Combining Sentiment Lexicons and Content-Based Features for Depression Detection

Raymond Chiong

Gregorious Satia Budhi

Sandeep Dhakal
The University of Newcastle, Australia

***Abstract*—Numerous studies on mental depression have found that tweets posted by users with major depressive disorder could be utilized for depression detection. The potential of sentiment analysis for detecting depression through an analysis of social media messages has brought increasing attention to this field. In this work, we propose 90 unique features as input to a machine learning classifier framework for detecting depression using social media texts. Derived from a combination of feature extraction approaches using sentiment lexicons and textual contents, these features are able to provide impressive results in terms of depression detection. While the performance of different feature groups varied, the combination of all features resulted in accuracies greater than 96% for all standard single classifiers and the best accuracy of over 98% with Gradient Boosting, an ensemble classifier.**

■ **THERE IS** a general agreement in the relevant literature that social media platforms, by allowing people to express their feelings or share their ideas and thoughts more freely, have become a vital source for monitoring health issues and trends [1], [2]. Posts on platforms, such as Twitter and Facebook, enable researchers to investigate multiple patterns of human behavior and their psychology [3]. Several studies on mental depression—a medical illness with symptoms such as persistent sadness, loss of interest, and an inability to carry out normal activities [4]—have found that tweets posted by users with major depressive disorder could be utilized to predict the possibility of future episodes of depression in those users [5], [6], [7], [8], [9]. Sentiment analysis, which is an automatic and systematic process of detecting the sentiment or emotional

© 2021 ACSA

tone of a given text, has been identified by various studies as a potential mechanism for detecting signs of depressive disorder [10], [11], [12]. Sentiment analysis has previously been successfully applied to predict the sentiment or emotional tone behind social media messages, online reviews or any other types of text messages [13], [14], [15], [16]. In addition to the detection algorithm applied, the performance of sentiment analysis is also significantly influenced by the features selected [17], [18], [19]. Therefore, in this study, we propose 90 unique features, through a combination of feature extraction using sentiment lexicons and content-based features from the social media messages themselves. Two sentiment lexicons, namely SentiWordNet [20] and SenticNet [21], are used for feature extraction. Similarly, the content-based features utilized for depression detection are formulated from the characteristics of the Twitter message content (e.g., the number of words, sentences, questions, exclamations), part-of-speech (POS) tags, linguistic traits, and readability scores. The combined features are then used as input for several machine learning models trained using publicly labeled depression/non-depression datasets comprising of tweets [6]. Results of our extensive experiments confirm the effectiveness of these features for depression detection. The rest of this paper is organized as follows: the following two sections discuss the datasets used and the design of the input features; next, we provide details about our framework and the measurements used; then, experimental results and discussions are presented; finally, we conclude the paper and highlight future research directions.

1. DATASETS

Two depression datasets, comprising of Twitter posts that have automatically been labeled as either ‘Depression’ or ‘Non-Depression’, were used for all the experiments in this study (Table 1). These datasets were used to train and test the proposed featuring approach for several machine learning models using 10-fold cross-validation. The first dataset, by Shen et al. [6], was constructed with the restriction that a record would be labeled as ‘Depression’ only if its anchor tweets satisfied the strict pattern “(I’m/I was/I am/I’ve been) diagnosed with depression”; the record would be labeled as ‘Non-Depression’ if the user had never posted any tweet containing the character string ‘depress’. Eye’s dataset[1], on the other hand, is less restrictive and was built by seeking the word ‘depression’ in the tweets. Any tweet containing the word ‘depression’ was labeled as ‘Depression’, and ‘Non-Depression’ otherwise. Eye’s dataset is highly imbalanced; depression class records account for only 22% of the total records.

Table 1. Datasets used in this study.

Dataset	Records		
	Total	Depression	Non-depression
Shen et al.	11877	54.67%	45.33%
Eye et al.	10314	22.44%	77.56%

Both are labeled and comprised of Twitter posts.

