Machine Learning Models for Analysis of User Credibility Index in the E-Marketplaces

Koray Gunel
Department of Management Information Systems
Faculty of Economics and Administrative Sciences
Pamukkale University
Denizli, Turkiye
koraygunel@pau.edu.tr

Selcuk Burak Hasiloglu
Department of Management Information Systems
Faculty of Economics and Administrative Sciences
Pamukkale University
Denizli, Turkiye
selcukburak@hasiloglu.com

***Abstract*—The paper is aimed at identifying the most suitable machine learning model for testing user credibility index in the E-marketplaces. The findings revealed that Gradient Boosting and Random Forest algorithms are the most suitable models for this study.**

***Keywords*— machine learning models, credibility index, web scraping, data analytics, sentiment analysis, e-marketplaces**

I. Introduction

Digital shopping platforms like e-marketplaces contain more uncertainties compared to traditional marketplaces. Hence, consumer trust perception holds a significant place in digital marketing. Perceived uncertainties for consumers also generate a sense of risk. There are numerous studies in the literature of digital marketing that use the concepts of trust and risk together or separately [1]. Some studies evaluate the relationships between these two concepts independently or interconnectedly [2].

Consumers resort to various factors to reduce uncertainty and risk. Among these factors, the opinions and reviews of other consumers are crucial. Reviews serve as guiding competitive tools for other consumers regarding the product or seller [3], [4]. Unfortunately, there are also fake reviews, commonly known as spam comments [5], [6], [7]. With the increase in fake reviews, trust in reviews becomes an important factor for consumers. Regardless of how many positive reviews a seller's product has on an e-marketplace, if they are not credible, they hold no meaning. In this study, machine learning was used to produce an index value for the credibility of reviews.

Sentiment analysis is used to examine user comments.

Sentiment analysis captures and analyzes people's opinions and attitudes toward a product, service, topic, or issue. Opinions and attitudes include judgments, beliefs, feelings, evaluations, emotional/emotional states, desires, etc. Also known as opinion mining or subjective analysis [8], [9], [10], [11].

Machine learning techniques for sentiment classification are interesting because they can model many features while capturing context [12], adapt more easily to changing inputs, and measure the likelihood of classification uncertainty. The most popular are supervised methods trained on manually classified examples. The most common approach here uses lowercase letters as features when describing training and test examples. Thought mining involves using two or more words to express a specific emotion and is important for accurately identifying negative emotions because it reverses the polarity.

In our study, 6 machine learning models were compared.

Gradient boosting is a potent machine learning algorithm used to construct predictive models. The fundamental concept is to minimize the model's loss function by adding new weak learners (decision trees) to compensate for the deficiencies of existing ones. Each iteration concentrates on samples that were previously deemed difficult and misclassified, thereby improving the model's accuracy. [13].

Random Forest is a high-performance machine learning algorithm and strengthens the decision-making process by combining many tree structures. Each tree is trained independently using random samples of the dataset, and then these trees are combined to make an overall prediction. This method is robust to overfitting and is known for its ability to evaluate the impact of various features. Random Forest is a widely used algorithm for classification and regression problems and provides effective results on large data sets [14].

The CN2 algorithm, used to build a rule-based classifier from a dataset, is a widely used rule induction technique in the fields of data mining and machine learning. This algorithm aims to learn simple and understandable sets of rules that describe patterns in the data set. CN2 is especially effective on small and medium-sized datasets and is widely preferred in data mining projects to improve classification accuracy. The algorithm determines the best rule set by going through the data set and makes classification based on this rule set, thus becoming a useful tool in data analysis [15].

k-NN (k-Nearest Neighbor) algorithm is a classification and regression method used in the field of machine learning and data mining. This algorithm uses the average around neighboring data points to classify or predict a new data point. The user-specified k value represents the number of data points to be used as neighbors. Although the distance metric used is usually the Euclidean distance, different distance metrics can also be used. The k-NN algorithm is especially effective in small and medium-sized data sets with its simple and understandable structure. However, for large data sets, the computational cost may increase and may be sensitive to noise in the data set [16].

