ISSN 2347 - 3983
Volume 8. No. 6, June 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter83862020.pdf
https://doi.org/10.30534/ijeter/2020/83862020

Classification of Twitter Data by Sentiment Analysis in the Malay Language

Abdul Karim Mohamad[1], Mailasan Jayakrishnan[1], Nurnajwa Hazwani Nawi[2]
[1]Centre for Advanced Computing Technology, Faculty of Information & Communication Technology, Universiti Teknikal Malaysia Melaka, karim@utem.edu.my
[2]Faculty of Information & Communication Technology, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100, Durian Tunggal, Melaka, Malaysia

ABSTRACT

A set of tweets are characterized manually using human clarification with their sentiments and contemplate as training data. Then an additional set of tweets that is live streaming, are composed hinge on the text mining on Twitter Streaming Application Programming Interface (API). The tweets are fetched and retain as a text data and later will be utilized as a testing set. Testing data will grasp from training data calculation to forecast sentiment significance. The scenario of this research is no Twitter dataset collection available in the Malay language with characterizing sentiment significance. Then, findings are filtered to search tweets in the Malay language that is from Malaysia. The crucial dispute about this research is to cumulate a characterize entity as a training set. Considering there is no characterize Twitter entity available in the Malay language, a database of sentences is manually characterized with sentiments utilizing human explanation and utilize tweet's geo-position to search for tweets display within Malaysia. The outcome of this research, Twitter entity utilizing Twitter Streaming API capable to be collected and tweets from Malaysia collected by utilizing tweet's geo-position capable to be acquired.

Key words: Artificial Neural Network, Big Data, Dataset, Machine Learning, Sentiment Analysis, Twitter.

1. INTRODUCTION

The Internet has revolutionized communications and computer nature like nothing before [1]. The internet has proven to be useful and has come with a lot of advantages and a lot of disadvantages too [2]. The Internet is at once a world-wide transmission efficiency, a structure for knowledge circulation, and a channel for interaction and collaboration between the entity and their analogs without mark for the geographic position [3]. To determine the emotion of a person by their writings is a challenging task in developing sentiment analysis [4]. Twitter is an online broadcast and community mingle section where the community interacts within precise notice termed tweets. Tweeting is displayed the precise message for everyone who pursues you on Twitter, with the desire that your word is interesting and useful to someone within your congregation. A lot of drivel is on Twitter but concurrent, there is a foundation about the suitable message and knowledgeable content. Twitter is a way to determine the world through another person's sentences.

We perform the sentiment analysis where the tweets undergo preprocessing phase, specifically the tweets are filtered to remove unnecessary strings such as emoticons and Http links. We aim to investigate how to collect tweets on Twitter using Twitter Streaming API, how to use tweets geolocation to search for tweets in the Malay language originated from Malaysia, and how to categorize Twitter datasets into their sentiment values using a decision tree.

2. LITERATURE REVIEW

Twitter is a social media application that has been launched since 2006 [5]. With social networking use surpassing web-based email utilizes in February 2009, a few about the connection are just not being created by humans. Business is taking advantage of the social networking medium to search for chances, promotions, workers, and details on how customers make use of their products and services. Supervised learning used existing machine learning methods to carry out sentiment analysis [6]. It includes constructing classifiers from responses such as movie reviews, for instance, are used as a training and testing set [7].

Machine learning methods that usually used are Support Vector Machine (SVM), Naïve Bayes (NB), Maximum Entropy (ME), and k-Nearest Neighbour (kNN). This method is initialized with data collection. Feature selection is the procedure to select a set of attributes or features that suits the process mining. According to [8], they applied Support Vector Machine, Naïve Bayes, and Maximum Entropy for

movie reviews. Naïve Bayes works best for smaller training data, however, for larger training data, Support Vector Machine (SVM) has the optimum execution compared to Naïve Bayes and Maximum Entropy.

Furthermore, knowledge representation and machine learning are the major contributions of tweet sentiment analysis [9]. The machine learning method uses a training set to classify the sentiment value of each word accurately [10]. This method does not need a database the same as knowledge representation, which is a good thing. Therefore, supervised learning is one of the machine learning tasks about building activity from designate training data. The training data includes a set containing training samples. According to [11], in supervised learning, each object is a combination subsist of an input item and a suitable output advantage. A supervised learning method performs analyzation of the training data and the outcome of implied action, which can be utilized for aligning advanced objects [12].