2. DESIGN OF INPUT FEATURES

The input features in this study have been defined based on two sentiment lexicons: SentiWordNet [20] and SenticNet [21]. These input features are categorized into three groups, namely Groups A, B and C (Table 2). Group A consists of 9 features created using SentiWordNet, whereas Group B consists of the same features extracted using SenticNet. Group C includes 4 features that were directly extracted using some sentiment values in SenticNet and represent the total introspection, temper, attitude, and sensitivity values of the terms in the text. The features from SenticNet have been split into Groups B and C to facilitate a fairer comparison of the effectiveness of the two lexicons for depression detection. Since SenticNet has four additional features, the initial comparison is first conducted using Groups A and B (same 9 features), following which the effect of the additional features in SenticNet (Group C) is investigated. To improve detection, another 68 features have been defined based on our previous study [22] (Table 2). These features were extracted based on the characteristics of the tweets, and are categorized into four groups (D, E, F, and G) as follows. The features in Group D are related to basic information that can be extracted from the text; Group E consists of 36 POS tags based on Penn POS [23]; Group F captures the linguistic traits of the text; and Group G is related to the readability of the text. Groups

[1] http://kaggle.com/bababullseye/depression-analysis

D-F were extracted using the Natural Language Toolkit [24] and additional custom functions and formulas written in Python, whereas the features in Group G, representing the readability scores, were extracted using functions from the TextStat project[2].

3. FRAMEWORK

Our framework, as depicted in Fig. 1, is straightforward. Once the dataset(s) and settings have been loaded, the input features are extracted based on the group settings. All input features are subsequently normalized to a scale of 0 to 1 using min-max normalization. Since all attributes have differing ranges, normalization ensures that all features have equal contribution toward the detection. Following the creation of training targets, the n-fold cross-validation process is run according to the assigned classifiers. Finally, the best classifier for detecting depression is determined, and all information and the detailed results are written to a file. In this study, we implemented and tested four standard single classifiers—Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), and Multilayer Perceptron (MLP)—and four ensemble models—Bagging Predictors (BP), Random Forest (RF), Adaptive Boosting (AB), and Gradient Boosting (GB)—for detecting depression from Twitter posts. These classifiers are often used in text analysis and have produced excellent performance in previous studies on textual-based sentiment analysis [19] and malicious web domain identification [25]. The performance of the featuring approach with the above classifiers was assessed using four common measurements for prediction or classification (Table 3): accuracy, precision, recall and F-measure (also known as F1 score). All machine learning classifiers, ensemble models and measurements were built using scikit-learn components [26]. Default parameters were used for all classification models to ensure that the results can only be affected by the implementation of our approach and not by the modification of classifier parameters.

4. EFFECTS OF SENTIMENT LEXICON FEATURES

In this section, we present the results of our investigation into the effects of sentiment lexicon feature groups on prediction performance. The 10-fold cross-validation experiments were run on the LR classifier, which was identified as one of the best classifiers in previous experiments [19], using the two datasets described above (Table 1). The results in Table 4 clearly indicate that, when similar sentiment features were compared (Group A vs. Group B), the features extracted using SenticNet (Group B) outperformed the features extracted using SentiWordNet (Group A) for both the datasets. Thus, we can conclude that the sentiment terms in SenticNet and their sentiment scores are more suitable for depression detection in Twitter texts. The performance of Group C, which consists of additional features that could only be provided by SenticNet, was also satisfactory. The accuracy, precision, recall and F1 scores for Group C were above 50% for Shen et al.'s dataset. In the case of Eye's dataset, the accuracy was even better ($> 81\%$), but the recall and F1 scores were much worse. These results indicate that the features in Group C are not suitable for detecting the target class, i.e., the depression class (in binary classification, recall and the accuracy of the target class are the same). It should be noted, however, that the results from combining Group C with other groups were marginally better than without Group C. Between the two datasets, the results show that the accuracy was always higher for Eye's dataset compared to Shen et al.'s dataset. However, the recall and F1 scores were always lower for Eye's dataset. This implies that the classifier trained using Eye's samples found it difficult to detect the target class (i.e., depression class) than the other class, and we suspect that this is due to the imbalanced nature of Eye's dataset (Table 1). This problem could be easily solved by applying sampling methods [25], [27] to the dataset, but in this study, we attempt to overcome the problem by implementing ensemble models (Fig. 3b).