The Naive Bayes algorithm is a statistical algorithm used for classification problems in machine learning. This algorithm is based on the basic principles of Bayes' theorem and is especially widely used in fields such as text classification. The basic assumption of Naive Bayes is the assumption of independence between attributes, that is, the value of one attribute does not depend on the values of other

This study was funded by the Scientific and Technological Research Council of Turkey (TUBITAK) ARDEB 1001 Grant No 122K017. We thank TÜBİTAK for its support.

attributes. That's why it's called "naive". This algorithm performs probabilistic classification of new samples by examining the distribution of features in the data set. This makes it possible to predict to which class a sample belongs. The Naive Bayes algorithm is generally known for its simple structure and good performance, but it is important to note that the independence assumption is not always valid in real data sets [17].

Logistic regression is a modeling technique used to solve classification problems in statistics and machine learning. This algorithm is used when the dependent variable is a categorical variable and focuses on predicting the probability of an event as output. Logistic regression attempts to model the relationship between independent and dependent variables using a linear regression model but limits the output (a value between 0 and 1) so it can be interpreted as a probability. The algorithm performs classification by multiplying input data by weights and comparing them to a threshold (cutoff). During the training phase, maximum likelihood estimation methods are usually used to estimate parameters. Logistic regression is widely used in binary classification problems and is based on linear probability modeling [18].

Support vector machine (SVM) is a powerful algorithm used for classification and regression problems in machine learning. SVM is considered a learning model that is particularly effective on nonlinear data sets. Essentially, it tries to find the best dividing line (or hyperplane) to divide the data points into different classes. The purpose of this dividing line is to separate the data points as much as possible while maximizing the separation between classes. SVM uses support vectors to accomplish this; that is, it identifies the data points that contain the points closest to the split line and uses these points to determine the best location of the split line. For nonlinear data sets, SVM uses the kernel method to transform the data into a high-dimensional space to make them linearly separable, thereby being able to handle more complex data structures. SVM has a wide range of applications and is known for its high accuracy and generalization capabilities [19].

II. Methodology

In this study, the aim was to train and test the credibility of classified comments using machine learning. Two consumers were asked to score the credibility level by examining product pages and comments on Amazon and Trendyol e-marketplaces for one month to create the dataset. In this way, we obtained a total of 5500 comment data and then cleaned the problematic data. We utilized Scott's pi analysis to measure the consistency between the cleaned data, and any inconsistent data were re-evaluated within the scope of common agreement.

We proceeded to determine the most suitable machine learning model for testing the generated credibility index. We conducted web scraping by using the link of each comment. The 10 parameters collected through web scraping are as follows:

1. Star: User comment star ratings ranging from 1 to 5.
2. Picture: Binary values (0 or 1) indicating whether users added photos to their comments (0 for not added, 1 for added).
3. Like: Number of likes received on user comments.
4. MainStar: Arithmetic mean of all users' star ratings for the product.
5. Rating: Total number of star ratings given for the product.
6. Total Review: Number of reviews containing comments for the product.
7. Neg. Ratio: Ratio of negative comments to all comments for the product.
8. Polarity: Compound values calculated based on sentiment analysis of comments.
9. Sentiment: Values indicating the emotional load of comments (1 for Positive, 0 for Neutral, -1 for Negative).
10. Credibility Index: Credibility score of the comment for the product.

After the web scraping process, the 9 parameters we used and credibility scores were added to the dataset for analysis. Among these parameters, polarity and sentiment variables are included. The results of sentiment analysis of comment data revealed the distribution of emotional evaluations of comments. By examining polarity values, it was determined whether comments tended to be positive, negative, or neutral. The Vader Lexicon dictionary was used for sentiment analysis, which calculates emotional charges of words as positive, negative, or neutral.