The main disadvantage of using a supervised method is that the classifiers' performance depends on the training data. Larger and higher quality training data produce a better classification [13]. Information lacking and minimal information about training data can result in misclassification [14]. This research study on the method to calculate the sentiment value of Twitter data in the Malay language. The significance of sentiment analysis is that it can be a good source of information and can supply a model that is beneficial to companies such as improving the quality of a product or service. Besides, it can prove whether a campaign is a success or not plus improve strategy making.

3. METHODOLOGY

The methodology should be associated with the literature toward illustrating why this research is utilizing conclusive approaches and the scholarly ground about the research preferred. Furthermore, it is a structure about deep rules or principles from which precise approaches or actions can be obtained via illustrating or clarify divergent dilemmas within the breadth of a precise practice. Unlike an algorithm, a methodology is not a procedure but a set of the process. It is the mechanism utilized to collect data and information for the determination of composing research outcomes. The methodology is the systematic steps and theoretically used in developing this research. Therefore, we have adopted the waterfall model as the methodology of this research. The waterfall model is a software development process. We have designed and developed a specific waterfall model for this research purpose as shown in Figure 1.

Figure 1: The Research Waterfall Model.

Figure 1 shows the research waterfall model that consists of five (5) phases; (1) Data Collection, (2) Data Analysis, (3) Experimental Design, (4) Evaluation and Testing, and (5) Conclusion. Every phase is vital and must be done from phase to phase to produce a high-quality result. By having this phase, it will keep the research on track to achieve high accuracy and quality. We have tabulated the activities involved in the phases as shown in Table 1.

Table 1: The Activities Involved in the Phases

No	Phase	Activities Performed	Deliverables	References
1	Data Collection	Focused on collecting data that is needed via performing the sentiment analysis about Twitter data in the Malay language.	Methods need to be executed are as follows: • Define the goal of collecting data, researchers must always focus on their problem statements and objectives so that the research can be successful. • Declare principle before collecting data so that unnecessary procedures would not be taken and affect the quality and the outcome of the data. • Initialize collection data by starting the process and follow the principle that has been started and every step is vital thus needs to be recorded in case there is a need for reference and documentation purpose. • Observe the quality pattern of the data by referring to the problem	[15], [16]

			statements and objectives. If the quality is not fulfilling the objectives of the research, the data collection phase must be repeated.	
2	Data Analysis	Tweets undergo the preprocessing phase.	• Filtering where the text attribute is extracted from the tweet. Other attributes such as user attributes, created at attribute, and language attribute are not needed in the process to determine the sentiment of the tweets. Then tweets are filtered to remove unnecessary strings such as emoticons and Http links. • Tokenization, where Tweets are tokenized, broken down into words to divide the text by spaces and punctuation marks and each word will be labeled as the neutral, positive, or negative hinge on the data dictionary which contains words with their respective sentiment analysis values. If there is no pairing match, the word will be considered as positive. • Stemming is removing the suffix and prefix such as "-lah", "-nya", "-kan", from a word, to gain only the root word. • Removing stop words such as "saya", "dia", "yang" is eliminated from the tweets and the space left is considered as a positive word.	[17], [18]
3	Experimental Design	The decision tree is chosen as the classifier to predict the sentiment values of testing data.	• To create a decision tree, information content, or can be called as the entropy of the sentiment column must be calculated first. By having the entropy, information gain can be searched for all the tokens. A decision tree can be formed after all information is calculated. By producing the decision tree, we can predict the sentiment values of tweets in the testing set.	[19], [20]
4	Evaluation and Testing	The sentiment values of the same testing data will be calculated by using the JCreator LE tool.	• To compare whether using decision tree classifier from human interpretation is as accurate as using a machine tool.	[21], [22]
5	Conclusion	The conclusion will be made after the result is obtained.	• The best method for this research besides revealing the outcome of the research.	[23], [24]

Table 1 summarized the activities involved in the phases of the research waterfall model that we have designed and developed based on five (5) phases; (1) Data Collection, (2) Data Analysis, (3) Experimental Design, (4) Evaluation and Testing and (5) Conclusion.