5. ADDITIONAL FEATURES BASED ON THE CONTENT-BASED APPROACH

The above results demonstrated that sentiment lexicon features can perform well in terms of detecting depression from Twitter posts. Next, we explore whether content-based features (Groups D, E, F and G) could further improve perfor-

[2] http://pypi.org/project/textstat

Table 2. Features.

Group	No.	Description
Sentiment lexicon features		
A: Sentiment lexicon features based on SentiWordNet	1 2-4 5, 6 7 8, 9	Total of sentiment items Total of (positive, neutral, negative) sentiment terms The ratio of (positive, negative) sentiment to neutral terms The ratio of negative to positive sentiment terms (Positive, negative) sentiment scores
B: Sentiment lexicon features based on SenticNet	10 11-13 14, 15 16 17, 18	Total of sentiment terms Total of (positive, neutral, negative) sentiment terms The ratio of (positive, negative) sentiment to neutral terms The ratio of negative to positive sentiment terms (Positive, negative) sentiment scores
C: Additional lexicon features based on SenticNet	19 20 21 22	Total introspection value Total temper value Total attitude value Total sensitivity value
Content-based features		
D: Basic text information	23-26 27 28 29 30, 31 32	Total (letters, words, stop words, sentences) in the text Total words with capitalized 1st letter Total negative terms (e.g., 'does not', 'do not', 'will not', etc.) Total elongated words (e.g., 'yesss', 'fiiine', 'yoouu', etc.) Total exclamation and question sentences The existence of weblink inside the text
E: POS	33-68	Total existence of 36 Tags of Penn POS
F: Linguistic characteristics	69 70 71 72 73 74 75-77 78-80	The ratio of adjectives and adverbs Average of number of words per sentence The ratio of word repetition to total words The average number of letters per word Average of words with 1st capital to total sentences The ratio of words with 1st capital to total words Total of (1st, 2nd, 3rd) person pronouns The ratio of (1st, 2nd, 3rd) person pronouns to total pronouns
G: Readability scores	81-87 88 89 90	Flesch Reading Ease, Simple Measure of Gobbledygook Index, Flesch Kincaid Grade, Coleman-Liau Index, Gunning Fog Index, DaleChall Readability and Linsear Write Formula. Automated Readability Index Difficult words Estimation of school grade level required to understand the text.

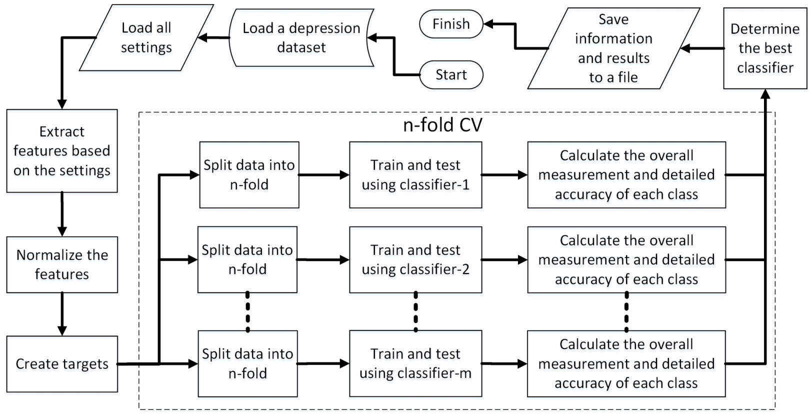

Load all settings
Load a depression dataset
Finish
Start
Save information and results to a file
Determine the best classifier
Extract features based on the settings
Normalize the features
Create targets
n-fold CV
Split data into n-fold
Train and test using classifier-1
Calculate the overall measurement and detailed accuracy of each class
Split data into n-fold
Train and test using classifier-2
Calculate the overall measurement and detailed accuracy of each class
Split data into n-fold
Train and test using classifier-m
Calculate the overall measurement and detailed accuracy of each class

