Utilizing Random Forest Algorithm for Sentiment Prediction Based on Twitter Data

Iwan Setiawan[1], Agung Mulyo Widodo[2], Mosiur Rahaman[3], Tugiman[4], Muhammad Abdullah Hadi[2], Nizirwan Anwar[2(✉)], Muhammad Bahrul Ulum[2], Erry Yudhya Mulyani[2], and Nixon Erzed[2]

[1] Nusa Putra University, Sukabumi 43152, Indonesia
iwansa@nusaputra.ac.id
[2] Esa Unggul University, Jakarta 11510, Indonesia
{agung.mulyo,nizirwan.anwar,m.bahrul_ulum,erry.yudhya,nikson}@esaunggul.ac.id,
muhammad.abdlhadi@student.esaunggul.ac.id
[3] Asia University, Taichung 413, Taiwan
mosiurrahaman@asia.edu.tw
[4] Buddhi Dharma University, Tangerang 15115, Indonesia

Abstract. Information sharing throughout the globe or universe has become a characteristic of social media. There has been a lot of research into the classification of sentiments. In this study, Twitter has been mined for unstructured GoFood Reviews data. It has been preprocessed to analyze the reviews' sentiment with polarity analysis, feature extraction with TF-IDF, and supervised learning with random forest. From June 1, 2022, to June 30, 2022, a total of 28763 tweets with the keyword GoFood were retrieved from Twitter. The data is processed by the Python programming language utilizing NLTK, Sastrawi for the Indonesian language, Textblob, TF-IDF, Random Forest Classification, and other algorithms. Twitter is a nearly limitless source for classifying text. This algorithm takes roughly five minutes to compute.

Keywords: Classification and Analysis of Sentiment · Random Forest Algorithm · Polarity Analysis · social media · Twitter

1 Introduction

Social media has gained notoriety in recent years for its ability to transmit information throughout the world or universe. They spread the information via Facebook, Twitter, and other social media platforms [1]. Through addition to conveying knowledge, teachers communicate their opinions on the subject in their comments, whether good or negative [2]. As e-commerce technology develops and spreads, an increasing number of consumers prefer to purchase on a variety of e-commerce platforms. In contrast to offline shopping in physical stores, customers can shop anytime, whenever, and do not need to wait until the weekend [3]. Since these platforms are virtual, there are a number of issues

I G. P. Suta Wijaya et al. (Eds.): MIMSE-I-C 2022, ACSR 102, pp. 446–456, 2022.
https://doi.org/10.2991/978-94-6463-084-8_37

with the things offered on them, including discrepancies between the descriptions of the goods and the real products, poor quality, inadequate after-sales service, and more [4].

The objective of sentiment analysis is to differentiate between text polarities. Sentiment analysis is one approach to the classification problem [2] In order to categorize the writing as positive, negative, or neutral, the work focuses mostly on three categories of issues. This analysis assists academics and decision-makers in gaining a deeper understanding of opinions and consumer satisfaction by automatically collecting diverse perspectives from a variety of platforms and classifying them according to sentiment. There is significant interest in the study of sentiment classification. Historically, most of it has concentrated on identifying longer works as reviews in [5]. Sentiment analysis for product reviews, also known as text orientation analysis or opinion mining, is the technique of automatically assessing the subjective comments text with the customer's emotional hue and determining their emotional propensity [6].

The benefits of sentiment analysis extend to both the provider and the client. Using user input on online stores or mobile applications, for instance, it enables the vendor to promote new products while assisting the client in locating unique items [7]. Subject classification and text analysis use significantly less vocabulary. In reality, it is frequently challenging to identify the dominant viewpoint in a text using the simplest methods, which merely use word frequency statistics [8]. This analysis uses unstructured GoFood Reviews data from Twitter. It has been filtered to remove distracting information such as emojis, typos, and improper punctuation. In the evaluation of the reviews' sentiment, feature extraction with TF-IDF, supervised learning with random forest, and polarity analysis were all employed during preprocessing.

2 Methodology

Figure 1 depicts the suggested method for sentiment analysis of tweets. Preprocessing processes include text cleaning and preprocessing. During the phase of training, tweets are trained. The concluding step will consist of the sentiment analysis of the test dataset.

