International Journal of Technology (2018) 8: 1609-1617
ISSN 2086-9614

A SENTIMENT ANALYSIS VISUALIZATION SYSTEM FOR THE PROPERTY INDUSTRY

Nurul Husna Mahadzir[1*], Mohd Faizal Omar[1], Mohd Nasrun Mohd Nawi[2]

[1]*School of Quantitative Sciences, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia*
[2]*School of Technology Management and Logistics, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia*

(Received: July 2018 / Revised: September 2018 / Accepted: December 2018)

ABSTRACT

The usage of social media platforms such as Facebook and Twitter, either by the public or by organizations, has been rapidly increasing. The decision-makers in the organizations use social media to engage with their customers since public users tend to express their opinions about certain products and services through this popular mechanism. Hence, this valuable data can be useful for marketing and business decisions. However, the main obstacle is obtaining meaningful information from these platforms due to the unstructured data they present. Sentiment analysis is seen as the best tool to analyze insights or opinions in this huge amount of data. In this article, we extract data on public opinion about property in order to understand the reason behind the imbalances of supply and demand currently faced by the property industry in Malaysia. In addition, we visualized the sentiment results in the form of a dashboard so that it may help property players to understand the public sentiments toward their housing or construction projects.

Keywords: Housing; Property; Sentiment analysis; Social media; Visualization

1. INTRODUCTION

The property scenario in Malaysia is currently facing a crucial supply and demand imbalance. As reported by the Central Bank of Malaysia (BNM), the supply and demand imbalances in the property market have increased since 2015, with unsold residential properties already at their highest in 10 years. According to the leading property consultant, Knight Frank Malaysia, Malaysia's property market is expected to be sluggish this year (NST Business, 2018). This problem is due to the oversupply situation and the imbalances in supply and demand for the properties. The causes of the supply and demand mismatch crucially need to be addressed to ensure that the property market can completely reach its target.

Therefore, the BNM proposed that the government should gather information about characteristics and preferences of the public in order to effectively meet the demand of households. The lack of that information has contributed to a large number of overhang properties, including the affordable housing projects in Johor, Selangor, and Kedah (Kay, 2018). In addition, Datuk Seri FD Iskandar Mohamed Mansor, the president of the Real Estate and Housing Developers' Association (REHDA), also suggested that the property and real estate players should have a one-stop data center to assist decision-makers in the property

*Corresponding author's email:nurul.husna.mahadzir@ahsgs.uum.edu.my,Tel. +60-13-3882842,Fax.+6049284756
Permalink/DOI: https://doi.org/10.14716/ijtech.v9i8.2753

industry to better bridge the gap between supply and demand and to predict future housing trends (Ling et al., 2017). It is crucial to have an integrated database on property supply and demand that can provide insights on the needs and preferences of households, their links to demographic shifts, and property gaps across Malaysia (Mustafa et al., 2017).

Currently, several studies involving traditional methods of research, such as surveys and questionnaires that targeted only certain groups of people have been conducted to gather public preferences on the property sector, including affordable housing projects (Jamaluddin et al., 2016). However, the traditional research methods are limited to a particular set of questions that are sometimes forced onto people who might not give candid answers. To address this gap, we proposed to gather information from social media platforms since the amount of data that can be captured through Twitter and Facebook is massive. Moreover, the information shared in social media is considered to be honest feedback from the public since they posted their opinions without being asked for it.

In order to analyze such a huge amount of unstructured data in social media, we used Sentiment Analysis (SA) as a tool to compute and analyze public opinions or sentiments written in the form of text. SA is a field of research in Natural Language Processing (NLP) that aims to automatically determine the attitude of a speaker or writer based on the subjective information shared on the Web (Pang & Lee, 2008; Liu, 2012). The importance of this field has been proven by the extensive number of methods and approaches that have been proposed in research as well as by the interest of organizations and companies that it has raised in recent years. Previous studies have reported that SA has been applied on wide variety of topics and issues such as online products reviews (e.g., mobile phones; Di Fabbrizio et al., 2013), hotel reviews (Kasper & Vela, 2011), political and financial analysis (Soelistio & Surendra, 2015; Chiong et al., 2018), housing (Mahadzir et al., 2016) and the prediction of real-word events (Rifai et al., 2015). Past research on the data visualization of SA in various domains, including political sentiments during 2012 United States presidential election (Wang et al., 2012), real-time monitoring and analysis of football (Saavedra, 2016), and customer reviews on products and services (Al Kubaizi et al., 2018; Chen & Zheng, 2018), have been presented.