Figure 1. Design of the proposed framework used for detecting depression in this study.

mance. As above, we conducted experiments on the LR classifier but, based on the above results, used only Shen et al.'s dataset to train it. The results in Table 5 show that each content-based group improved the detection measurements when combined with sentiment lexicon features. Group E (POS) provided the best improvement, followed by Group G (readability scores), Group D (basic text information), and Group F (linguistic characteristics). However, the overall best improvement was achieved when all sentiment lexicon and content-based features were used at the same time; all measurements were higher than 95%. It is also worth mentioning that the F1 scores obtained with our approach are better than the baseline results (85%) in Shen et al. [6].

6. IDENTIFYING THE BEST CLASSIFIER

The following set of experiments was conducted with the best feature setting from the above experiments (Groups A to G) on all single

Table 3. Measurement functions and formulas.

Name	Function	Formula
Accuracy	accuracy_score()	$Accuracy(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} 1(\hat{y_i} = y_i)$, where y is the set of predicted pairs, $\hat{y}$ is the set of true pairs, and $n_{samples}$ is the total number of samples.
Precision	precision_score()	$Precision(y_i, \hat{y_i}) = \frac{TP}{TP+FP}$, where i is the set of classes, y_i is the subset of y with class i, TP is true positive, and FP is false positive.
Recall	recall_score()	$Recall(y_i, \hat{y_i}) = \frac{TP}{TP+FN}$, where FN is false negative.
F-measure/F1	f1_score()	$F1(y_i, \hat{y_i}) = \frac{2 \times Precision(y_i, \hat{y_i}) \times Recall(y_i, \hat{y_i})}{Precision(y_i, \hat{y_i}) + Recall(y_i, \hat{y_i})}$.

Table 4. Effects of sentiment lexicon features on depression detection.

Sentiment lexicon group(s)	Dataset	Measurements (%) *				Class accuracy(%) *	
		Acc	Pre	Rec	F1	Dep	Non-Dep
A	Eye's	82.33	74.04	32.76	45.36	32.76	96.68
	Shen et al.'s	76.50	78.79	78.07	78.41	78.07	74.61
B	Eye's	88.51	82.74	61.59	70.57	61.59	96.28
	Shen et al.'s	80.34	83.80	79.37	81.52	79.37	81.49
C	Eye's	81.30	84.98	20.32	32.70	20.32	98.94
	Shen et al.'s	63.80	64.42	75.50	69.50	75.50	49.71
B, C	Eye's	89.03	83.67	63.49	72.16	63.49	96.42
	Shen et al.'s	84.20	85.76	85.28	85.51	85.28	82.91
A, B, C	Eye's	89.58	84.09	66.11	73.96	66.11	96.38
	Shen et al.'s	84.91	85.75	86.82	86.27	86.82	82.61

*: Acc = Accuracy; Pre = Precision; Rec = Recall; F1 = F-measure; Dep = Depression; Non-Dep = Non-Depression.

Table 5. Effects of content-based features on depression detection when trained using Shen et al.'s dataset.