4. ANALYSIS

The analysis is done to make sure the information is meaningful and can be used as a reference in developing strategies in assuring the success or failure of an experiment. We will discuss assessing the accuracy of the result of this research. This is an important process to prove that this research is functional and practical. Data need to be analyzed so that to ensure whether the outcomes are high in quality and consistent or not. We will insert the data into a tool which is JCreator LE, to calculate the precision of sentiment analysis of the data by using the Multilayer Perceptron (MLP) method. JCreator LE uses Java programming language. MLP source code is obtained from the internet which is freely used and is modified to enable the coding to calculate sentiment analysis using Artificial Neural Network as shown in Figure 2.

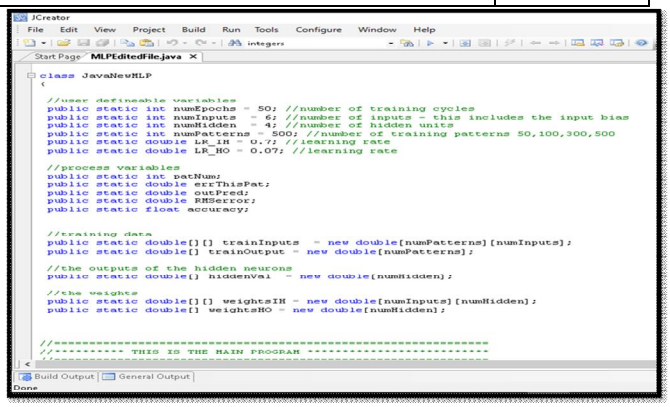

Figure 2: The MLP source code in JCreator LE.

Figure 2 shows the MLP source code in JCreator LE that been utilize for the data analysis purpose. Some of the variables that are used in this code are *numEpochs*, *numInputs*, *mum hidden*, *numPatterns*, *LR_IH*, and *LR_HO*. The first term, *numEpochs*, is defined as how many times data are trained. In the neural network, one epoch means one forward pass and one backward pass for the whole set of data. It cannot be sure whether 10 epochs or 100 epochs is sufficient for the data to be well-trained. In this code, the number of training cycles as 50 at first. Then, the number of training patterns are raised to

observe the performance. For *numInputs,* it means how many inputs that are inserted.

For this research, there are 5 tokens, or also known as inputs. In a neural network, a bias which is always 1 is inserted in the number of inputs to create a hyperbolic tangent (tanh) curve with the range from -1 to 1, so that it can handle the value of the sentiment of the data which are -1, 0 and 1. Therefore, the number of inputs in this research is 6, which includes bias. Next is the *numHidden,* explained as several hidden units. Apart from that is the number of training patterns, which is *numPatterns.* The significance of the term is the number of patterns that can be trained, which can be any number until 1000 because there are only 1000 data. Plus, the meaning for the term *LR_IH* is the learning rate from the input layer to the hidden layer while LR_HO is the learning rate from the hidden layer to the output layer. The learning rate set for *LR_IH* is 0.7 and the learning rate for *LR_HO* is 0.07. The learning rate is also known as the weight in the neural network diagram as shown in Figure 3.

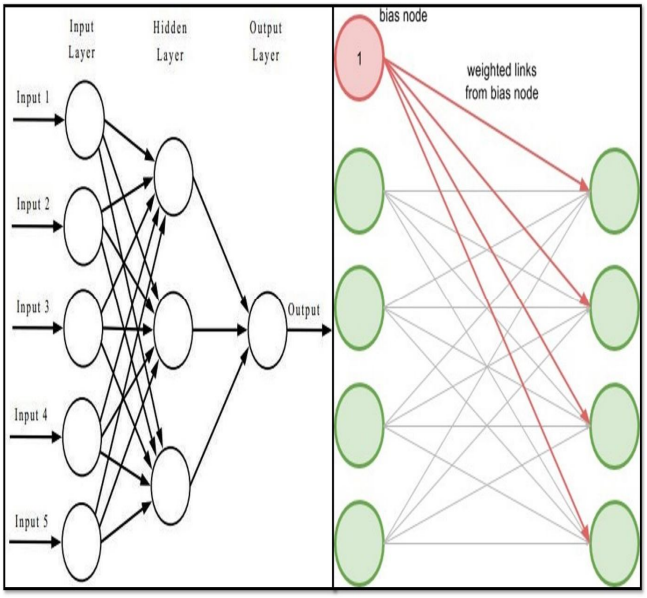

Figure 3: The Artificial Neural Network Diagram with Bias node sample

Figure 3 shows the artificial neural network diagram with the bias node sample. The performance of this method can be observed by looking at the different number of training patterns. For this research, the number of training patterns chosen is 50, 100, 200, 300, 400, and 500. Table 2 is the Artificial Neural Network (ANN) Model Prediction table for 50 patterns, their actual sentiment value and the Artificial Neural Network Prediction value, and the status whether it is Correct or Wrong.