SA is commonly divided into two main techniques, which are machine learning and lexicon based. The machine learning technique attempts to train a sentiment classifier based on the occurrence frequencies of the various words in the datasets (Feldman, 2013; Santosh & Vardhan, 2015). There are several well-known machine learning methods that have been applied such as Maximum Entropy, Support Vector Machine (SVM), and Naïve Bayes. Based on previous study, it has been demonstrated that the Naïve Bayes method leads to better performance and accurate classification (Kunal et al., 2018). Meanwhile, the sentiment lexicon requires sentiment dictionaries consisting of sentiment words and their polarity to classify words. For example, the polarity for “best” and “worst” are positive and negative respectively. Various sentiment lexicons have been constructed either manually or semi-automatically such as SentiWordNet (Medagoda et al., 2015) and SenticNet (Cambria et al., 2014).

A huge amount of previous research has been done in mining the sentiments written in the English language. Despite the fact that SA research in English is rather mature, SA studies in other languages such as Malay have just set sail (Al-Moslmi et al., 2017).

In this paper, a case study method is presented that uses Twitter data to analyze public sentiments toward the Perumahan Rakyat 1Malaysia (PR1MA) project. We proposed the implementation of SA toward the property industry using the machine learning method. We used Naïve Bayes as a classifier due to its accurateness and effectiveness as demonstrated in previous Malay SA research (Alshalabi et al., 2013). In order to make the analysis results readable and understandable by the property players, we visualized the results in the form of a

dashboard. The dashboard is able to help the property players in understanding public preferences and refining their marketing strategies to enable the industry to better bridge the gap between supply and demand in this sector.

2. METHODS

Our objective was to analyze public opinion toward property and to visualize the results in the form of a dashboard. For this study, we applied a case study method that studied the government affordable housing project known as PR1MA and we used Twitter as the data source. We collected the data using the Twitter API that mentioned two keywords: PR1MA or #PR1MA. Only tweets written in Malay and English were extracted for this research. Figure 1 show several raw tweets posted regarding PR1MA.

Figure 1 Sample tweets mentioning PR1MA

We divided the research activities into two phases: (1) the SA, which involves data pre-processing and sentiment classification; and (2) the data visualization, which involves the design and development of the dashboard. Figure 2 illustrates the overview of our system architecture. Next, we elaborate on each phase and the activities involved in detail.

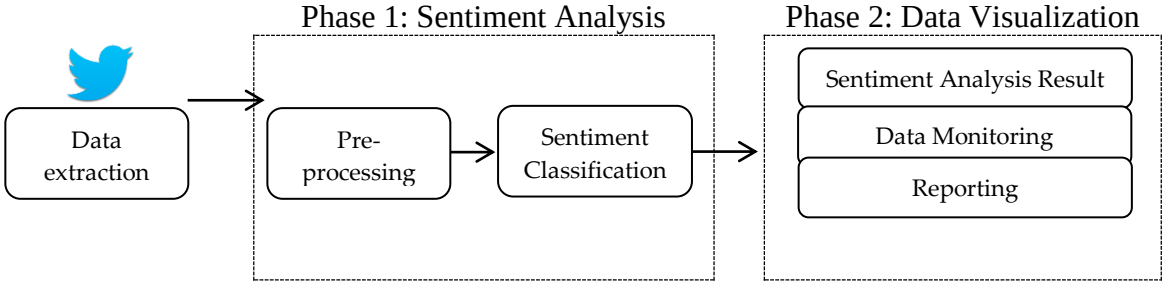

Figure 2 System architecture

2.1. Data Extraction

The data used for this study were extracted from the Twitter platform. The tweets were limited to the "PR1MA" keyword between January 2018 and April 2018. Only 2018's tweets were

collected in order to analyze the most current tweets about the PR1MA project. We have collected 745 tweets consist of the reviews written in Malay and English languages.

2.2. Phase 1: Sentiment Analysis

During this phase, we dealt with two main processes, data pre-processing and sentiment classification. The data extracted from Twitter was analyzed and classified into three categories which were positive, neutral, and negative.