Sentiment lexicon group(s)	Content-based group(s)	Measurements (%) *				Class accuracy(%) *	
		Acc	Pre	Rec	F1	Dep	Non-Dep
A, B, C	D	88.73	89.21	90.31	89.75	90.31	86.82
A, B, C	E	94.62	94.51	95.72	95.11	95.72	93.31
A, B, C	F	86.63	87.29	88.40	87.84	88.40	84.48
A, B, C	G	91.30	90.58	93.85	92.18	93.85	88.23
A, B, C	D, E	95.21	95.10	96.19	95.64	96.19	94.02
A, B, C	D, F	88.99	89.55	90.42	89.97	90.42	87.27
A, B, C	D, G	92.62	92.31	94.37	93.32	94.37	90.51
A, B, C	D, E, F	95.32	95.19	96.30	95.74	96.30	94.14
A, B, C	D, E, G	96.34	96.37	96.95	96.66	96.95	95.59
A, B, C	D, F, G	92.71	92.50	94.32	93.40	94.32	90.77
A, B, C	D, E, F, G	96.50	96.48	97.14	96.81	97.14	95.73

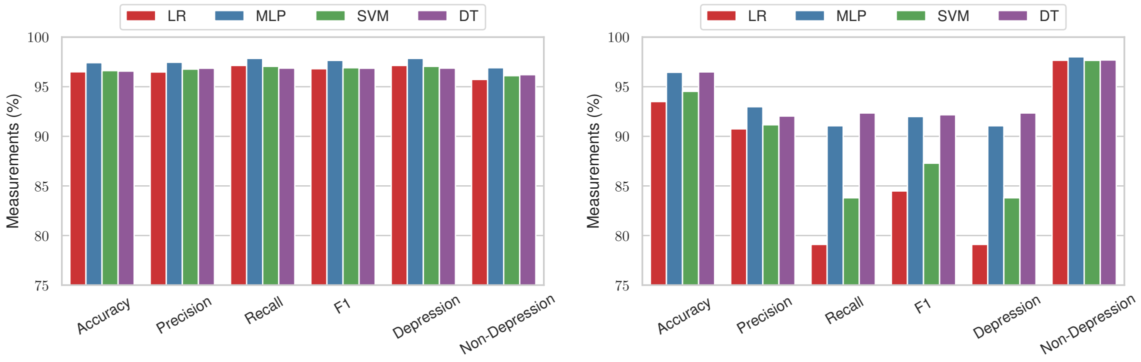
*: Acc = Accuracy; Pre = Precision; Rec = Recall; F1 = F-measure; Dep = Depression; Non-Dep = Non-Depression.

classifiers (LR, MLP, SVM, and DT) and both datasets to further investigate the performance of the best feature setting. We can see in Fig. 2a and Fig. 2b that the features performed better with other single classifiers than LR. The MLP was the best single classifier for Shen et al.'s dataset, whereas the DT was the best for Eye's dataset.

It is important to note that both the MLP and DT performed significantly better than other classifiers in terms of recall; their recall scores were at least 90%, for Eye's dataset. Similar experiments were also conducted with the ensemble models (AB, BP, GB, and RF) for both datasets, and the results can be seen in Fig. 3a and Fig. 3b. The results show that all ensemble models performed well on both datasets; GB achieved the highest measurements for both datasets, with an accuracy of more than 98%. The results in Fig. 3b also show that all ensemble models provided recall values of around 95% despite Eye's dataset being heavily imbalanced, thus, obviating the need for any further action to overcome the class imbalance issue.