Table 2: The ANN Model Prediction for first 50 Patterns

Patter	Actual	ANN Model	Final	Status
n		Prediction		
1	-1	-0.877928593	-1	Correct
2	0	-0.46871618	0	Correct
3	-1	-0.962273327	-1	Correct
4	1	1.026502288	1	Correct
5	-1	-0.853502349	-1	Correct
6	1	1.026488646	1	Correct
7	0	0.825133267	1	Wrong
8	1	0.875166093	1	Correct
9	1	0.940172608	1	Correct
10	-1	-0.849268001	-1	Correct
11	1	1.026805407	1	Correct
12	1	1.025979735	1	Correct
13	1	0.809613603	1	Correct
14	1	1.026769593	1	Correct
15	1	0.608767767	1	Correct
16	1	1.026835039	1	Correct
17	1	0.333243477	0	Wrong
18	1	1.026839489	1	Correct
19	1	1.026242709	1	Correct
20	0	0.355435571	0	Correct
21	-1	-0.421700227	0	Wrong
22	-1	-0.134637251	0	Wrong
23	1	1.026769593	1	Correct
24	1	1.026488646	1	Correct
25	1	1.026688511	1	Correct
26	1	1.026839489	1	Correct
27	0	0.176021747	0	Correct
28	1	0.986868751	1	Correct
29	1	1.026502288	1	Correct
30	1	1.001415662	1	Correc

				t
31	1	0.338600892	0	Wrong
32	-1	-0.929553852	-1	Correct
33	1	1.026839489	1	Correct
34	1	0.868098436	1	Correct
35	1	0.901873212	1	Correct
36	1	1.026834078	1	Correct
37	1	1.010272152	1	Correct
38	1	1.022762247	1	Correct
39	1	0.995194735	1	Correct
40	1	1.006886955	1	Correct
41	1	1.008863713	1	Correct
42	1	0.987463394	1	Correct
43	1	0.914498317	1	Correct
44	-1	-0.48910277	0	Wrong
45	1	0.836121501	1	Correct
46	-1	-0.424680518	0	Wrong
47	1	0.995194735	1	Correct
48	-1	-0.123619005	0	Wrong
49	1	1.022944187	1	Correct
50	0	1.026823769	1	Wrong

Table 2 summarizes the ANN model prediction when *numPatterns* = **50**. There are 9 incorrect predictions of sentiment value for 50 patterns. This proves that the method is showing a good performance in precision. Table 3 is the ANN Model Prediction table for 100 patterns, their actual sentiment value, and the Artificial Neural Network Prediction value and the status whether it is Correct or Wrong.

Table 3: The ANN Model Prediction for first 100 Patterns

Pattern	Actual	ANN Model Prediction	Final	Status
1	-1	-0.647112108	-1	Correct
2	0	-0.437728575	0	Correct
3	-1	-1.10199158	-1	Correct
4	1	0.980036544	1	Correct
5	-1	-1.096011071	-1	Correct
6	1	0.979950878	1	Correct
7	0	0.889833041	1	Wrong
8	1	0.97988345	1	Correct
9	1	0.980450126	1	Correct
10	-1	-0.640969349	-1	Correct
11	1	0.979903475	1	Correct
12	1	0.97991431	1	Correct
13	1	0.791245013	1	Correct
14	1	0.979914687	1	Correct
15	1	0.904597236	1	Correct
16	1	0.979903681	1	Correct
17	1	0.406006526	0	Wrong
18	1	0.979903313	1	Correct
19	1	0.979902883	1	Correct
20	0	0.661714718	1	Wrong
21	-1	-0.682780205	-1	Correct
22	-1	-1.074875115	-1	Correct
23	1	0.979914687	1	Correct
24	1	0.979950878	1	Correct
25	1	0.979903303	1	Correct
26	1	0.979903313	1	Correct
27	0	0.983469859	1	Wrong
28	1	0.976664288	1	Correct
29	1	0.980036544	1	Correct
30	1	0.979918377	1	Correct
31	1	0.637090384	1	Correct
32	-1	-1.146961964	-1	Correct
33	1	0.979903313	1	Correct
34	1	0.854541647	1	Correct