2.1 Data Collection

Twitter was utilized to obtain data for sentiment analysis in this study. By utilizing the powerful search feature of the Python application SNScrape, we can more quickly harvest real-time user Tweet data based on keywords. Between June 1 and June 30, 2022, a total of 28763 tweets in Indonesian with the hashtag GoFood were collected from Twitter. The data is analyzed using Python programming, NLTK, Sastrawi for Indonesian, Textblob, TF-IDF, Random Forest Classification, and other tools. A number of steps are required to do an effective sentiment analysis. Collecting, pre-processing, feature extraction, and sentiment analysis are the steps.

2.2 E-Commerce

It is logical that if e-commerce increases rapidly, constraints that restrict its dynamics will emerge. Numerous studies have demonstrated that e-commerce enterprises face a

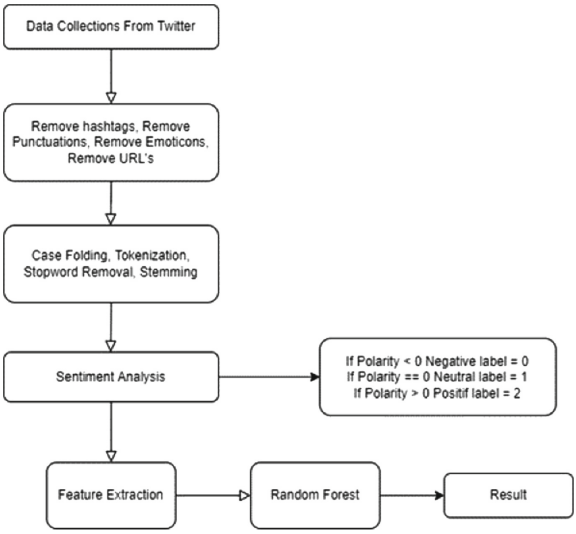

Fig. 1. Sentimen Analysis Process.

number of obstacles and problems. Customer service, supply chain, payment processing, tax and regulatory difficulties, infrastructure, the state of the economy, and marketing are the primary barriers limiting the e-commerce industry from fully utilizing the internet [9]. This is a less common scenario than that of food producers, who due to shorter delivery times and routes, typically sell their products to customers in their near vicinity. Deliveries of perishable food do not usually require particular routes [10]. Occasionally, a catalyst is necessary for the rise of the online food trade, such as an unanticipated event [11] or a global COVID-19 epidemic [12, 13].

2.3 Social Network

By providing communication services, social networking sites (SNS) aim to enhance social interaction between users [14]. Users of social networking sites create the sites' content. It contains a significant amount of user data, shared ideas and thoughts, and real-time information on user statuses and discussions. In addition to the rise of SNS users [9], the rate of data in SNSs demonstrates that SNSs play a vital role in real-time analysis and forecasting in numerous domains. Twitter, a social networking service, was founded formally on July 13, 2006 [15]. Twitter's principal function is tweeting, which can be done on a desktop or mobile device. A tweet may include no more than 280 characters. Twitter, a nearly limitless resource, is used for text categorization. Multiple features are offered for tweets on Twitter [16].

2.4 Text Cleaning and Data Preprocessing

Data from Twitter was gathered by researchers during the proposed approach stage. Data preparation steps include data cleaning, case folding, tokenization, stop words

removal, and stemming after data collection. Next, carry out sentiment analysis once more, applying a random forest technique [17].

1. Data cleansing is the process of removing unimportant tweet data to create relevant data.
2. Case folding is the transformation of words into similar-looking forms, such as lowercase or uppercase letters.
3. Tokenization is the division of phrases into smaller units known as tokens. Words, phrases, or other meaningful elements can be formed into tokens.
4. Stopwords are words that are widely used and frequently found in sentences but are removed. The stopwords list claims that Indonesian stop words like dan (and), atau (or), etc., can be found in Twitter tweets.
5. By removing prefixes and suffixes, a word's base is obtained through stemming [18]. Use the random forest technique to continue performing sentiment analysis after that.

2.5 Text Mining

Text mining is the quickly evolving practice of identifying and extracting information from large unstructured textual resources. Text mining can utilize unstructured data sources [19]. Text mining involves three steps.