We utilized the Orange Data Mining software in the analysis process, which is based on Python and developed in collaboration with the University of Ljubljana in Slovenia and the open-source community for research on machine learning and data mining [20]. Cross-validation was used in the machine learning data sampling process, where the data was split into 5 folds for cross-validation. The algorithm was tested by holding out examples from one fold at a time, while the model was induced from the other folds. The held-out fold examples were then classified, and this process was repeated for all folds.

Table 1 presents the comparison results of the models based on the Area under ROC method. As seen from the table, the row where the Gradient Boosting model is located has the value closest to 1 in comparison with all other models.

TABLE I. COMPARISON OF THE MODELS ACCORDING TO THE AREA UNDER ROC METHOD

	Gradient Boosting	Random Forest	CN2 rule inducer	kNN	Naive Bayes	Logistic Regression	SVM
Gradient Boosting		0.814	1.000	0.993	0.999	0.994	1.000
Random Forest	0.186		0.737	0.974	0.982	0.996	1.000
CN2 rule inducer	0.000	0.263		0.918	0.997	0.988	1.000
kNN	0.007	0.026	0.082		0.871	0.971	1.000
Naive Bayes	0.001	0.018	0.003	0.129		0.946	0.996
Logistic Regression	0.006	0.004	0.012	0.029	0.054		0.822
SVM	0.000	0.000	0.000	0.000	0.004	0.178	

In Table 2, the comparison results of the models based on the classification accuracy method are presented. As seen from the table, the row where the Random Forest model is located has the value closest to 1 in comparison with all other models.

Classification accuracy is a metric used in machine learning and statistics to measure the accuracy of a model. This indicator represents the proportion of correctly predicted samples to the total samples. It is usually expressed as a percentage, with higher accuracy values meaning better model performance. However, classification accuracy alone is insufficient and can be misleading in class-imbalanced datasets. Therefore, various performance metrics and evaluation methods (such as sensitivity, specificity, F1 score, etc.) should be used to more comprehensively evaluate the performance of the model.

TABLE II. COMPARISON OF THE MODELS ACCORDING TO THE CLASSIFICATION ACCURACY METHOD

	Gradient Boosting	**Random Forest**	**CN2 rule inducer**	**kNN**	**Naive Bayes**	**Logistic Regression**	**SVM**
Gradient Boosting		0.338	0.992	0.366	0.999	0.998	0.992
Random Forest	0.662		0.988	0.558	0.999	0.999	0.995
CN2 rule inducer	0.008	0.012		0.020	0.996	0.955	0.973
kNN	0.634	0.442	0.980		1.000	0.999	0.995
Naive Bayes	0.001	0.001	0.004	0.000		0.001	0.959
Logistic Regression	0.002	0.001	0.045	0.001	0.999		0.959
SVM	0.008	0.005	0.027	0.005	0.439	0.041	

Table 3 displays accuracy values for 7 different machine learning models. According to the findings, Gradient Boosting and Random Forest models have the highest accuracy.

TABLE III. ACCURACY VALUES OF THE MODELS

	AUC	**CA**	**F1**	**Precision**	**Recall**	**LogLoss**
Gradient Boosting	0.962	0.942	0.939	0.938	0.942	0.135
Random Forest	0.950	0.944	0.944	0.944	0.944	0.305
CN2 rule inducer	0.941	0.927	0.927	0.926	0.927	0.172
kNN	0.922	0.944	0.942	0.941	0.944	0.581
Naive Bayes	0.897	0.878	0.889	0.906	0.878	0.338
Logistic Regression	0.854	0.909	0.888	0.887	0.909	0.239
SVM	0.796	0.875	0.882	0.890	0.875	0.287

In Table 4, performance values based on the Proportion of Actual for Gradient Boosting and Random Forest machine learning models are provided. According to the table, the proportion of correctly predicting positive credibility index is high in the Gradient Boosting model, while the proportion of correctly predicting negative credibility index is high in the Random Forest model. However, the predictive power of the Random Forest model (for negative credibility index) is more pronounced.