35	1	0.979500416	1	Correct
36	1	0.979903342	1	Correct
37	1	0.983941466	1	Correct
38	1	0.979946598	1	Correct
39	1	0.979907008	1	Correct
40	1	0.966745736	1	Correct
41	1	0.979333553	1	Correct
42	1	0.979717145	1	Correct
43	1	0.978988242	1	Correct
44	-1	-0.724602228	-1	Correct
45	1	0.971124312	1	Correct
46	-1	-0.240048058	0	Wrong
47	1	0.979907008	1	Correct
48	-1	-0.930924969	-1	Correct
49	1	0.979903193	1	Correct
50	0	0.979903602	1	Wrong
51	-1	-0.759480771	-1	Correct
52	0	-0.317123536	0	Correct
53	0	0.979903313	1	Wrong
54	0	0.979937426	1	Wrong
55	1	0.406006526	0	Wrong
56	1	0.979673878	1	Correct
57	0	0.665352562	1	Wrong
58	1	0.955374105	1	Correct
59	1	0.980036544	1	Correct
60	1	0.979903207	1	Correct
61	1	0.979903681	1	Correct
62	1	0.979907575	1	Correct
63	1	0.406006526	0	Wrong
64	1	0.980263482	1	Correct
65	0	0.689803612	1	Wrong
66	0	0.979717145	1	Wrong
67	1	0.978856775	1	Correct
68	1	0.979906116	1	Correct
69	1	0.979630655	1	Correct
70	1	-0.194234196	0	Wrong
71	1	0.979896373	1	Correct
72	1	0.947694177	1	Correct
73	-1	-1.114371383	-1	Correct
74	0	0.979892841	1	Wrong
75	1	0.653437441	1	Correct
76	1	0.979903342	1	Correct
77	1	0.979903321	1	Correct
78	1	-0.240048058	0	Wrong
79	1	0.979903313	1	Correct
80	1	1.023167432	1	Correct
81	0	0.97990222	1	Wrong
82	1	0.979903303	1	Correct
83	1	1.023440299	1	Correct
84	1	0.97990222	1	Correct
85	1	0.623256503	1	Correct
86	1	0.980036544	1	Correct
87	1	0.980036544	1	Correct
88	1	0.979903207	1	Correct
89	1	0.974085559	1	Correct
90	1	0.979903313	1	Correct
91	1	0.979950878	1	Correct
92	1	0.979903602	1	Correct
93	1	0.979900288	1	Correct
94	1	0.931460419	1	Correct
95	1	0.617797026	1	Correct
96	0	-0.365901885	0	Correct
97	0	0.995400908	1	Wrong

98	1	0.668326211	1	Correct
99	1	0.979907008	1	Correct
100	1	0.109519176	0	Wrong

Table 3 summarizes the ANN model prediction when *numPatterns* = **100**. There are 19 incorrect predictions of sentiment value for 100 patterns. This proves that the method is good in terms of performance in precision. Furthermore, we will tabulate the Root Mean Square (RMS) error for 50 epoch as shown in Table 4.

Table 4: The RMS Error Based on Default Epoch 50.

Epoch	RMS Error
0	0.232256681
1	0.169777041
2	0.136815503
3	0.142983149
4	0.121719607
5	0.127634747
6	0.13245719
7	0.119434683
8	0.116522973
9	0.12212544
10	0.11380472
11	0.118205043
12	0.113363844
13	0.11512509
14	0.113989982
15	0.105713225
16	0.105983094
17	0.107277329
18	0.105867267
19	0.10567526
20	0.105668706
21	0.107487643
22	0.115689916
23	0.123468761
24	0.12601211
25	0.124700187
26	0.132661519
27	0.133530619
28	0.113846899
29	0.110621334
30	0.122383744
31	0.106180001
32	0.108704947
33	0.104217983
34	0.103365898
35	0.130305335
36	0.104847937
37	0.122885893
38	0.123343549
39	0.116702608
40	0.121160942
41	0.108620215
42	0.106167749
43	0.108385342
44	0.101760424
45	0.100882879
46	0.101595852
47	0.12950327
48	0.129572958
49	0.100759955
50	0.10108852

Table 4 summarizes the RMS error based on default epoch 50 is 0.10108852. We utilize the RMS error to measure the difference between the fitted line to data points and the

difference between the Artificial Neural Network model prediction on sentiment value and the genuine sentiment value over the number of training patterns. Furthermore, the best performance is where the RMS Error is the most minimal. Therefore, we have tabulated the RMS Error based on a different number of epochs as shown in Table 5.