2.2.1. Pre-processing

As Twitter data contains a lot of idiosyncratic idioms, such as emoticons, URLs, RT for re-tweet, @ for user mentions, # for hashtags, and repetitions, it was necessary to pre-process and normalize the data. Among the activities involved during pre-processing was the removal of the repetition tweets (re-tweets) and unwanted tags, tokenization, corrections of spelling, and stop words for both the Malay and English language. Figure 3 shows the output of the pre-processing activity.

Sample tweet	@winstonetco RT myedgeprop: Target of 1mil #PR1MA home off by 983,318 https://buff.ly/2PdEmXO #myedgeprop
Post-processing	target 1mil pr1ma home off by 983318

Figure 3 The output after the pre-processing activity

2.2.2. Sentiment classification

We performed two levels of classifications, which were the overall sentiments and the aspect-based SA. Due to the lack of publicly available Malay sentiment resources and the limited amount of Malay text classification research, all Malay tweets were first translated into English prior to the classification activity.

The experiment was conducted with the commonly used sentiment classifier, Naïve Bayes, and the R application as the analytic tool. A total of 600 tweets were used for the sentiment classification purpose. We split the dataset using 70% for training and 30% for testing as summarized in Table 1.

Table 1 Training and testing data

Training data	420
Testing data	180
Total dataset	600

Since the classifier needs to be trained, we requested two annotators to perform a sentence level annotation for each post. Each tweet was given a score of either -1 (negative), 0 (neutral), or 1 (positive). The mean Cohen's kappa coefficient of the inter-annotator agreement between the sets of annotations was around 0.87. The kappa coefficient is a reliable and robust measure of the agreement between two users (Yu et al., 2018). We have used this manual classification as a baseline to measure the accuracy of our machine learning classifier.

2.3. Data Visualization

In order to visualize the results of the analysis performed in the first phase, we designed a dashboard using the Tableau application to display the SA's results, data monitoring, and reporting. It consists of three components:

1) The main dashboard gives the overall sentiments of the analysis and the aspect-based analysis. It contains the amount of data being analyzed, the overall sentiments of positive,

neutral, and negative classes, the most trending words mentioned in Twitter, and the sentiments for each aspect/features in the property domain.

2) In the second component, we display the real-time Twitter monitoring consisting of the daily statistic, the number of re-tweets and likes, and the most frequent tweets on a particular subject.
3) Finally, a third component shows the details of tweets based on their feature/aspect and polarity, which may help the property players to view the details of what people talked about.

3. RESULTS AND DISCUSSION

We have collected all the relevant tweets in real-time that mention PR1MA. Below, we presented the results for sentiment classification and visualized the results in the form of a dashboard.

3.1. Sentiment Analysis

Table 2 shows the results for the overall sentiments obtained from this experiment using the three classification categories of positive, neutral, and negative.

Table 2 Sentiment classification's result

Classifier	Positive	Neutral	Negative
Manual annotation	42.27%	10.40%	47.33%
Naïve Bayes	39.17%	8.33%	52.50%

Table 3 Precision, recall, and accuracy

Evaluation Measures	
Precision	0.93
Recall	0.88
Accuracy	0.90

Based on the data presented in Table 3, our classification shows an accuracy of 90%. The accuracy measure is shown as Equation 1 where True Positive (TP) and True Negative (TN) denote correct classifications and False Positive (FP) and False Negative (FN) denote incorrect classifications.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

However, the learning algorithm was slightly biased toward a positive classification, which is evident from the confusion matrix. Most of the errors are due to negative posts being identified as positive. In future works, we will use a larger dataset to train the system in order to eliminate the bias toward positive and neutral sentiments.

3.2. Data Visualization

Figures 4–6 present the visualization system for the results we obtained from the previous phase. Figure 4 shows the main dashboard. The total number of data being analyzed is displayed in the upper left corner. It is followed by the overall sentiments for the project and the most mentioned keywords in the right corner. In the middle of the page, we display a word

cloud that represents the frequency of words that appear in the datasets. At the bottom of the page, we show the SA results for each feature, such as the price, design, and location.