7. CONCLUSION AND FUTURE WORK

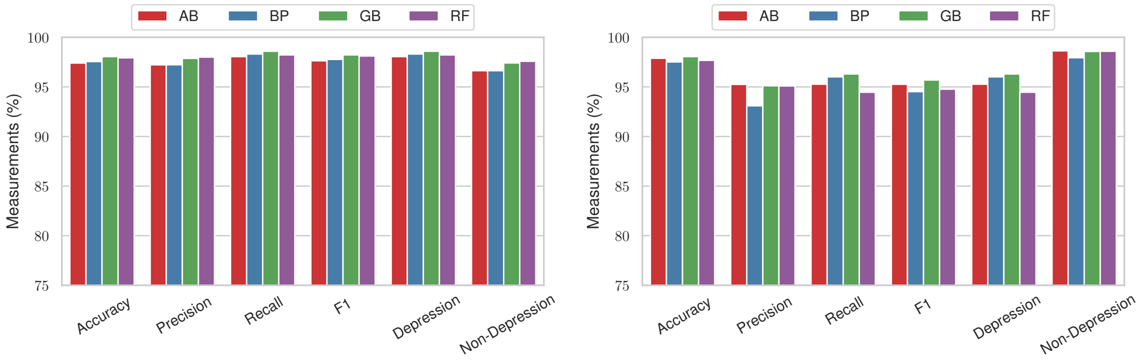
In this study, we proposed 90 different features that can be used by machine learning classifiers for detecting depression by analyzing the social media messages of users. These features were extracted using the combination of sentiment lexicons and content-based approaches. While our experiments were conducted using datasets comprising of Twitter posts, these features can be used for any textual content. Through extensive experiments, involving two datasets of Twitter posts, four single classifiers and four ensemble models, we were able to verify the effectiveness of these features.



(a) trained using Shen et al.'s dataset

(b) trained using Eye's dataset

Figure 2. Results for feature groups A-G on four single classifiers trained using two different datasets.



(a) trained using Shen et al.'s dataset

(b) trained using Eye's dataset

Figure 3. Results for feature groups A-G on ensemble models trained using two different datasets.

The best results were obtained when all the proposed features were utilized together for depression detection; however, the effectiveness of different feature groups greatly varied. In particular, the content-based features were able to improve the accuracy to >96% for both datasets. Whereas all single classifiers and ensemble models provided excellent results, the GB ensemble was able to provide accuracies >98% for both datasets. Our analysis also revealed that the ensemble models were able to overcome the data imbalance issue, which the single classifiers were unable to do. As future work, we plan to investigate a novel idea about the combination of multiple classifiers for improving accuracy. We will also investigate the possibility of using sentiment analysis datasets, which can be easily constructed in larger sizes, for depression detection in social media texts.

REFERENCES

1. O. Edo-Osagie et al., “A scoping review of the use of Twitter for public health research,” *Computers in Biology and Medicine*, vol. 122, 2020, 103770.
2. S. Ji et al., “Suicidal Ideation Detection: A Review of Machine Learning Methods and Applications,” *IEEE Transactions on Computational Social Systems*, vol. 8, no. 1, 2021, pp. 214–226.
3. J. Hussain et al., “Exploring the dominant features of social media for depression detection,” *Journal of Information Science*, vol. 46, no. 6, 2019, pp. 739–759.
4. S. A. Qureshi et al., “Multitask representation learning for multimodal estimation of depression level,” *IEEE Intelligent Systems*, vol. 34, no. 5, 2019, pp. 45–52.
5. M. D. Choudhury et al., “Predicting depression via social media,” *ICWSM*, 2013, pp. 128–137.
6. G. Shen et al., “Depression detection via harvesting social media: A multimodal dictionary learning solution,” *IJCAI*, 2017, pp. 3838–3844.

7. H. S. Alsagri and M. Ykhlef, “Machine learning-based approach for depression detection in Twitter using content and activity features,” *IEICE Trans on Information and Systems*, vol. E103.D, no. 8, 2020, pp. 1825–1832.
8. S. Ji et al., “Suicidal Ideation and Mental Disorder Detection with Attentive Relation Networks,” *Neural Computing and Applications*, vol. 33, 2021.
9. Q. Chen et al., “Sequential Fusion of Facial Appearance and Dynamics for Depression Recognition,” *Pattern Recognition Letters*, 2021.
10. A. U. Hassan et al., “Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression,” *ICTC*, 2017, pp. 138–140.
11. M. R. Islam et al., “Depression detection from social network data using machine learning techniques,” *Health Information Science and Systems*, vol. 6, 2018, 8.
12. Y. Chen et al., “Sentiment analysis based on deep learning and its application in screening for perinatal depression,” *IEEE DSC*, 2018, pp. 451–456.
13. S. L. Lo et al., “A multilingual semi-supervised approach in deriving Singlish sentic patterns for polarity detection,” *Knowledge-Based Systems*, vol. 105, 2016, pp. 236–247.
14. L. Yang et al., “Sentiment analysis for e-commerce product reviews in Chinese based on sentiment lexicon and deep learning,” *IEEE Access*, vol. 8, 2020, pp. 23522–23530.
15. Y. Susanto et al., “Ten Years of Sentic Computing,” *Cognitive Computation*, vol. 13, 2021.
16. F. Xing, F. Pallucchini, and E. Cambria, “Cognitive-Inspired Domain Adaptation of Sentiment Lexicons,” *Information Processing and Management*, vol. 56, no. 3, 2019, pp. 554–564.
17. A. Yousefpour, R. Ibrahim, and H. N. A. Hamed, “Ordinal-based and frequency-based integration of feature selection methods for sentiment analysis,” *Expert Systems with Applications*, vol. 75, 2017, pp. 80–93.
18. P. Bansal and R. Kaur, “Twitter sentiment analysis using machine learning and optimization techniques,” *International Journal of Computer Applications*, vol. 179, no. 19, 2018, pp. 5–8.
19. G. S. Budhi et al., “Using machine learning to predict the sentiment of online reviews: A new framework for comparative analysis,” *Archives of Computational Methods in Engineering*, 2021.
20. S. Baccianella, A. Esuli, and F. Sebastian, “SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining,” *LREC, vol. 10*, 2010, pp. 2200–2204.
21. E. Cambria et al., “SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis,” *CIKM*, 2020, pp. 105–114.
22. G. S. Budhi et al., “Using a hybrid content-based and behaviour-based featuring approach in a parallel environment to detect fake reviews,” *Electronic Commerce Research and Applications*, 2021.
23. S. Buchholz, *Memory-based grammatical relation finding*, Thesis, 2002.
24. S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, O’Reilly Media, Inc., USA, 2009, ISBN 9780596516499.
25. Z. Hu et al., “Malicious web domain identification using online credibility and performance data by considering the class imbalance issue,” *Industrial Management & Data Systems*, vol. 119, no. 3, 2019, pp. 676–696.
26. F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, no. 85, 2011, pp. 2825–2830.
27. G. S. Budhi, R. Chiong, and Z. Wang, “Resampling imbalanced data to detect fake reviews using machine learning classifiers and textual-based features,” *Multimedia Tools and Applications*, vol. 80, no. 9, 2021, pp. 13079–13097.

Raymond Chiong is the corresponding author and is currently an Associate Professor with the School of Electrical Engineering and Computing, The University of Newcastle, Australia. His research interests include machine learning, data analytics, evolutionary optimization, and modeling of complex adaptive systems. He is the Editor-in-Chief of the *Journal of Systems and Information Technology*, an Editor of *Engineering Applications of Artificial Intelligence*, and an Associate Editor of *Engineering Reports*. Contact him at Raymond.Chiong@newcastle.edu.au.

Gregorious Satia Budhi is currently a PhD Student with the School of Electrical Engineering and Computing, The University of Newcastle, Australia. He is also an academic staff member with Petra Christian University in Indonesia. His research interests include sentiment analysis, machine learning and data/text mining. Contact him at Gregorious.Satiabudhi@uon.edu.au.

Sandeep Dhakal is currently a PhD Student with the School of Electrical Engineering and Computing, The University of Newcastle, Australia. His research interests include modeling of complex adaptive systems, evolutionary game theory, and sentiment analysis. Contact him at Sandeep.Dhakal@newcastle.edu.au.