Nelson B. Wekesa, 11:6
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

Embedded Dynamic Feature Selection Methods in the Detection of Fake Reviews in Social Media.

Nelson B. Wekesa, Dr. Kennedy Ogada, Dr. Tobias Mwalili
Department of Computing,
Jomo Kenyatta University of Agriculture and Technology, Kenya

Abstract - Technological advancement has led to the growth of internet users, and hence social media usage. Social media plays a pivotal role in society today as compared to in the past. However, there are chances of deceptive social media reviews with massive usage; Hence, it is imperative that there is a need for improved authenticity and robust fake social media reviews detection tools. Embedded feature selection methods seem to be more effective in detecting fake social media reviews owing to the massive content generated daily. This study aims to assess the use of feature selection methods to detect fake reviews in social media data. LASSO, RIDGE, and Random Forest classification methods are experimented. Thstudy findings are that the three feature selection methods perform the same. Classification models using methods experimented had an accuracy of ninety percent (90%). Classification models without feature selection (all features present) recorded the lowest accuracy of eighty-nine percent (89%). The classification model using LASSO outperforms RIDGE irrespective of the penalization technique used. Feature selection is critical to improve the classification model prediction accuracy and precision, reduce training time, and test time.

Keywords - Fake; Features; LASSO; RIDGE; Random Forest; Social Media Reviews.

I. INTRODUCTION

A feature refers to a stand-alone characteristic that is measurable about an instance under observation [1]. Machine learning algorithms often use several features presented in a dataset for classification tasks. However, feature selection is significant as it enables to get rid of irrelevant features, enhancing the learning methods performance akin to the interpretability of the results therein. Primarily, feature selection entails the choosing of a relevant subset from given features that are relevant. Feature selection is a process of eliminating unwanted or redundant features conducted in both high dimension and low dimension datasets. Redundant features undermine the performance of the model as well as affect the search for valuable knowledge and classification accuracy. [2] states that redundant features in a dataset increases computation burden in high dimension dataset. There are three classifications of supervised feature selection methods. 1. Filter, 2. Wrapper and 3. Embedded methods. According to [3], filter method is a category of feature selection that is independent of any machine learning algorithms. Features selected do not provide feedback on the improvement of the algorithms. In relation to [4], filter methods measure the relevance of features by their correlation with the dependent variable. Features with meaningful relationships are the only ones included in a classification model. The Wrapper method on the other hand, is a category of feature selection method that is dependent on machine learning algorithm used. It provides feedback about the algorithm and option for improving the algorithm to optimize the outcome. [4] argues out that, wrapper method of feature selection, evaluates usefulness of a subset of features while training a model. Consistent with [3] and [4], embedded feature selection is a category of feature selection method that integrates both filter and wrapper methods' qualities. The objective of embedded feature selection method is to optimize the performance of both filter and wrapper methods. For example, computational time of wrapper feature selection method and taking into consideration learning algorithm characteristics in the filter feature selection method. Two common ways of implementing embedded feature selection methods are LASSO and RIDGE.

In relation to works done by [5], LASSO and RIDGE regression are the critical embedded methods used text classification because they have inbuilt penalization functions to lessen overfitting. Feature selection and extraction can be conducted either in isolation or in combination to enhance the estimation performance and data visualization and make the knowledge learned more comprehensible. The features therein can be relevant, irrelevant, or redundant. Subsets with the least dimensions are those that contribute to accuracy in the learning and thus considered best during the feature selection. Unfortunately, feature extraction is disadvantageous in that original features' linear combinations usually are uninterpretable, and the information detailing the contribution of each feature is often lost. With technological advancements, massive volumes of data get generated every day from diverse channels such as social media, among other internet platforms. Unfortunately, there are chances that some of the massive information published is fake. According to [6], fake news can be published in articles, social media, and journals, while fake reviews are primarily published for organizations on social media and websites. Fake reviews often entice people into believing in market information about institutions, thereby causing social, political, or economic influence. More people use the internet today due to the increased usage of mobile phones to access the internet remotely. As a result, the news industry is seeking robust means of detecting fake news or reviews. Natural Language Processing (NLP) has been used in the past, replacing human fact checking for both reviews and news releases. However, in relation to works done by [7], using NLP with machine learning has been quite problematic in terms of accuracy. This thus makes it possible to manually fact checking fake reviews using deep learning approach. To improve accuracy differentiating the fake/deceptive reviews from the authentic/truthful reviews, different deep learning approaches can be experimented. This study demonstrates how deep learning embedded methods of feature selection and extraction in LASSO, RIDGE, and Random Forest could improve accuracy in distinguishing between true and false reviews in social media platforms.

II. RELATED WORKS

1. Fake Reviews in the Social Media

Fake reviews in the social media channels for organizations have grown exponentially in the past decade, posing threats of trustworthiness in the marketplace. In relation to [8] among the most compelling issues in social media, marketing is the increased fake reviews about organizations in the market. Consequently, customers' buying decisions are highly influenced by company reviews on social media platforms amid the high market competition posited by globalization. However, limited literature is available exploring the methodologies and techniques of determining fake reviews. This study proposes the use of deep learning approaches to distinguish fake from truthful reviews in social media. [9] demonstrated that the prevalence of social media for personal and professional use had rendered more concerns on the authenticity of the content therein. Social media usage was prevalent during the peak of the Covid-19 crisis through 2020 and early 2021. Fake social media content comprises fake news, fake reviews, spam and even comments, and rumor, hoax, and engagement. Fake social media content often results from practices of illegal marketing to generate some commercial advantage. [10] demonstrated that 77% of consumers and 75% would increase their sending on the brands they follow on social media. The implications are that the quality of the social media content significantly influences buyer consumption behaviors. The adverse effects of fake social media reviews are that, if unattended, they lead to wrong consumer purchasing decisions and further compel them to mistrust social media-generated content [11]. In addition, social media channels would not be a safer place in the marketplace despite the increased centrality of social media marketing and sales. Therefore, the ability to detect fake reviews on social media platforms has innumerable practical value in the community. Diverse factors undermine the detection of fake social media content in contemporary society. For example, the rate at which new content gets generated on social media platforms is beyond the cognitive capacity of human beings. About 3.8 billion people were in social media globally, while 4.5 billion were using the internet. Summarily, about 60% of the world's population is on the internet [12]. The Digital 2020 global report also points out that more than 5.19 billion people use smartphones, which has increased by 2.4% from the previous year. The report further projected that internet users typically spend on average of 6 hours and 43 minutes online each day. About 53.3% access the internet using mobile phones, 44% access the internet via laptops and desktops, while 2.7% use tablets [13] and [12]. Besides, about 0.07% access the internet via online gaming tools. The implications are that the rate at which social media content is generated is relatively high, hence the possibility of more fake content being consumed. Robust methods of detecting fake content (reviews) in social media are thus inevitable. However, manual fact-checking of the massive data generated in social media rarely meets practical demand [14]. On the other hand, if sufficient labor were available to cross-examine and verify the reviews in social media, human beings would often suffer from cognitive bias. Similarly, the rate of deception from online platforms is relatively high.

2. Feature Selection

According to [15] feature extraction is an approach for extracting a collection of new features from a given set of features established in the feature selection stage. [16] describes feature selection as a critical process to simplify the model through data reduction, lessen the storage burden, enhance visualization, and render Occam's razor. The Occam's razor is an evaluation theory or principle that stipulates "plurality should not be posited without necessity". In this regard, through feature selection, Occam's razor guideline on the general rule of prudence and conservative in the data investigation support simplicity amongst multiple explanations or approaches to a model in a study. Furthermore, in relation to [17], the statistical significance of a dataset is sharpened through feature selection. Also feature selection, reduces data training time, overfitting, and overall, the accuracy of the chosen model. Besides, the curse that comes with high data dimensionality is avoided. In this regard, an embedded feature selection method has been adopted for use in this study. As stated by [1] diverse feature filtering techniques can be used comprising wrappers, filters, and the embedded method deployed in this study. Often, the feature selection method deployed largely depends on the classifier used in the study. The wrapper classifier often outperforms the others; however, it is costly for vast feature spaces, especially due to its high computational needs. Therefore, every feature collection method is tested against the classifier used, which slows down the feature selection.

The filter method tends to be less costly, requires less computation, and is hence time savvy. Unfortunately, the filter method has lower classification efficiency; hence, it is only suitable for high-dimensional data sets [18]. Nevertheless, the embedded techniques outperform the former as they integrate the benefits from both the wrapper and the filter techniques. Moreover, the embedded technique is the most recent employing subset of the features of independent tests and the assessment function's output. Besides, the filter technique is further split into two; subset search algorithm and element-weighing algorithm. The filter method is beneficial because it is comparatively cheaper than the wrapper and the embedded methods, apart from the fact that its running time is much shorter than others [19]. The filter method also has lower overfitting risks, and it espouses higher abilities of generalization. Further, the filter method is much easier to scale for high-dimensional datasets. On the other hand, the filter method does not interact with the classification model in selecting the features [20]. The filter method also ignores dependency between features as it considers each feature separately, especially in the case of the Univariate techniques. Univariate techniques are techniques used where low computational performance may be recorded, unlike other feature selection methods. The filter selection method entails choosing variables irrespective of the model. The model exhibits general features like the variable association for prediction purposes. The filter method works by suppressing variables that are of least interest in the modeling. Variables that are not suppressed are thus the only ones used in the model or regression to predict the pertinent concepts [21]. For prediction models, filter methods are not the best candidates are they estimate relevant scores. For example, filter methods assume equal sample distribution for diverse classes like ANOVA, Bayesian, and Chi-squared, which cannot be relied upon on some datasets.

The Wrapper method of feature selection interacts effectively with the classifiers during the process. It tends to be more comprehensive in feature-set space selection and considers the dependency of features. Further, it is a better generalization method than the filter method [15]. On the other hand, the wrapper method has a very high computational cost, has a longer running time, and espouses high overfitting risks compared to both embedded and filter methods. Besides, the wrapper method is not computationally feasible where there are many features. The evaluation of the wrapper method is done on the subsets for the variables therein. The classifier accuracy is the basic tool for the variable communication observation in the model [22]. The wrappers achieve higher accuracy by choosing subsets of features. As such, the robust discriminative powers enshrined in the selection process in the wrapper method promote the achievement of higher accuracy results. Besides, the wrapper method is classifier-dependent; hence, the accuracy of the results largely depends on the classifiers chosen for the modeling. Therefore, different classifiers is should best be used in the feature selection to boost the accuracy of the performance.

Lastly, the embedded method is beneficial because it is computationally less intensive, especially when compared with the wrapper method. Furthermore, the running time for the embedded method is much shorter than the Wrapper method [18]. In addition, the method interacts with the classifier model in the feature selection effectively akin to the wrapper method. There are lower overfitting risks than the wrapper method, and that the method outperforms the filter method in the error generalization where data points are increased. However, the embedded method is quite problematic in the feature selection identification for a smaller set of features. Summarily, the embedded method outperforms the wrapper and the filter feature selection methods as it combines the benefits of each of the latter. The embedded method is mostly the same as the wrapper method because it relies on learning algorithms. The method is, however, less computationally intensive than the wrapper method. Thus, the higher accuracy of the embedded method partially emerges from combining the filter's efficiency with the wrapper's accuracy. Besides, the feature selection for the embedded methods is inbuilt, which aid in the reduction of the features [1]. Popular examples of embedded methods that render higher accuracy in the modeling are the LASSO and RIDGE regressions. A comprehensive overview of the accuracy and performance of the two approaches in detecting fake reviews in social media is explored in this study.

3. Regularization

The embedded feature selection methods preferred were LASSO, RIDGE, and Random Forest. The embedded methods were iterative, thereby taking care of each iteration of the model during the data training. Besides, the embedded methods were considered to aid in effectively extracting only the features that significantly contribute the most to the model for a given set of iterations. The regularization embedded method herein penalized features given a specific coefficient threshold. Also called the penalization method, the regularization method introduced new constraints in the predictive regression algorithm, thereby lessening the complexity of the model by virtue of reduced coefficients [23]. In the study, the penalization methods chosen were LASSO and RIDGE, given their inbuilt penalization functions to counter overfitting problems. LASSO regression performs regularization, which adds penalty equivalent to the absolute value of the coefficients' magnitude. On the other hand, the RIDGE regression is primarily a model tuning method used to analyze a dataset that suffers multi-collinearity. The RIDGE method performs L2 regularization, unlike the LASSO regression. Further, the Random Forest offers general predictive performance with low overfitting and renders better interpretability of the data [24]. Where the Random Forest is used, computing the contribution of each variable to the decision is much easier. Besides, the measure of impurity used was Gini impurity and information entropy.

[25] used a random forest classification model to assess the risk factors contributing to stunting among children below five years of age in Kenya. The random forest demonstrated desirable results in feature identification to explain critical factors contributing to the stunting among the select group. Some of the risk factors identified comprised of being underweight, age, religion, age of the mother, and ethnicity [25]. This study revealed that random forest, among other feature selection embedded methods, effectively determines the factors contributing to the decision in a model. For example, the study uses LASSO and RIDGE, which have inbuilt penalization functions for reducing overfitting, visualized feature importance in the model. Their individual contribution was succinctly defined.

On the other hand, elastic Net is a regularization model that, akin to different embedded approaches, prevents model overfitting by artificially penalizing model coefficients. Elastic Net is the hybrid of both LASSO and RIDGE; hence it outperforms the individual models in prediction accuracy. Furthermore, LASSO and RIDGE regression penalties are combined in Elastic Net, making it a superior prediction tool. Notably, unlike the respective constituents of Elastic Net (LASSO and RIDGE), the latter shrinks parameters associated with pertinent correlated variables in a model to an equation or ultimately remove them in entirety. Thus, in place of LASSO and RIDGE, the Elastic Net model can be deployed where the variables denote a significantly high level of correlation. The implications are that, while in the study, LASSO and RIDGE had an accuracy level of about 0.90 on average, the Elastic Net would register about 0.98 in predicting fake reviews in social media.

In addition, RIDGE reduces overfitting by shrinking regression coefficients such that variables with the insignificant contribution to the decision are tended towards zero; however, none of them is zeroed (equal to zero) upon shrinking [2]. As a result, while RIDGE shrinks the least contributing variables, every feature is incorporated in the final model in which the final decision is made. The L2 norm only reduced the overfitting but did not eliminate it entirely, as with the elastic net model. Contrary to RIDGE, LASSO does the shrinkage of regression coefficients to zero by means of L1 regularization or penalization. RIDGE retains every variable after penalization, unlike the LASSO [25]. As such, LASSO plays a pivotal role in feature selection to boost the model precision. The findings explain why RIDGE has lower accuracy than RIDGE in this study, where LASSO accuracy was 0.90 while RIDGE recorded 0.90. The model with all features present in this study reported prediction accuracy, which was insignificantly different from RIDGE; All features present model prediction accuracy was 0.90. The implications, commensurate to this study findings, are that retention of features upon shrinking in RIDGE classification model contributes to the lower prediction of the model due to the pertinent noise.

[5] denoted that model accuracy can be improved using the LASSO penalization approach in a semi-parametric mixed method (SPMM) framework instead of the RIDGE penalty. In addition, a generalized additive model using the R Gam package helps improve model fitting. However, model fitting in GAM is often done to a specified degree of freedom. Spline fitting can be applied in a model to enhance the efficiency of LASSO and RIDGE classification models [5]. Notably, penalized spline in RIDGE and LASSO mixed methods significantly improves the model performance through reduced mean average squared distances. However, the spline penalized model in LASSO outperformed the spline penalized model in RIDGE. Generally, LASSO outperforms RIDGE in all penalization methods for all penalization boundaries less than 10% and greater than 90%.

III. METHODOLOGY

The dataset used for experimentation in this study is the Deceptive Opinion Spam Corpus [26]. Data shuffling was done prior to model training to create more representative training and testing sets. In addition, data cleaning and description are followed to enhance productivity and foster the highest quality information integration in the modeling [27]. Some of the data cleaning activities undertaken comprised removing rows with missing values, fixing errors noted in the structures, data reduction for efficient handling, and finding numbers with null values. Extensive data cleaning done was to remove special characters comprising making texts lowercase, removing texts engulfed in square brackets, removing hyperlinks, and eliminating words with numbers akin to any special characters. Count Vectorizer was used to transform the text into vector-based on their frequency or the count of each text occurrence in the entire text. Count Vectorizer is a useful tool by the Scikit Learn Python library for use where there are multiple texts in which each word within a text needs to be converted into vectors [28]. Similarly, feature extraction done was using Count Vectorizer. An n-gram (1, 2) specification was used in the model because it aided in understanding the frequency of words' effects on the polarity of the social media reviews therein (truthful or fake). Logistic regression models using the select features of LASSO, RIDGE, and Random Forest are evaluated herein.

IV. RESULTS AND DISCUSSION

1. LASSO, RIDGE, and Random Forest Performance

The ease of massive information dissemination and accessibility has made social media a significant source of information for a wide range of socioeconomic and political purposes [8]. As a result, there is a need for a robust process for detecting fake social media content, especially the social media reviews owing to their market significance. Owing to the contemporary challenge, embedded feature selection methods of LASSO, RIDGE, and the Random Forest demonstrated significant results in detecting fake reviews in social media. Table 1 below represents the feature selection classification for detecting fake reviews in social media using the LASSO. For total features 92191 and selected features 2426, the mode classification precision was on average 0.9, akin to the F1 score and recall. Similar reports were also recorded for RIDGE as denoted in table 2, however, the Random Forest classification registered significantly different precision, recall and F1-Score are denoted in table 3.

Table 1: LASSO Feature Selection Classification Report

Item	Precision	Recall	F1 Score	Support
0	0.90	0.90	0.90	240
1	0.90	0.90	0.90	240
Accuracy			0.90	480
Macro Avg	0.90	0.90	0.90	480
Weighted Avg	0.90	0.90	0.90	480

Table 2: RIDGE Feature Selection Classification Report

Item	Precision	Recall	F1 Score	Support
0	0.81	0.91	0.90	240
1	0.91	0.88	0.90	240
Accuracy			0.90	480
Macro Avg	0.90	0.90	0.90	480
Weighted Avg	0.90	0.90	0.90	480

Table 3: Random Forest Feature Selection Classification Report

Item	Precision	Recall	F1-score	Support
0	0.91	0.90	0.90	240
1	0.90	0.91	0.90	240
Accuracy			0.90	480
Macro Avg	0.90	0.90	0.90	480
Weighted Avg	0.90	0.90	0.90	480

Table 4: All Features Present Classification Report

Item	Precision	recall	F1-score	Support
0	0.89	0.90	0.89	240
1	0.90	0.88	0.89	240
Accuracy			0.89	480
Macro Avg	0.89	0.89	0.89	480
Weighted Avg	0.89	0.89	0.89	480

In this study, it is noted that RIDGE and LASSO feature selection classifiers recorded similar results. Notably, it is normally a challenging task to obtain important features in any given dataset.

The most notable trend in the classification performance in tables 1-4 above is that the model's level of precision varies significantly for the LASSO, RIDGE, and Random Forest classification models. However, the precision declines significantly, as denoted in Table 4, where all features are present. Performance of the model improved with the LASSO classification than the RIDGE classification due to the qualities of the features selected, which impose performance efficiency on the model. The implications are that the LASSO classifier performs better than RIDGE; however, the Random Forest classifier outperforms both LASSO and RIDGE.

2. Accuracy, Train Time, and Test time

Table 5 below summarizes the accuracy, train time, and test time performances for LASSO, RIDGE, Random Forest feature selection methods as well as when all features are present. Based on the comparative assessment, it is eminent that LASSO had the highest accuracy in testing the truthful reviews, followed by Random Forest then RIDGE. On the other hand, the classifier with all features presents had the lowest accuracy level recorded. Further, LASSO took the least training time, followed by Random Forest, then RIDGE. Finally, classification with all features presents took the most time to train.

Table 5: Comparative Model Performance

Item	Accuracy	Train Time	Test Time
LASSO	0.900	0.451	0.100
RIDGE	0.896	4.578	0.412
Random Forest	0.904	0.575	0.800
All Features Present	0.894	10.644	0.937

Table 5 above demonstrates significant differences between the test times for the diverse embedded methods therein. For example, the model with all features presents took the longest test time (0.94), followed by a Random Forest model (0.80). LASSO took the shortest test time (0.10), followed by the RIDGE with the total test time (0.41). Compared to [21] and [29] in their works, they demonstrated that Random Forest has superior performance than the LASSO and RIDGE. Also, in their study, [30] recorded that the prediction accuracy of Random Forest, LASSO, and elastic Net were 0.539, 0.431, and 0.587, respectively. The findings are akin to the findings in this study, as demonstrated in table 5 above, where Random Forest prediction is superior to LASSO. [31] reported on the superiority of the Random forest-based regression in predicting housing prices given some features. The study compared the performance of Random Forest, RIDGE, LASSO, Naïve Byes, and SVM regression. The conclusion from that study is that the Random Forest prediction algorithm can be used to predict fake social media reviews due to their accuracy in testing positive/truthful reviews.

[32] demonstrated that machine learning and predictive analysis backs high accuracy in project performance analysis. A case in point, Bayesian analysis is more effective than the traditional multiple regression analysis. Similarly, neural networks perform better in prediction analytics than linear regression models. Generally, machine learning models offer higher accuracy in phenomenon prediction than the traditional statistical regression analysis methods. The training of the models in the machine learning approaches offers more compelling findings than the regression analysis. Besides, ambiguous functions are better solved by means of machine learning approaches [33]. Commensurate to these findings, this study demonstrated the prediction accuracy of the embedded approaches. The feature selection in embedded approaches demonstrated high reliability for the classification to predict truthful and fake social media reviews with ambiguous dataset features.

Figures 1, 2, and 3 provide a graphical presentation of the test results for the embedded methods with and without feature selected. Notably, according to figure 1, the Random Forest model produced the highest accuracy; however, the difference is relatively insignificant compared to the rest; LASSO, RIDGE. Failure to perform feature selection causes noise in the model, and hence the accuracy of the model is affected. For example, in figure 1, while all the models produced a reliable level of accuracy in the classification, all the features present produced a lower accuracy level.

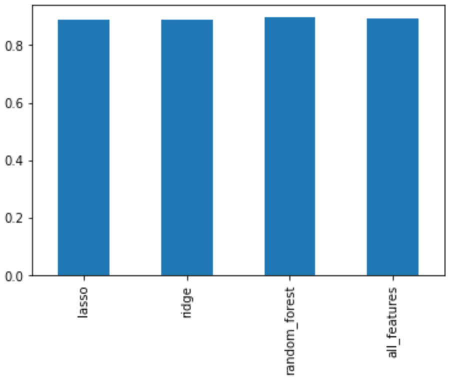

Figure 1: Results Plot on Accuracy

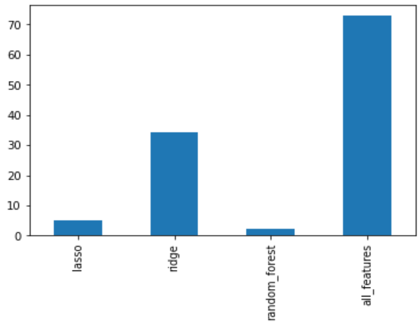

Figure 2: Results Plot on Train Time

Figure 1 denotes that the embedded approaches generally have a higher prediction accuracy because none of the models reported accuracy less than 0.9. On the other hand, Figure 2 summarily denote that modeling with all the features took the most training time. Thus, the training time among the different classifiers therein differs considerably, with models with all features taking the longest time. In addition, figure 3 in the next page demonstrates that, summarily, modeling with all the features had the most testing time; the Random Forest model had the shortest testing time. The study revealed that the accuracy score after using the three embedded feature selection techniques was similar. Nevertheless, the Random Forest approach when used for feature selection had an insignificant edge over the rest.

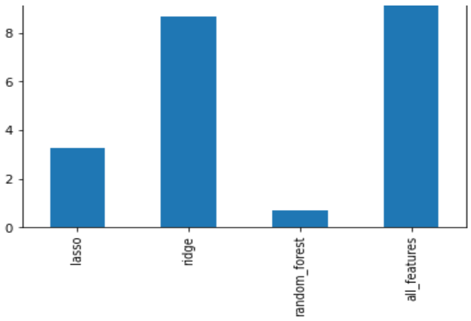

Figure 1: Results Plot Testing Time

3. Comparison of Linear Regression, Machine Learning Algorithms, and Regularization

The standard linear regression models herein may over fit data where limited sites are used for data training and situations where many predictor variables are used. Moreover, the algorithm may not capture some complex relationships between the datasets as it operates on linear relationships. This is often the case because the relationship between the predictor variables might not be dependent in all instances. As such, commensurate to the findings of this study, the embedded approaches perform better in unearthing not only the dependent (interactional) relationships within the data but also the non-interactional/complex relationships within the dataset. On the other hand, Random Forest has a higher predictive ability for datasets with low overfitting, and the contribution of each variable (predictor) in the overall outcome is better determined. Besides, LASSO, contributes to high interpretable output, unlike the standard regression algorithms, which may result in coefficient estimates that are not easy to interpret. In this regard, the embedded approaches are more reliable for the computation of datasets in which the relationship between variables is complex.

V. CONCLUSION

The use of the internet and the centrality of social media channels in marketing have posited the need for trustworthy social media content. On the other hand, massive data get generated in social media, some of which are meant to mislead for social, economic, or political gains. Therefore, there is a need for reliable and robust techniques to distinguish truthful from deceptive social media content, especially social media reviews. Embedded feature selection methods are the most promising tools for detecting fake social media reviews owing to the massive social media content generated daily across the globe. LASSO, RIDGE, and Random Forest are among the embedded methods with the highest precision for predicting fake/negative social media reviews and hence render the online content more reliable to the consumers. Among the vital processes in embedded approaches is the feature selection to boost the prediction accuracy in which the model is significantly simplified. Feature selection methods such as filter, wrapper methods influence a model accuracy, training, and testing time. LASSO, RIDGE, and Random Forest classification models presented depicted different precision rates, training, and test times. Performance of models improved with the LASSO classification than the RIDGE classification due to the qualities of the features selected, which impose performance efficiency on the model. Models with all features presents took the most or the longest test time, followed by the Random Forests approach. LASSO took the shortest test time of 0.10, followed by RIDGE with the total test time of 0.41. In summary, Random Forest classification outperformed LASSO, RIDGE, and the model in when all the dataset features were present. The second-best performing model is LASSO, followed by RIDGE. LASSO often outperforms RIDGE irrespective of the feature selection method used, but Random Forest generally performs better. Again, the penalization technique used has a significant influence on the overall model performance.

ACKNOWLEDGEMENTS

I want to express my sincere appreciation to everyone who played a role in bringing this paper to completion. A special acknowledgment goes to my supervisors, Dr. Ogada and Dr. Mwalili, whose insightful guidance and expertise significantly enhanced the content. I am also grateful for their valuable feedback and constructive suggestions during both the writing and review stages. My heartfelt thanks extend to my family and friends for their unwavering support and understanding throughout the entire research and writing process; their encouragement has been a continuous wellspring of inspiration. The successful realization of this work is attributed to the collective efforts of all those who contributed in various capacities. I extend my gratitude to each one of you for being an indispensable part of this journey.

REFERENCES

1. S. S. Hameed, O. O. Petinrin, A. O. Hashi, and F. Saeed, "Filter-wrapper combination and embedded feature selection for gene expression data," International Journal of Advances in Soft Computing and its Applications, vol. 10, no. 1, pp. 90–105, 2018.
2. A. Ahrens, C. B. Hansen, and M. E. Schaffer, "LASSOPACK: Stata module for lasso, square-root lasso, elastic net, ridge, adaptive lasso estimation and cross-validation," Statistical Software Components, 2020.
3. U. Stańczyk and B. Zielosko, "Heuristic-based feature selection for rough set approach," International Journal of Approximate Reasoning, vol. 125, pp. 187–202, 2020, doi: 10.1016/j.ijar.2020.07.005.
4. S. Matharaarachchi, M. Domaratzki, and S. Muthukumarana, "Assessing feature selection method performance with class imbalance data," Machine Learning with Applications, vol. 6, no. April, p. 100170, 2021, doi: 10.1016/j.mlwa.2021.100170.
5. M. A. S. Mullah, J. A. Hanley, and A. Benedetti, "LASSO type penalized spline regression for binary data," BMC Medical Research Methodology, vol. 21, no. 1, pp. 1–14, 2021, doi: 10.1186/s12874-021-01234-9.
6. S. Sridhar, G. Deena, T. