, and emoticons. Moreover, words often be written incorrectly with special characters, for example "ดี ดีย์ ดี๊ดีย์" (good). These makes word segmentation more difficult. [9]

2.3 Noise Reduction

Noise reduction is necessary. The task removes unnecessary elements in the text before using it for sentiment analysis. This task is critical since social media text is usually written informally with "noises" as special characters, or extra repetitions of characters. Two main approaches used to reduce noises are a heuristic and machine learning approach.

2.3.1 Heuristic Approach

A heuristic approach uses a set of rules constructed by analyzer's experience. In general, different rules are designed for different types of noises; such as advertisement, emoticons, extra repetitions of characters, spam, special characters, and URLs. [10]

2.3.2 Machine Learning Approach

A machine learning approach deploys a machine learning method to reduce noises. The examples of research which employ machine learning approach to analyzed Thai words in social media are:
Analyzing text in social network by using TLex [4]. The authors attempt to detect repetitions consonantal and vowel characters.
Detecting spelling errors [5].

2.3.2.1 Conditional Random Fields

Conditional random fields (CRFs) deploys a supervise learning algorithm. CRFs calculates conditional probability based on discriminative undirected probabilistic graphical model. CRFs is developed from Maximal Entropy Markov Models (MEMMs). It is claimed that CRFs solves the label bias problem of MEMMs [11]. CRFs is often deployed for labeling, segmenting, and tagging part of speech. . The examples of research in Thai texts using CRFs are "A Supervised Learning based Chunking in Thai using Categorical Grammar" [12], and "LexToPlus: A Thai Lexeme Tokenization and Normalization Tool" [13].CRFs equation is shown by Equation 1.

$$P(y|x) = \frac{1}{Z(x)} exp \quad (\sum_{t=1}^{T} \sum_{k=1}^{K} \lambda kfk(yt-1, yt, x, t) \quad (1)$$

http://dx.doi.org/10.29322/IJSRP.13.04.2023.p13609

Where $P(y|x)$ is a conditional probability.
x, y is a sequence of result.
K is a number of function
T is a position. λk is a weight of feature function. fk is a feature function.
$Z(x)$ is a normalization function can be written as shown in Equation 2.

$$Z(x) = \sum_y exp\ (\sum_{t=1}^{T} \sum_{k=1}^{K} \lambda kfk(yt-1, yt, x, t) \qquad (2)$$

2.4 Sentiment Analysis

Sentiment analysis, opinion analysis, or opinion mining, is the analysis of text to understand emotions and tones. "Sentiment analysis" is a method that "uses the natural language processing (NLP), text analysis and computational techniques to automate the extraction or classification of sentiment from sentiment reviews" [14]. The result of the analysis is the classification of emotions and opinions into polarity categories as positive, neutral, negative, or more types [15].

The examples of sentiment analysis research in the domain of social media are those on Facebook [16] [17], Twitter [18] [19] [20], News [21] [22], and Products and services [23] [24]. All these researches aim to develop a sentiment dictionary for the targeted domain. Every single word in the dictionary has polarity values, positive, neutral, negative, or others. There are two approaches to score sentiment, coarse(or binary) sentiment and fine-grained sentiment analysis [25]. The coarse sentiment distinguishes sentiment into positive, neutral, or negative polar. The fine-grained approach distinguishes sentiment as many subgroups. For example, Socher et al. (2013) categorize sentiment into five groups: very negative, negative, neutral, positive, and very positive [26].

The examples of sentiment analysis research conducted on Thai social media textual domain are:
Categorizing posts in twitter: Vateekul and Koomsubha (2016) use two deep learning techniques, Long short Term Memory (LSTM) and Dynamic Convolutional Neural Network (DCNN) to analyze sentiment in posts [27]. Meeprasert and Rattagan (2021) analyze customers on Twitter by using Random Forest for topic classification, and WangchanBERTa (pre-trained Thai Language model) for sentiment classification [28].

Categorizing posts in twitter and web board: Haruechaiyasak et al. (2018) propose S-Sense as an intention and sentiment analysis framework for Thai social media. The researchers employ a social media corpus in mobile service domain which obtained from Twitter and Pantip web board, and deploy Naive Bayes as the classification algorithm [29].

The sentiment value of a text resulted from a summation of the frequency of the words in the sentiment dictionary. Because sentiment analysis is based on a sentiment dictionary, the available and reliable sentiment dictionary for a targeted textual domain is essential.