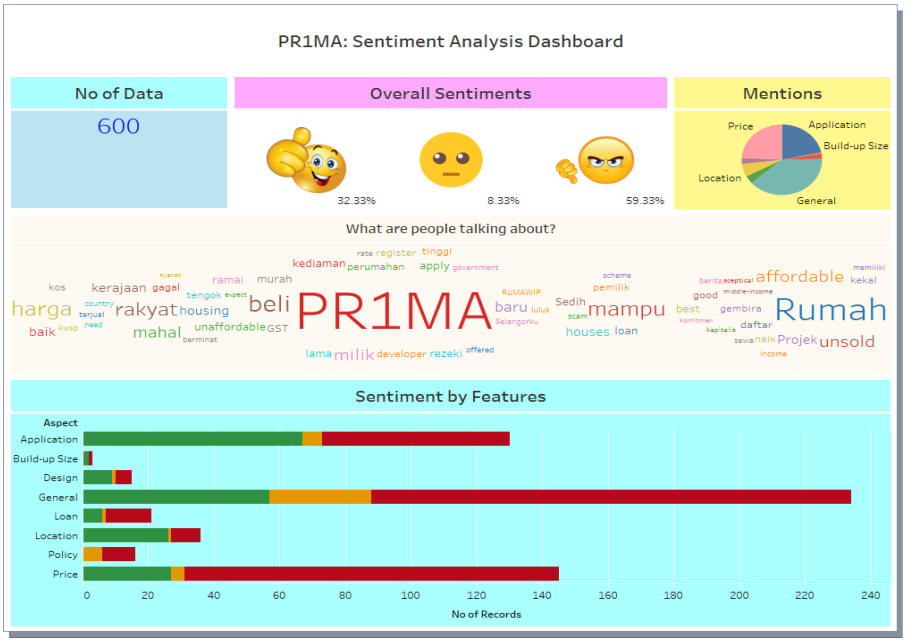

Figure 4 First page: Sentiment analysis results

Figure 5 demonstrates a real-time data monitoring system from the Twitter platform. We provide this monitoring system to assist the decision-makers in keeping up with what's happening online and to see what people are talking about regarding their projects or services. We presented a detail report to display all the posts with their aspect/feature and polarity categories in Figure 6.