TABLE IV. PERFORMANCE TABLE OF GRADIENT BOOSTING AND RANDOM FOREST MODELS

CREDIBILITY INDEX Proportion of actual			Predicted			
			Positive		Negative	
			Gradient Boosting	Random Forest	Gradient Boosting	Random Forest
Actual	Positive	Gradient Boosting	98.0%		2.0%	
		Random Forest		97.0%		3.0%
	Negative	Gradient Boosting	41.2%		58.0%	
		Random Forest		30.3%		69.7%

III. CONCLUSION

In this study, machine learning models were compared to generate an index value for the credibility of comments. According to the created model, variables such as the star rating of the product, the total number of comments received by the product, the ratio of negative comments to the product, the star rating of each comment, the number of likes received on the comment, the presence of photos in the comment, and the sentiment value of the comment contribute to forming the credibility index of the comment.

The results of comparing machine learning models with each other were evaluated according to the "area under ROC" method. Subsequently, the results of comparing models with each other were examined according to the classification accuracy method. Accuracy values were calculated for 7 different machine learning models, and performance values were examined based on the proportion of actual. It was found that the actual positive prediction rate is high in the Gradient Boosting model, while the actual negative credibility index prediction rate is high in the Random Forest model. However, the prediction power (negative credibility index) of the Random Forest model is more significant. Based on the findings, Gradient Boosting and Random Forest algorithms were identified as the most suitable models for our study.

REFERENCES

[1] S.Gurkaynak-Gurbuzer, and S.B. Hasiloglu, “Dijital Pazarlama Güven Ölçeğinin Geliştirilmesi”, *İnternet Uygulamaları ve Yönetimi Dergisi*, vol.15, no. 1, pp.1-15, 2024

[2] D. Gefen, V. S. Rao, and N. Tractinsky, “The Conceptualization of Trust, Risk and Their Relationship in Electronic Commerce: The Need for Clarifications.,” in *HICSS*, 2003, p. 192. Accessed: Mar. 11, 2024. [Online]. Available: http://www.ise.bgu.ac.il/faculty/noam/Papers/03_gefen_rao_tractinsky.pdf

[3] R. A. Westbrook, “Product/Consumption-Based Affective Responses and Postpurchase Processes,” *Journal of Marketing Research*, vol. 24, no. 3, pp. 258–270, Aug. 1987, doi: 10.1177/002224378702400302.

[4] M. R. Jalilvand and N. Samiei, “The impact of electronic word of mouth on a tourism destination choice: Testing the theory of planned behavior (TPB),” *Internet research*, vol. 22, no. 5, pp. 591–612, 2012.

[5] E. D. Wahyuni and A. Djunaidy, “Fake review detection from a product review using modified method of iterative computation framework,” in *MATEC web of conferences*, EDP Sciences, 2016, p. 03003. Accessed: Mar. 12, 2024. [Online]. Available: https://www.matec-conferences.org/articles/matecconf/abs/2016/21/matecconf_bisstech2016_03003/matecconf_bisstech2016_03003.html

[6] A. Heydari, M. ali Tavakoli, N. Salim, and Z. Heydari, “Detection of review spam: A survey,” *Expert Systems with Applications*, vol. 42, no. 7, pp. 3634–3642, 2015.

[7] E. Kauffmann, J. Peral, D. Gil, A. Ferrández, R. Sellers, and H. Mora, “A framework for big data analytics in commercial social networks: A case study on sentiment analysis and fake review detection for marketing decision-making,” *Industrial Marketing Management*, vol. 90, pp. 523–537, Oct. 2020, doi: 10.1016/j.indmarman.2019.08.003.

[8] B. Liu, M. Hu, and J. Cheng, “Opinion observer: analyzing and comparing opinions on the Web,” in *Proceedings of the 14th international conference on World Wide Web - WWW ’05*, Chiba, Japan: ACM Press, 2005, p. 342. doi: 10.1145/1060745.1060797.