2.5 Sentiment Dictionary

Sentiment dictionary, or sentiment lexicon, is critical for sentiment analysis. The dictionary consists of words having polarity value; positive, neutral, negative, or others. Many researchers have devised construction techniques for sentiment dictionary. The examples of the researches include:
"Connotation Lexicon: A Dash of Sentiment Beneath the Surface Meaning" builts sentiment dictionary based on the purposed induction algorithms. The algorithms are based on three types of widely adopted algorithmic frameworks for sentiment lexicon induction; Constraint Optimization via Integer Linear Programming , HITS & PageRank, and Label/Graph Propagation. [30]
"Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora" proposes SentProp algorithm which uses PMI and learns orthogonal transformations of word vectors to create sentiment lexicons. [31]
"Ntusd-fin: a market sentiment dictionary for financial social media data applications." constructs a financial sentiment analysis lexicon based on PMI, frequency-inverse document frequency, and the cosine similarity as features [32]

2.5.1 PMI-IR Algorithm

PMI-IR is a statistical technique used to estimate semantic orientation of words. PMI-IR calculates associated target word to a reference word or "seed word". Seed word has polarity value varied from strong positive to strong negative [33] [34]. PMI defines sentiment value of word as Equation 3, semantic orientation (SO) is defined in Equation 4.

$$PMI(x, y) = \log_2 \left(\frac{P(x,y)}{P(x)P(y)}\right) \qquad (3)$$

PMI is a measure of association between target word (x) and seed words (y)
$P(x, y)$ is the probability that both x and y appear in data
$P(x)$ is the probability that x appears in data.
$P(y)$ is the probability that y appears in data
$\log_2 \left(\frac{P(x,y)}{P(x)P(y)}\right)$ is statistical dependence between (x) and (y) per amount of data acquired

$$SO(x) = \Sigma PMI(x, positive\ seed\ word) - \Sigma PMI(x,\ negative\ seed\ word) \qquad (4)$$

SO is a calculation of sentiment value of target word (x) derived by sum PMI positive minus sum PMI negative.

In addition, SO-PMI-IR has various of "AND" operation and "NEAR" operation. "AND" operation uses total number of comments containing both target word and seed word (Equation 5). "NEAR" operation counts number of comments containing both target word and seed word within the pre-defined distance in a comment (Equation 6).

$$\mathrm{SO}(x) = \log_2 \left(\frac{hit(x\ AND\ positive\_list)hit(negative\_list)}{hit(x\ AND\ negative\_list)hit(positive\_list)}\right) \qquad (5)$$

$$\mathrm{SO}(x) = \log_2 \left(\frac{hit(x\ NEAR\ positive\_list)hit(negative\_list)}{hit(x\ NEAR\ negative\_list)hit(positive\_list)}\right) \qquad (6)$$

2.6 TF-IDF

TF-IDF (Term Frequency Inverse Document Frequency) is the most popular technique of feature calculation of text classifier. TF-IDF computes values for each word in a document by using an inverse proportion of the frequency of the word in the document to the percentage of documents in which the word appears. This implies words with high TF-IDF numbers have a strong relationship with the document they appear in. TF-IDF could be used as a measure of the important of word and document regarded to the corpus. TF-IDF presented words in matrix format which help reducing length of document. TF-IDF consists two parts, Term Frequency (TF) and Inverse Document Frequency (IDF) [35] [36] [37].

TF is a frequency of term, denoting a weight of each word (or term) that occurred in a document. TF is calculated as Equation 7.

$$Wi,j = TFi,j \qquad (7)$$

Where i is term. j is document. Wi,j denotes weight of term i in document j

IDF is an inverse document frequency. IDF is a measure of the important of word (or term) related to a whole document. IDF is calculated as Equation 8.

$$IDFi = \log_2 \left(\frac{N}{ni}\right) + 1$$

(8)

Where N is the total number of documents. ni is the number of documents that term i is appeared.

$$TFi,j * IDFi = TFi,j * \log_2 \left(\frac{N}{ni}\right) + 1$$

(9)

TF-IDF is obtained by multiplying components as shown by Equation 9. TF-IDF shows how often a term occurred in all documents. TF-IDF weighting schemes is a popular method for text classification [38].