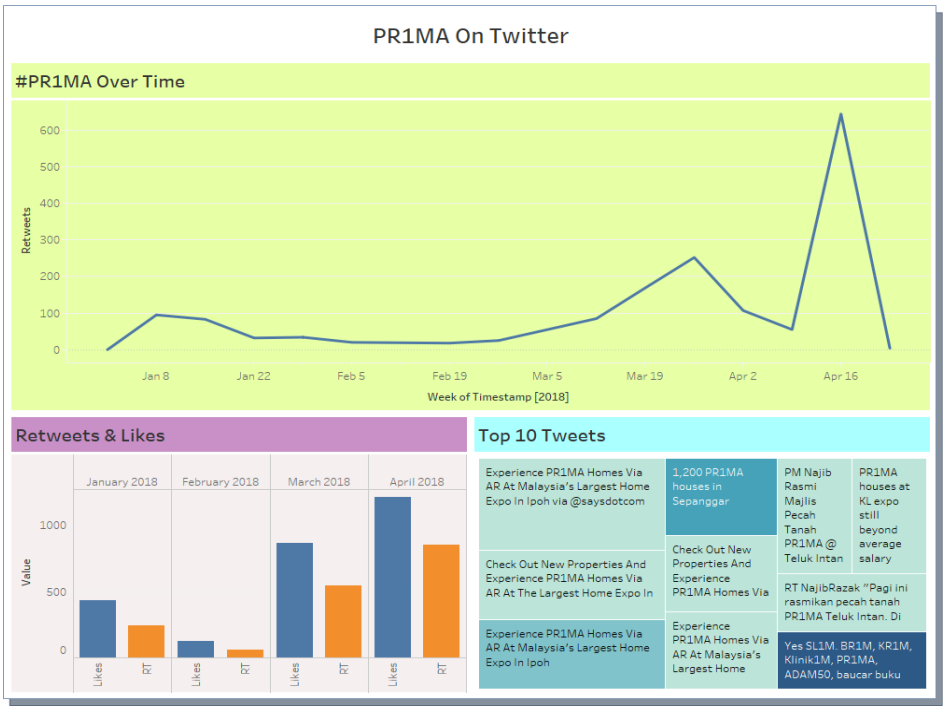

Figure 5 Second page: Data monitoring

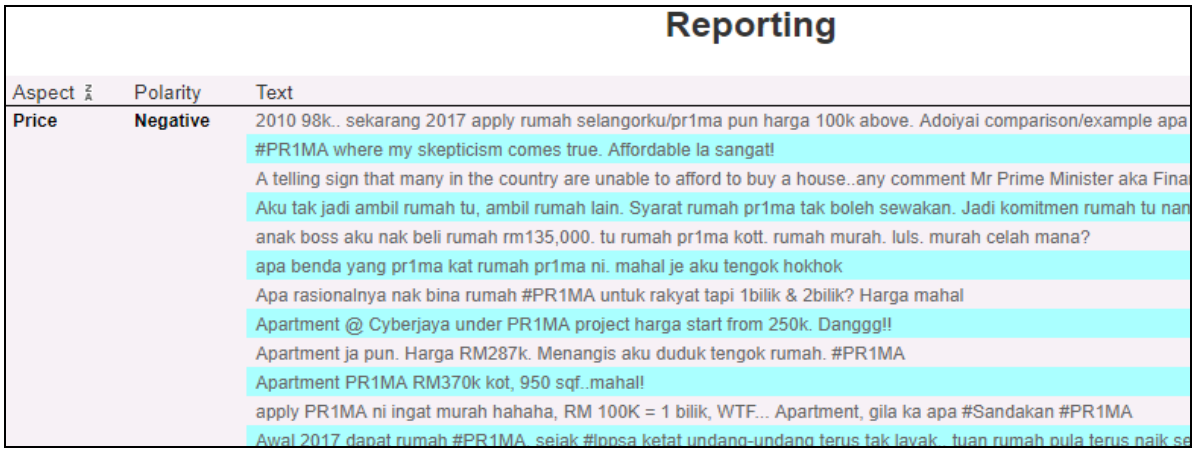

Figure 6 Third page: Detail reporting

4. CONCLUSION

This SA visualization system is meant to assist the property players in understanding the public views about their housing or construction projects. Hence, decision-makers could make subtle and informed decisions with a better understanding of outside information.

In this paper, we proposed the implementation of SA toward the property industry. A case study involving Twitter data in analyzing public sentiments toward the PR1MA project has been presented. We implemented a machine learning algorithm and used Naïve Bayes to carry out the sentiment classification process. To make the analysis results readable and understandable by the property players, we visualized the results in the form of a dashboard. Our dashboard system consists of three main components which are the overall and feature-based SA, Twitter data monitoring, and reporting.

However, there is a room for improvement. In the future, we are planning to apply various algorithms or to use the sentiment lexicon to perform the classification instead of machine learning alone. Our goal is to find the best classification technique for Malay and English text. We also will extend our work by extracting data from other social media platforms such as Facebook and online forums.

5. ACKNOWLEDGEMENT

This research is supported by the University Grant (S/O Code: 13879), University Utara Malaysia, 2018.

6. REFERENCES

Al Kubaizi, R., Al-Otaibi, S., Al Washigry, B., Al Suhaim, E., Al Sughayer, J., Al Jumaiah, R., 2018. Mining Expertise using Social Media Analytics. *In:* 1st International Conference on Computer Applications & Information Security (ICCAIS), IEEE, pp. 1–5

Al-Moslmi, T., Omar, N., Albared, M., Alshabi, A., 2017. Enhanced Malay Sentiment Analysis with an Ensemble Classification Machine Learning Approach. *Journal of Engineering and Applied Sciences*, Volume 12(20), pp. 5226–5232

Alshalabi, H., Tiun, S., Omar, N., Albared, M., 2013. Experiments on the Use of Feature Selection and Machine Learning Methods in Automatic Malay Text Categorization. *Procedia Technology*, Volume 11, pp. 748–754

Cambria, E., Olsher, D., Rajagopal, D., 2014. SenticNet 3: A Common and Common-sense Knowledge Base for Cognition-driven Sentiment Analysis. *In:* AAAI'14 Proceedings of

the Twenty-eighth AAAI Conference on Artificial Intelligence, Québec City, Québec, Canada, pp. 1515–1521

Chen, C.I.P., Zheng, J., 2018. Improved Big Data Analytics Solution using Deep Learning Model and Real-time Sentiment Data Analysis Approach. *In:* International Conference on Brain Inspired Cognitive Systems, Springer, Cham, pp. 579–588

Chiong, R., Fan, Z., Hu, Z., Adam, M.T.P., Lutz, B., Neumann, D., 2018. A Sentiment Analysis-based Machine Learning Approach for Financial Market Prediction Via News Disclosures. *In:* Proceedings of the Genetic and Evolutionary Computation Conference Companion, Kyoto, Japan, pp. 278–279