Table 5: The RMS Error Based on Epochs

Number of Patterns	Number of Epochs	Precision Percentage	RMS Error
500	**50**	74%	0.45654375746099235
	100	76%	0.45929934502840647
	200	76%	0.4773323848486768
	300	72%	0.4450825489116451
	400	74%	0.4517595588134799
	500	77%	0.47461938184957225

Table 5 summarizes the RMS error based on epochs. We have tabulated the comparison table between the methods, which are the decision tree, simple summation, and artificial neural network. The selected epochs are set as 50 and the number of patterns is 50, 100, 200, 300, 400, and 500 as shown in Table 6.

Table 6: The Precision Percentage Comparison.

Number of Epochs	Number of Patterns	ID3-1	ID3-2	Simple Summation	Artificial Neural Network
50	**50**	42%	52%	66%	86.0%
	100	52%	48%	52%	88.0%
	200	42%	35%	35%	72.0%
	300	36%	32%	31%	73.0%
	400	33%	29%	28%	73.5%
	500	33%	27%	27%	73.4%

Table 6 summarizes the precision percentage comparison. We can conclude that the best method for sentiment analysis by using Artificial Neural Network where when the first 500 patterns are calculated its sentiment analysis, the ID3-1 precision percentage is 33%, ID3-2 is 27%, the simple summation is 27% and Artificial Neural Network is 73.4%. It is proven that Artificial Neural Network is the overall best method to predict sentiment analysis.

5. CONCLUSION

Artificial Neural Network (ANN) is the highest according to the precision percentage. In the future, coding that can calculate the precision percentage of testing data can be created precisely. Moreover, can be developed a system with high precision by improving and adding new features such as a website system that can predict Malay sentences sentiment analysis just by typing in can increase the value of this research and is not impossible to proceed in the future. It will ensure that the system is more reliable and can be used and high in user satisfaction. We aim to describe the outcome of using the tool called JCreator LE for the Multi-Layer Perceptron coding.

Apart from that, to define a scope for tweets data that are retrieved from Twitter. As for now, there is still no coding for the scope to fetch together with Malay language and Malaysia as the location in one single source code. Sentences should be in scope so that the training and the testing will be easier and more accurate. Sentences discussing different topics are more difficult to handle because the testing data is not compatible with the training data. Further research needed in the future to overcome these scenarios.

ACKNOWLEDGMENT

The authors would like to thank the editor and reviewers for their recommendation to strengthen the standard of this paper. Our thanks are also forwarded to the Centre for Advanced Computing Technology, Faculty of Information and Communication Technology, UTeM for the encouragement and support for the work and this publication.

REFERENCES

[1] J. Zhu, Y. Wang, and C. Wang, "A comparative study of the effects of different factors on firm technological innovation performance in different high-tech industries," *Chinese Manag. Stud.*, vol. 13, no. 1, pp. 2–25, Apr. 2019. https://doi.org/10.1108/CMS-10-2017-0287

[2] M. B. A.M., "Machine Learning-Structural Equation Modeling Algorithm: The Moderating role of Loyalty on Customer Retention towards Online Shopping," *Int. J. Emerg. Trends Eng. Res.*, vol. 8, no. 5, pp. 1578–1585, May 2020. https://doi.org/10.30534/ijeter/2020/17852020

[3] A. K. Mohamad, M. Jayakrishnan, and N. H. Nawi, "Employ Twitter Data to Perform Sentiment Analysis in the Malay Language," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 2, pp. 1404–1412, 2020. https://doi.org/10.30534/ijatcse/2020/76922020

[4] S. Tofighy and S. M. Fakhrahmad, "A proposed scheme for sentiment analysis," *Kybernetes*, vol. 47, no. 5, pp. 957–984, May 2018.

[5] I. Xie and J. A. Stevenson, "@Digital libraries: harnessing Twitter to build online communities," *Online Inf. Rev.*, vol. 43, no. 7, pp. 1263–1283, Nov. 2019.