2.7 Classification Algorithm

Classification Algorithm is used for text classification. There are three algorithms which are commonly used in sentiment analysis: Decision Tree, Naïve-Bay, and Support Vector Machine [39] [40] [41] [42]

2.7.1 Decision Tree

Decision Tree model consists of nodes and branches. Node represents attributes. Root node is the topmost node which has no branch comes into, but has branches go out to other nodes. Internal node is the node that branch comes in and goes out. Terminal node (leaf node) is the last node which has no branch going out. Branch is the connection between nodes. Each branch denotes the result of an attribute's value.

To construct a decision tree, all the attributes are estimated. The attribute influencing most the highest classification will be the root node, then the attribute influencing the second most the highest classification will be the next node, and so on. It is noted that, feature selection is critical for decision tree. The algorithms devised to create decision tree are ID3, C4.5, C5.0 and CART. ID3, C4.5, and C5.0 are devised by Quinlan (Quinlan 2014) who firstly develops ID3 algorithm. ID3 computes the Information Gain of each of attributes, then selects and adds attribute nodes with the highest value. This process will repeat until the data is completely categorized. C4.5 algorithm is invented to solve the bias problem existing in ID3. C5.0 is a commercial software claimed to outperform C4.5 in the sense that it better utilizes memory. Thus C5.0 is more suitable for big data analysis.

CART (Classification and Regression Trees algorithm) deploys a binary tree of which each node has mostly two branches. CART employs Gini Index to select an attribute to be the root node, internal node, or leaf node. [43]

In the context of Thai textual domain, many NLP classification researches using decision tree algorithm.

2.7.2 Naïve-Bayes

Naïve-Bayes (simple Bayesian) is based on Bayes' theorem which uses probability calculation called conditional probability. The algorithm analyzes the relationship between independence variables to evaluate probability condition for each relationship. Theoretically, the result will be accurate if all features are independence. The conditional probability of independent feature A and feature B is calculated by

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

(10)

B is a feature (attribute) used to calculate the posterior probability of feature A

P(A|B) is the posterior probability of A conditioned on B

P(B|A) is the posterior probability of B conditioned on A

P(A) is the prior probability of A

P(B) is the prior probability of B

Equation 11 presents Naïve- Bayes' calculation of conditional probability. X denotes a a class onsisting of attributes N number X1, X2, X3, …, Xn. X can have M subclasses C1, C2, C3, …, Cm.

$$P(Ci \;|X) = \frac{P(X|Ci)P(Ci)}{P(x)}$$

(11)

The examples of Thai textual research using Naïve-Bayes include "Multi-stage annotation using pattern-based and statistical-based techniques for automatic Thai annotated corpus construction" [44], and "S-Sense: A sentiment analysis framework for social media sensing" [9].

2.7.3 Support Vector Machine (SVM)

SVM is a supervised learning algorithm invented by Cortes and Vapnik (1995) [45]. Given a set of training examples, each example will be defined as a member of one of two categories. A training algorithm constructs a model which assign examples to the points in a space of one category to maximize the margin between the two categories. Taking into consideration the width of the gaps between two categories, the new examples are mapped into the same space and predicted to belong to the category. The classification could be linear or non-linear. For non-linear data, SVM deploys Kernel function to transform low dimensional input space to a higher dimensional space. The commonly used kernel functions of SVM are Linear kernel, Radial basis kernel (RBF), Polynomial kernel, and Sigmoid Kernel.

The examples of sentiment analysis of textual data researches are "An Empirical Study on Multi-Dimensional Sentiment Analysis from User Service Reviews" [46] and "Sentiment Analysis of Thai

Online Product Reviews using Genetic Algorithms with Support Vector Machine” [47].

2.8 VECM

This research employs VECM to examine whether there is a predictive relationship between crypto valuations and social media variables. VECM is used to evaluate the interdependencies across time-series. [48]

VECM is used rather than multiple regression [49] because VECM enables modeling of recursive relationship between interdependent variables. Additionally, VECM does not required a knowledge of the mechanisms influencing variables needed by structural models. Furthermore, VECM allows cross-correlation and autocorrelation. This could lead to better understanding of the dynamic relationships among variables. In addition, causality between VECM variables can be assessed by using Granger causality. In this research, VECM is deployed to test whether the sentiment values of social media variables are helpful for predicting crypto values. In this research, crypto variables include daily price, number of transactions, and trading volume. Additionally, social media variables used are number of posts each of which are classified as positive (POS) and negative (NEG). VECM with p variables, k lags, and co-integration order r is presented in Equation 12:

$$\Delta \boldsymbol{Y}_t = \sum_{j=1}^{k-1} \Gamma_j \Delta \boldsymbol{Y}_{tt-k} + \alpha\beta' \boldsymbol{Y}_{tt-1} + \boldsymbol{\mu} + \epsilon_t \qquad (12)$$

Where Δ is the first difference operator.
Yt is a p x 1 vector with order of integration 1.
μ is a p x 1 constant vector denoting the linear trend.
Γj is a p x p matrix representing short-term relationships among variables.
€ is the residual vector. k is the lag length.
α is a p x r matrix presenting the speed with which the variables adjust to the long-term equilibria.
β is a p x r matrix representing the long-term relationships between the co-integrating vectors.
β′Yt-1 is the error correction term, this term does not exist in VAR model.