1. Information acquisition

    a. Collect, select, and filter database documents (Twitter, Facebook, or another database)

2. Extraction of information

    a. Analysis of the language that is incomplete, superficial, and in-depth
    b. Identify important elements and data-related objects

3. Data mining

    a. Integrate and connect facts
    b. Learn new information and facts

2.6 Natural Language Processing (NLP)

For example, language processing for sentiment analysis uses using computing technology known as natural language processing (NLP), text can be processed naturally at one or more levels of linguistic analysis in order to reach human-like comprehension and sentence-essence abilities. Functions that can be used in a variety of applications are provided by natural language processing [20]. Bhuvneshwar Kumar et al. [21] focus on Natural Language Processing (NLP) issues to distinguish between positive and negative customer reviews for products on the online market. Amazon.com, Rediff.com, and Flipkart.com were used to gather the information for this.

2.7 Feature Extraction

By choosing the ideal characteristics, machine learning may achieve its highest level of accuracy. In each classification task, feature selection is an essential step. The set of characteristics in text a small group of words called classification can be used to distinguish between several classes [22]. The chosen terms need to offer pertinent information that can be applied to classification. Therefore, it's crucial to consider several methods for putting the content in a format that can be used to get the information needed. In this study, term-based, sentiment-based, and GoFood-related keyword features are weighted as useful features.

Features Based on Sentiment. Contextual polarity is used in sentiment analysis to extract a text's sentiment. It is frequently used to categorize online reviews of various things, like the mood of movies. We used TextBlob to assign in this investigation, add a sentiment to each tweet library [23]. To analyse textual data, TextBlob is a Python library. A sentiment value—positive, neutral, or negative—is assigned to a tweet based on its polarity score.

The only IDF and TF-IDF word weighting algorithms ponder whether a term or phrase is relevant over the entire corpus as opposed to just one particular text. In [12], it was demonstrated that TF-IDF is more accurate than IDF. Where NJ is the number of documents where the term j appears, and n is the total number of documents.

1. Term Frequency-Inverse Document Frequency (TF-IDF)
   TF-IDF is a widely used weighting method whose performance is still on par with cutting-edge methods despite its widespread use. The phrase "weighting" refers to factors, which include documents. The primary preprocessing step necessary to index the texts is taken into consideration while choosing the feature for the feature selection procedure [24]. In applications for retrieving information connected to supervised learning, such as text categorization and filtering, the features are weighted using supervised term weighting (STW) [25]. While [26] developed the term frequency-inverse document frequency (TF-IDF) strategy, which is regarded as the conventional method for weighting the keywords in feature extraction. To generate a composite weight for each term in each document, the terms "term frequency" and "inverse document frequency" are combined [27].

$$tf - idf = tfxidf \tag{1}$$

2. Data Processing
   In the data processing stage, the researcher first performs feature extraction using TF-IDF by reducing the maximum number of features by 10000. Then separate the training data from the test data. Training data is 80%, and testing data is 20%. The random forest algorithm is then applied to the training data to perform the classification. After that, the test data is subjected to the categorization outcomes from the training data. Performance evaluation is done after the data has been processed.
3. Sentiment Analysis
   Sentiment analysis is a type of data mining that examines people's opinions through text analysis, computational linguistics, and natural language processing [28]. In recent years, sentiment analysis in social media has been a useful technique for

gathering an overview of the general public's opinion on a certain topic. Numerous studies have been conducted on sentiment, ranging from hotel evaluations [29] to movie reviews [30], with the goal of eliciting thoughts on subjects, trends, etc.

4. Random Forest
The RF classification algorithm was first presented by Breiman in 2001 [17]. The RF is a collection of unpruned classification and regression trees [31]. To build an RF, the Bootstrap sample is extracted. Afterward, the Bootstrap sample does recursive partitioning. At each node, the q predictors are chosen at random from the p predictors. After finishing the recursive partitioning, a tree is created. Until a forest is created, the aforementioned stages are repeated. When all trees cast their votes in the majority, a classification based on forests is created [32].
5. Metric Evaluation
We analyze the metrics to examine how the classification model we have developed performs and is represented by several metrics such as accuracy, precision, recall, and F-measure after we train the data and load the random forest algorithm-based classification model. If the dataset is not balanced, the accuracy metric is flawed. When dealing with unbalanced datasets, precision and recall are preferable metrics.