Di Fabbrizio, G., Stent, A.J., Gaizauskas, R., 2013. Summarizing Opinion-related Information for Mobile Devices. *In:* Mobile Speech and Advanced Natural Language Solutions, Springer, New York, pp. 289–317

Feldman, R., 2013. Techniques and Applications for Sentiment Analysis. *Communications of the ACM*, Volume 56(4), pp. 82–89

Jamaluddin, N.B., Abdullah, Y.A., Hamdan, H., 2016. Encapsulating the Delivery of Affordable Housing: An Overview of Malaysian Practice. *In:* MATEC Web of Conferences, Volume 66

Kasper, W., Vela, M., 2011. Sentiment Analysis for Hotel Reviews. *In:* Computational Linguistics-Applications Conference, Jachranka, Poland, pp. 45–52

Kay, L.K., 2018. *Rehda: Malaysia's Property Industry Must Tap into Big Data*. Available Online at https://www.edgeprop.my/content/1289419/rehda-malaysia's-property-industry-must-tap-big-data, Accessed on 10 September 2018

Kunal, S., Saha, A., Varma, A., Tiwari, V., 2018. Textual Dissection of Live Twitter Reviews using Naive Bayes. *Procedia Computer Science*, Volume 132, pp. 307–313

Ling, C.S., Almeida, S.J., Wei, H.S., 2017. *Affordable Housing: Challenges and the Way Forward*. Bulletin. Bank Negara Malaysia. Available Online at http://www.bnm.gov.my/files/publication/qb/2017/Q4/p3ba1.pdf, Accessed on 15 September 2018

Liu, B., 2012. Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, Volume 5(1), pp. 1–167

Mahadzir, N. H., Omar, M. F., Nawi, M. N. M. 2016. Towards Sentiment Analysis Application in Housing Projects. *Journal of Telecommunication, Electronic and Computer Engineering*, Volume 8(8), pp. 145-148

Medagoda, N., Shanmuganathan, S., Whalley, J., 2015. Sentiment Lexicon Construction using SentiWordNet 3.0. *In:* 11th International Conference on Natural Computation (ICNC), Zhangjiajie, China, pp. 802–807

Mustafa, A., Adnan, N., Nawayai, S.S.M., 2017. The Influence of Product Quality and Service Quality on House Buyer's Satisfaction in Prima Home. *Pertanika Journal of Social Science and Humanities*, Volume 25(4), pp. 1841–1851

NST Business, 2018. Local Property Market Continues to Self-correct: Knight Frank Malaysia. Available Online at https://www.nst.com.my/business/2018/01/329939/local-property-market-continues-self-correct-knight-frank-malaysia, Accessed on July 20, 2018

Pang, B., Lee, L., 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, Volume 2(1-2), pp. 1–135

Rifai, A.I., Hadiwardoyo, S.P., Correia, A.G., Pereira, P., Cortez, P., 2015. The Data Mining Applied for the Prediction of Highway Roughness due to Overloaded Trucks. *International Journal of Technology*, Volume 6(5), pp. 751–761

Saavedra, A.P., 2016. *Development of a Dashboard for Sentiment Analysis of Football in Twitter based on Web Components and D3. js*. Master's Thesis, ETSI Telecommunication, Universidad Politécnica de Madrid, Spain

Santosh, D.T., Vardhan, B.V., 2015. Obtaining Feature- and Sentiment-based Linked Instance RDF Data from Unstructured Reviews using Ontology-based Machine Learning. *International Journal of Technology*, Volume 6(2), pp. 198–206

Soelistio, Y.E., Surendra, M.R.S., 2015. Simple Text Mining for Sentiment Analysis of Political Figure using Naive Bayes Classifier Method. *In:* The Proceedings of The 7th ICTS, Bali, Indonesia

Wang, H., Can, D., Kazemzadeh, A., Bar, F., Narayanan, S., 2012. A System for Real-time Twitter Sentiment Analysis of 2012 US Presidential Election Cycle. *In:* Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju, Republic of Korea, pp. 115–120

Yu, L.C., Lee, C.W., Pan, H.I., Chou, C.Y., Chao, P.Y., Chen, Z.H., Tseng, S.F., Chan, C.L., Lai, K.R., 2018. Improving Early Prediction of Academic Failure using Sentiment Analysis on Self-evaluated Comments. *Journal of Computer Assisted Learning*, Volume 34(4), pp. 358–365