III. RESEARCH METHODOLOGY

The research objectives are 1) To propose a method useful to construct a Thai crypto sentiment dictionary, 2) To propose a method to evaluate sentiment of the posts on social network sites, and 3) To find out whether there is a predictive relationship between crypto valuation and social media.

The designed system architecture has six modules as shown in figure 1. The data will be collected from three main Thai crypto Facebook pages; Bitcoin Crypto Thailand – members 6.1 hundred thousand, Bitcoin Thai Community– members 4.9 hundred thousand, and Bitcoin Thailand Community– members 1.8 hundred thousand.

3.1 *System Architecture*

User interface will be designed properly by using bootstrap. The backend part is developed by using Facebook-scrapers, MySQL, python, RapidMiner, and TLex+. The system modules are as follows:

A. Data Acquisition: this module extracts daily textual posts about crypto, along with other relevant information, source and date/ time of the posts. Also, the module performs data cleansing, and exports posts to the database in the appropriate text format.

B. Data Preprocessing: data preprocessing employs TLex+ to remove noises and do word segmentation.

C. Sentiment Dictionary Construction: this module deploys SO-PMI-IR based algorithms and the constructed set of seed words to calculate association value between target words and seed words.

D. Sentiment Estimation: sentiment estimation module is part of the sentiment dictionary construction module. This module computes polarity value of target words, and updates the polarity value of the word maintained in the dictionary.

D. Sentiment Categorization: sentiment categorization module constructs feature vectors by using TF-IDF, and polarity value of words in the sentiment dictionary. In addition, this module classifies posts by using three different machine learning methods; Decision Tree, Naïve Bayes, and SVM.

E. Time series Analysis: this module deploys VECM to estimate the relationship between crypto valuation, and the post sentiment. This approach works the best in guidance of fellow researchers. In this the authors continuously receives or asks inputs from their fellows. It enriches the information pool of your paper with expert comments or up gradations. And the researcher feels confident about their work and takes a jump to start the paper writing.

IV. CONCLUSION

Crypto currencies have emerged and played an important role in financial ecosystems. However, there have not been many studies conducted on relationship between social media and crypto value. The research is carried out to construct an appropriate Thai sentiment dictionary for crypto valuation, to improve a textual sentiment analysis, and to examine the relationship between social media and crypto valuation. The expected research outcomes are an appropriate Thai crypto sentiment dictionary, a method to evaluate sentiment of posts on social network sites, and the understanding of the relationship between crypto valuation and social media.

REFERENCES

[1] X. Luo and J. Zhang, “How do consumer buzz and traffic in social media marketing predict the value of the firm?,” Journal of Management Information Systems, 2013, 30, 2, pp.213–238. doi:10.2753/MIS0742-1222300208.

[2] United Nations, “New UN report examines links between population growth, socioeconomic development and environmental change,” 2022, Available from: https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/undesa_pd_2022_media_advisory.pdf.

[3] S. Dixon, “Number of global social network users 2018-2027,” 2022, Available from: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/.

[4] P. Hirankan, A. Suchato, and P. Punyabukkana, “Detection of Wordplay Generated by Reproduction of Letters in Social Media Texts,” Proceedings of

10th International Joint Conference on Computer Science and Software Engineering (JCSSE), 2013.

[5] N. Kunpattanasopon, N. Tongtep, and K. Hashimoto, "Noise Reduction Effect on Thai Social Texts Sentiment Analysis," Proceedings of Artificial Intelligence and Natural Language Processing (iSAI-NLP), 2017.

[6] P.T. Daniels, Writing Systems, in The handbook of Linguistics, 2003. Edited by Mark Aronoff and Janie Rees-Miller, Blackwell Publishers.

[7] S. Iwasaki and P. Ingkaphirom, A Reference Grammar of Thai, 2009, Cambridge University Press.

[8] R. Irfan, C. King, D. Grages, S. Ewen, S. Khan, S. Madani, H. Li, "A survey on text mining in social networks," The Knowledge Engineering Review, 30, 2, 157-170. doi:10.1017/S0269888914000277

[9] C. Haruechaiyasak, A. Kongthon, P. Palingoon, and K. Trakultaweekoon, "S-Sense : A Sentiment Analysis Framework for Social Media Sensing," in Workshop on Natural Language Processing for Social Media, Nagoya, Japan. 2013, pp.6–13.

[10] A. K. Nassirtoussi, et al., "Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment," Expert Systems with Applications, 2015, 42, 1, pp.306-324.

[11] C. Sutton and A. McCallum, "An Introduction to Conditional Random Fields," Foundations and Trends in Machine Learning, 2012, 4, 4, pp.267-373.

[12] T. Supnithi, P. Porkaew, T. Ruangrajitpakorn, et al., "A Supervised Learning based Chunking in Thai using Categorial Grammar," Proceedings of the 8thWorkshop AsianLanguage Resources, 2010, pp.129-136.

[13] C. Haruechaiyasak and A. Kongthon, "LexToPlus: A Thai Lexeme Tokenization and Normalization Tool," Proceedings of the Fourth Workshop on South and Southeast Asian Natural Language Processing, 2013.

[14] A. Basant, M. Namita, B. Pooja, and G. Sonal, Sentiment Analysis Using Common-Sense and Context Information, Hindawi Publishing Corporation Computational Intelligence and Neuroscience, 2015.

[15]. Hussein Doaa Mohey El-Din Mohamed, "A survey on sentiment analysis challenges," Journal of King Saud University - Engineering Sciences, 2018, 30, 4, pp.330-338,

[16] A. Ortigosa, J. M. Martín, R. M. Carro, "Sentiment Analysis in Facebook and its Application to E-learning," Computers in Human Behavior, 2014, 31, pp.527-541.

[17] F. Lucie, R. Eugen, and P. Daniel, "Analysing domain suitability of a sentiment lexicon by identifying distributionally bipolar words," Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2015.

[18] X. Bing and Z. Liang, "Improving twitter sentiment analysis with topic-based mixture modeling and semi-supervised training," Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), Association for Computational Linguistics, 2014.

[19] S. Rosenthal, N. Farra, and P. Nakov, "SemEval-2017 Task 4: Sentiment Analysis in Twitter," Proceedings of the 11th International Workshop on Semantic Evaluation, 2017.

[20] H. K. Sul, A. R. Dennis, L. Yuan, "Trading on Twitter: Using Social Media Sentiment to Predict Stock Returns," Decision Sciences, 2017, 48, 3, pp.454-488.

[21] C. Ouyang, W. Zhou, Y. Yu, Z. Liu, and X. Yang, "Topic sentiment analysis in Chinese news," International Journal of Multimedia and Ubiquitous Engineering, 2014, 9, 11, pp.385-396.

[22] M. Day, and C. Lee, "Deep learning for financial sentiment analysis on finance news providers," IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2016, pp.1127-1134.

[23] M. Subhabrata, and B. Pushpak, Feature Specific Sentiment Analysis for Product Reviews, in CICLing 2012, part I, LNCS 78181, Springer-Verlag, Berlin Heidelberg, 2012.

[24] X. Fang and J. Zhan, "Sentiment analysis using product review data," Journal of Big Data, 2015, 2, 5, https://doi.org/10.1186/s40537-015-0015-2

[25] R. Kohtes, From valence to emotions: How coarse versus fine-grained online sentiment can predict real-world outcomes, Anchor Academic Publishing, 2015.

[26] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, et al., "Recursive deep models for semantic compositionality over a sentiment treebank," Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2013, pp.1631–1642.

[27] P. Vateekul and T. Koomsubha, "A Study of Sentiment Analysis using Deep Learning Techniques on Thai Twitter Data," Proceedings of the 13th International Joint Conference on Computer Science and Software Engineering, 2016.

[28] W. Meeprasert and E. Rattagan, "Voice of Customer Analysis on Twitter for Shopee Thailand," Journal of information systems in Business JISB, 2564, 7, 3, pp.7-18.

[29] C. Haruechaiyasak, A. Kongthon, P. Palingoon, and K. Trakultaweekoon, "S-Sense: A Sentiment Analysis Framework for Social Media Monitoring Applications," Information Technology Journal, 2018, 14, 1.

[30] S. Feng, J. Kang, P. Kuznetsova and C. Yejin, "Connotation Lexicon: A Dash of Sentiment Beneath the Surface Meaning," ACL 2013 51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 2013.

[31] W. Hamilton, K. Clark, J. Leskovec and D. Jurafsky, "Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora," Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp.595–605.

[32] C. C. Chen, H. H. Huang, and H. H. Chen, "Ntusd-fin: a market sentiment dictionary for financial social media data applications," Proceedings of the eleventh international conference on language resources and evaluation (LREC), 2018.

[33] P. D. Turney, "Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL," Proceedings of European Conference on Machine Learning, 2001.

[34] S. Kalamdhad, S. Dubey, and M. Dixit, "Feature Based Sentiment Analysis of Product Reviews using Modified PMI-IR method," International Journal of Computer Trends and Technology (IJCTT), 2016, 34, 2, pp.115-121.

[35] A. Aizawa, "An information-theoretic perspective of tf–idf measures," Information Processing and Management, 2003, 39, pp.45–65.

[36] J. Ramos, Using Tf-idf to Determine Word Relevance in Document Queries, 2003, https://pdfs.semanticscholar.org/b3bf/6373ff41a115197cb5b30e57830c16130c2c.pdf?_ga=1.183821606.1606390940.1482600858.

[37] H. K. Yadla and P. P. Rao, "Machine Learning Based Text Classifier Centered On TF-IDF Vectoriser," International Journal of Scientific & Technology Research, 2020, 9, 3, pp.583-586.

[38] B. Das and S. Chakraborty, "An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation," ArXiv, 2018, abs/1806.06407.

[39] A. Ortigosa, J. M. Martín, R. M. Carro, "Sentiment Analysis in Facebook and its Application to E-learning," Computers in Human Behavior, 2014, 31, pp.527-541.

[40] G. Wang, J. Sun, J. Ma, K. Xu, and J. Gu, "Sentiment Classification: The Contribution of Ensemble Learning," Decision Support Systems, 2014, 57, pp.77-93.

[41] Y. Al-Amrani, M. Lazaar, K. E. Elkadiri, "Sentiment Analysis using supervised classification algorithms," Proceedings of the 2nd international Conference on Big Data, Cloud and Applications (BDCA'17), Association for Computing Machinery, 2017, Article 61, pp.1–8.

[42] M. Guia, R. Silva, and J. Bernardino, "Comparison of Naïve Bayes, Support Vector Machine, Decision Trees and Random Forest on Sentiment Analysis," Proceedings of the 11th International Conference on Knowledge Discovery and Information Retrieval, 2019, pp.525-531.

[43] L. Breiman, Classification and Regression Trees, Routledge, New York, 2017.

[44] N. Tongtep and T. Theeramunkong, "Multi-stage Annotation using Pattern-based and
Statistical-based Techniques for Automatic Thai Annotated Corpus Construction," Proceedings of the 9th Workshop on Asian Language Resources, 2011.

[45] C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, 1995, 20, 3, pp.273-297.

[46] S. Thanangthanakij, E. Pacharawongsakda, N. Tongtep, P. Aimmanee, and T. Theeramunkong, "An Empirical Study on Multi-Dimensional Sentiment Analysis from User Service Reviews," Proceedings of the Seventh International Conference on Knowledge, Information and Creativity Support Systems, 2012.

[47] R. Tesmuang and N. Chirawichitchai, "Sentiment Analysis of Thai Online Product Reviews using Genetic Algorithms with Support Vector Machine," Progress in Applied Science and Technology, 2020, 10, 2, pp.7-13.

[48] Y. Lee and J. H. Rhee, "A VECM analysis of Bitcoin price using time-varying cointegration approach," Journal of Derivatives and Quantitative Studies, 2022, 30, 3, pp.197-218. https://doi.org/10.1108/JDQS-01-2022-0001

[49] W. Antweiler and M. Z. Frank, "Is all that talk just noise? The information content of internet stock message boards," Journal of Finance, 2004, 59, 3, pp.1259–1294. doi:10.1111/ j.1540-6261.2004.00662.x.

AUTHOR

Asst.Prof.Dr.Sasiphan Nitayaprapha

Ph.D. in Informatics (MBS,The University of Manchester)
Faculty of Liberal Arts, Thammasat University,
contact: sasiphan.nit@gmail.com

http://dx.doi.org/10.29322/IJSRP.13.04.2023.p13609