[6] V. Shailaja, "Predictive Analytics of Performance of India in the Olympics using Machine Learning Algorithms," *Int. J. Emerg. Trends Eng. Res.*, vol. 8, no. 5, pp. 1829–1833, May 2020.

https://doi.org/10.30534/ijeter/2020/57852020

[7] H. K, "Efficient Image Compression by Machine Learning," *Int. J. Emerg. Trends Eng. Res.*, vol. 8, no. 5, pp. 1672–1677, May 2020. https://doi.org/10.30534/ijeter/2020/29852020

[8] N. Khamis, "Corpus-based Data for Determining Specialised Language Features," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 1, pp. 36–41, Feb. 2020. https://doi.org/10.30534/ijatcse/2020/07912020

[9] G. J. Wu, Z. "Jimmy" Xu, S. Tajdini, J. Zhang, and L. Song, "Unlocking value through an extended social media analytics framework," *Qual. Mark. Res. An Int. J.*, vol. 22, no. 2, pp. 161–179, Apr. 2019.

[10] Z. Huang and Y. Liang, "Research of data mining and web technology in university discipline construction decision support system based on MVC model," *Libr. Hi Tech*, p. LHT-09-2018-0131, Jun. 2019.

[11] G. Casalino, C. Castiello, N. Del Buono, and C. Mencar, "A framework for intelligent Twitter data analysis with non-negative matrix factorization," *Int. J. Web Inf. Syst.*, vol. 14, no. 3, pp. 334–356, Aug. 2018.

[12] Z. Zhang and Y. Dai, "Combination classification method for customer relationship management," *Asia Pacific J. Mark. Logist.*, vol. ahead-of-p, no. ahead-of-print, Jul. 2019. https://doi.org/10.1108/APJML-03-2019-0125

[13] C. Udanor and C. C. Anyanwu, "Combating the challenges of social media hate speech in a polarized society," *Data Technol. Appl.*, vol. 53, no. 4, pp. 501–527, Sep. 2019.

[14] V. Diamantopoulou and H. Mouratidis, "Applying the physics of notation to the evaluation of a security and privacy requirements engineering methodology," *Inf. Comput. Secur.*, vol. 26, no. 4, pp. 382–400, Oct. 2018.

[15] B. M. I. I. Hughes, M.A., Skorpik, J.R., Brambley, M.R., Gonzalez, E.G. and Huang, Y., "Building environment data collection systems.," 10/161,833., 2018.

[16] D. Levine, S., Pastor, P., Krizhevsky, A., Ibarz, J. and Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection.," *Int. J. Rob. Res.*, vol. 37, no. 4, pp. 421–436, 2018. https://doi.org/10.1177/0278364917710318

[17] E. Mitzenmacher, M. and Upfal, *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis.* Cambridge university press., 2017.

[18] B. W. Silverman, *Density estimation for statistics and data analysis.* Routledge., 2018.

[19] D. C. Montgomery, *Design and analysis of experiments.* John Wiley & Sons., 2017.

[20] D. J. Curtis, M.J., Alexander, S., Cirino, G., Docherty, J.R., George, C.H., Giembycz, M.A., Hoyer, D., Insel, P.A., Izzo, A.A., Ji, Y. and MacEwan, "Experimental design and analysis and their reporting II: updated and simplified guidance for authors and peer reviewers.," *Br. J. Pharmacol.*, vol. 175, no. 7, pp. 987–993, 2018.

[21] M. Almseidin, M., Alzubi, M., Kovacs, S. and Alkasassbeh, "Evaluation of machine learning algorithms for intrusion detection system.," in *In 2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY)*, 2017, pp. 000277–000282.

[22] H. Sun, J., Lang, J., Fujita, H. and Li, "Imbalanced enterprise credit evaluation with DTE-SBD: decision tree ensemble based on SMOTE and bagging with differentiated sampling rates.," *Inf. Sci. (Ny).*, vol. 42, no. 5, pp. 76–91, 2018. https://doi.org/10.1016/j.ins.2017.10.017

[23] Susan Lang and Craig Baehr, "Data Mining: A Hybrid Methodology for Complex and Dynamic Research," *Coll. Compos. Commun.*, vol. 1, no. September, pp. 172–194, 2012.

[24] H. Akbar, M.A., Sang, J., Khan, A.A., Shafiq, M., Hussain, S., Hu, H., Elahi, M. and Xiang, "Improving the quality of software development process by introducing a new methodology–AZ-model.," *IEEE Access*, vol. 6, no. 1, pp. 4811–4823, 2017. https://doi.org/10.1109/ACCESS.2017.2787981