Calculating the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) the chosen metrics. TP stands for the %of incidents that are appropriately labeled as positive, TN for instances that were accurately identified as negative, and FP for instances that were mistakenly labeled as positive. The percentage of cases that are wrongly labeled as negative is known as FN. Accuracy. It is a metric used to assess how well a prediction model performs. It measures how often labels are correctly classified. Equation 2 is used to compute it

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

Precision. It gauges true positive predictions. Equation 3 is used to calculate a model's precision:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

Recall. A sensitivity measure is this one. It is developed to evaluate how well a model predicts favourable labels. It is established by using Eq.:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

F-Measure. In this statistic, recall and precision are also taken into account. With values ranging from 0 (worst) to 1, it can be conceptualized as the weighted average of the recall and precision measurements (best). Equation 5 is used to calculate the F-measure.

$$\text{F-Measure} = 2 \times \frac{Precision \times \text{Re}call}{Precision \times \text{Re}call} \tag{5}$$

Table 1. Examples of classified Tweets.

Tweets	Sentiment	Polarity
wkwkwkwk really remember when fasting, people suggest good GoFood	Neutral	0
let's go for the rest of the GoRide GoCar GoFood vouchers, GoFood snack zone, money zone	Neutral	0
GoFood Yogyakarta, Taiwan, the shipping is still	Neutral	0
it was fixed after I got to college, where I had to learn to cook seriously in junior high. At least I've made taken seasoning when I'm boarding GoFood, then	Negative	−0,3
seliwer mulu mad coco ig, I'm going to go to GoFood, the crowd is bust, fortunately, gojek brother is waiting	Negative	−0,625
Huhu, sorry, GoFood has sold out, yes, it's only open in the evening	Negative	−0,25
gilee GoFood tokopedia brani using bts is cool, what ads do you use bts already tw lahh sultan babang gojek application very lucky gess jimin when he pouts so cute GoFood tokopedia bts advertising bts	Positive	0,33
recommended menu GoFood accompanying watching bts dimsum accompanying watching make it right	Positive	0,28
recommended menu to GoFood accompanying watching bts like army already a really favorite mixue if you watch it at noon, it's really mandatory, goFood is ready, just one set, one set, come drink, watch it's really fresh	Positive	0,5

3 Result and Discussion

This section summarizes the findings of our experiment and evaluates the effectiveness of the random forest classification technique. 5753 tweets are used as the test dataset in this study, and 23010 tweets are used as the training dataset. While the random forest is categorized as machine learning, training and testing data are split 80:20 to reduce confusion. If 0> is positive, 0 is negative, and 0 is neutral, the data is labeled according to polarity analysis. As shown in Table 1, certain instances of these tweets have been classified with various emotions.

3.1 Metric Evaluations and Classification Report

For the data results in the table below, a total of 5753 comments/tweets generated 220 true positive comments, 92 true negative comments, and 5364 true neutral comments as shown in Table 2. This algorithm takes about 5 min to computer.

Table 2 findings allow us to assess the percentage of categorization metrics belonging to each class, with 0 indicating a negative class, 1 neutral, and 2 positives. The percentage of estimated recall, precision, and F1 score for each class is visualized in Fig. 2.

The random forest approach gave a classification accuracy of 98.6% for a dataset containing 28763 records. This dataset was divided 80 percent into the training set and

Table 2. Metrics Classification Predictive

Classification Categgories	True Positive	True Neutral	True Negative
Prediction Positive	220	33	2
Prediction Neutral	2	5364	1
Prediction Negative	3	36	92

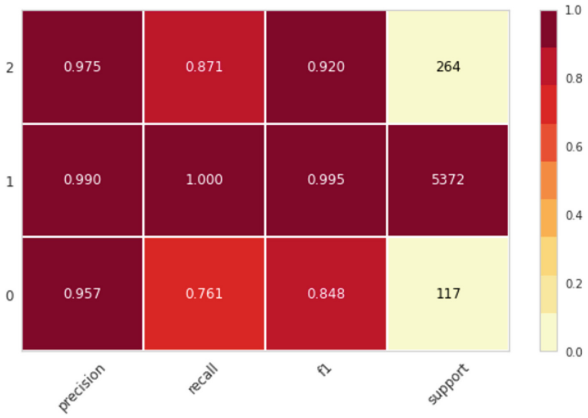

Fig. 2. Metrics Evaluations Percentage for Each Class.

Table 3. Metrics Evaluations for Each Class.

Accuracy	Precision	Recall	F-Measure
98.6%	98%	85.3%	90.6%

20 percent into the test set. Other measurements include precision, recall, and f1-score, as given in Table 3.

Figure 3 illustrates the ROC curve, which can be used to analyze the performance of the model. The ROC curve is utilized to show the relationship between sensitivity and 1-specificity. In this paper, the ROC curve is used extensively to characterize diagnostic accuracy and to find the best cut-off value for a model trained with the random forest approach.

From the perspective of business and application management, the results of the sentiment analysis indicate that the performance of the GoFood platform is relatively good, as the percentage of positive sentiment analysis results is still greater than the percentage of negative sentiment analysis results.

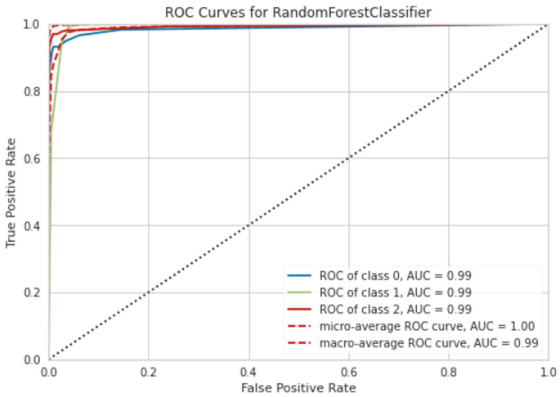

Fig. 3. ROC Curves from Model.

4 Conclusion

Management and product managers must consider sentiment the most while making decisions. This is useful for calculating, identifying, and articulating concepts regarding the product or application platform being built. In this article, we illustrate how the random forest algorithm creates classification rules efficiently and evaluate the model's performance using metrics such as accuracy, precision, recall, and the ROC curve. Using this method requires a considerable amount of time to calculate, approximately 5 min.

References

1. A. Nurwidyantoro and E. Winarko.: Event detection in social media: A survey. In: Proc. - Int. Conf. ICT Smart Soc. 2013 Think Ecosyst. Act Converg, pp. 307–311 (2013).
2. A. M. Ramadhani and H. S. Goo.: Twitter sentiment analysis using deep learning methods," 2017 7th International Annual Engineering Seminar (InAES), pp. 1–4 (2017).
3. L. Yang, Y. Li, J. Wang, and R. S. Sherratt.: Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning. IEEE 8, 23522–23530 (2020).
4. P. Ji, H. Y. Zhang, and J. Q. Wang.: A Fuzzy Decision Support Model with Sentiment Analysis for Items Comparison in e-Commerce: The Case Study of http://PConline.com. IEEE Trans. Syst. Man, Cybern. Syst. 49(10), 1993–2004 (2019).
5. S. J. Lewis.: Thumbs up. Am. Journal Orthod. Oral Surg. 31(9), 481–482 (1945).
6. D. Zeng, Y. Dai, F. Li, J. Wang, and A. K. Sangaiah.: Aspect based sentiment analysis by a linguistically regularized CNN with gated mechanism. Journal Intell. Fuzzy Syst. 36(5), 3971–3980 (2019).
7. P. Karthika, R. Murugeswari, and R. Manoranjithem.: Sentiment Analysis of Social Media Network Using Random Forest Algorithm. In: IEEE Int. Conf. Intell. Tech. Control. Optim. Signal Process. INCOS 2019, pp. 1–5 (2019).
8. Y. Al Amrani, M. Lazaar, and K. E. El Kadirp.: Random forest and support vector machine based hybrid approach to sentiment analysis. Journal Procedia Comput. Sci. 127, 511–520 (2018).

9. A. M. Florio, D. Feillet, and R. F. Hartl.: The delivery problem: optimizing hit rates in e-commerce deliveries. Journal Transp. Res. Part B Methodol. 117, 455–472 (2018).
10. C. I. Hsu, S. F. Hung, and H. C. Lin.: Vehicle routing problem with time-windows for perishable food delivery. Journal Food Eng. 80(2), 465–475 (2007).
11. A. K. Keshri, B. K. Mishra, and B. P. Rukhaiyar.: When rumors create chaos in e-commerce. Journal Chaos, Solitons and Fractals 131(xxxx), 109497 (2020).
12. X. Cui et al..: Chinese social media analysis for disease surveillance. Journal Pers. Ubiquitous Comput. 19(7), 1125–1132 (2015)
13. X. Gao, X. Shi, H. Guo, and Y. Liu.: To buy or not buy food online: The impact of the COVID-19 epidemic on the adoption of e-commerce in China. Journal PLoS One 15(8), 1–14 (2020).
14. P. Perner.: Machine Learning and Data Mining in Pattern Recognition. Springer International Publishing (2018).
15. M. M. Mostafa.: More than words: Social networks' text mining for consumer brand sentiments. Journal Expert Syst. Appl. 40(10), 4241–4251 (2013).
16. Go, Huang, & Bhayani.: Twitter Sentiment Analysis (final project results). Journal Inform (2009).
17. Z. Jin, J. Shang, Q. Zhu, C. Ling, W. Xie, and B. Qiang.: RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis. Lect. Notes Comput. Sci. (including Subsea. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12343 LNCS, pp. 503–515, (2020)
18. S. Prayoginingsih and R. P. Kusumawardani.: Klasifikasi Data Twitter Pelanggan Berdasarkan Kategori myTelkomsel Menggunakan Metode Support Vector Machine (SVM). Journal Sisfo. 7(2) (2018).
19. L. Deng and D. Yu.: Deep learning: Methods and applications. Journal Trends Signal Process 7(3–4), 197–387 (2013).
20. V. A. Fitri, R. Andreswari, and M. A. Hasibuan.: Sentiment analysis of social media Twitter with case of Anti-LGBT campaign in Indonesia using Naïve Bayes, decision tree, and random forest algorithm. Journal Procedia Comput. Sci. 161, 765–772 (2019).
21. C. W. K. Leung.: Sentiment Analysis of Product Reviews. Journal Encycl. Data Warehouse. Mining, Second Ed. 3(5), 1794–1799 (2011).
22. T. Joachims.: A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In: Proc. ICML97 (1997).
23. S. Loria, P. Keen, M. Hannibal, R. Yankovsky, D. Karesh, et al.: TextBlob: simplified text processing. Secondary TextBlob: Simplified Text Processing 2014 (2019).
24. M. Ramya and J. A. Pinkas.: Different Type of Feature Selection for Text Classification. International Journal Computer Trends Technology 10(2), 102–107 (2014).
25. F. Debole and F. Sebastiani.: Supervised Term Weighting for Automated Text Categorization. 81–97 (2004).
26. J. Chen, C. Chen, and Y. Liang.: Optimized TF-IDF Algorithm with the Adaptive Weight of Position of Word. 133, 114–117 (2016).
27. J. T. Medler.: The types of Flatidae (Homoptera) in the Stockholm Museum described by Stål, Melichar, Jacobi, and Walker. Journal Insect Syst. Evol. 17(3), 323–337 (1986).
28. V. Bonta, N. Kumaresh, and N. Janardhan.: A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis. Asian Journal Computer Science Technology 8(S2), 1–6 (2019).
29. R. A. Priyantina and R. Sarno.: Sentiment analysis of hotel reviews using Latent Dirichlet Allocation, semantic similarity, and LSTM. International Journal Intell. Eng. Syst. 12(4), 142–155 (2019).

30. Suhariyanto, A. Firmanto, and R. Sarno.: Prediction of Movie Sentiment based on Reviews and Score on Rotten Tomatoes using SentiWordnet. In: Proc. - 2018 Int. Semin. Appl. Technol. Inf. Commun. Creat. Technol. Hum. Life, iSemantic 2018, pp. 202–206 (2018).
31. A. Smith.: Image segmentation scale parameter optimization and land cover classification using the random forest algorithm. Journal Spat. Sci. 55(1), 69–79 (2010).
32. J. Ibrahim, M.-H. Chen, and D. Sinha.: Springer Series in Statistics, 27(2) (2009).