[9] J. Wiebe, R. Bruce, and T. P. O’Hara, “Development and use of a gold-standard data set for subjectivity classifications,” in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, 1999, pp. 246–253. Accessed: Mar. 19, 2024. [Online]. Available: https://aclanthology.org/P99-1032.pdf

[10] M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger, “Pulse: Mining Customer Opinions from Free Text,” in *Advances in Intelligent Data Analysis VI*, vol. 3646, A. F. Famili, J. N. Kok, J. M. Peña, A. Siebes, and A. Feelders, Eds., in Lecture Notes in Computer Science, vol. 3646. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 121–132. doi: 10.1007/11552253_12.

[11] M. Karamibekr and A. A. Ghorbani, “Verb oriented sentiment classification,” in *2012 IEEE/WIC/ACM international conferences on web intelligence and intelligent agent technology*, IEEE, 2012, pp. 327–331. Accessed: Mar. 19, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6511904/

[12] L. Polanyi and A. Zaenen, “Contextual Valence Shifters,” in *Computing Attitude and Affect in Text: Theory and Applications*, vol. 20, J. G. Shanahan, Y. Qu, and J. Wiebe, Eds., in The Information Retrieval Series, vol. 20. , Berlin/Heidelberg: Springer-Verlag, 2006, pp. 1–10. doi: 10.1007/1-4020-4102-0_1.

[13] Z. Yan and H. Wen, “Comparative Study of Electricity-Theft Detection Based on Gradient Boosting Machine,” in *2021 IEEE INTERNATIONAL INSTRUMENTATION AND MEASUREMENT TECHNOLOGY CONFERENCE (I2MTC 2021)*, in IEEE Instrumentation and Measurement Technology Conference. New York: IEEE, 2021. doi: 10.1109/I2MTC50364.2021.9460035.

[14] J. Han, Y. Liu, and X. Sun, “A Scalable Random Forest Algorithm Based on Map Reduce,” in *PROCEEDINGS OF 2013 IEEE 4TH INTERNATIONAL CONFERENCE ON SOFTWARE ENGINEERING AND SERVICE SCIENCE (ICSESS)*, M. S. P. Babu and L. Wenzheng, Eds., in International Conference on Software Engineering and Service Science. New York: IEEE, 2012, pp. 849–852. Accessed: Mar. 19, 2024. [Online]. Available: https://www.webofscience.com/wos/woscc/summary/375da509-ca46-4da7-9e82-957dd574ca2f-d690256a/relevance/1

[15] N. Kumar and U. Kumar, “Comparative analysis of CN2 rule induction with other classification algorithms for network security,” *Multimed. Tools Appl.*, vol. 81, no. 26, pp. 37119–37135, Nov. 2022, doi: 10.1007/s11042-022-13542-3.

[16] B. Lazzerini and F. Marcelloni, “K-NN algorithm based on neural similarity,” in *2002 IEEE INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE SYSTEMS, PROCEEDINGS*, V. G. Zakharevich and V. M. Kureichik, Eds., Los Alamitos: IEEE Computer Soc, 2002, pp. 67–70. doi: 10.1109/ICAIS.2002.1048054.

[17] J. Han, M. Kamber, and J. Pei, “Data mining concepts and techniques third edition,” *University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University*, 2012, Accessed: Mar. 19, 2024. [Online]. Available: https://www.academia.edu/download/43034828/Data_Mining_Concepts_And_Techniques_3rd_Edition.pdf

[18] Y. Fan *et al.*, “Privacy preserving based logistic regression on big data,” *J. Netw. Comput. Appl.*, vol. 171, p. 102769, Dec. 2020, doi: 10.1016/j.jnca.2020.102769.

[19] H. Tamura and K. Tanno, “Midpoint-validation method for Support Vector Machine classification,” *IEICE Trans. Inf. Syst.*, vol. E91D, no. 7, pp. 2095–2098, Jul. 2008, doi: 10.1093/ietisy/e91-d.7.2095.

[20] J. Demšar *et al.*, “Orange: data mining toolbox in Python,” *the Journal of machine Learning research*, vol. 14, no. 1, pp. 2349–2353, 2013.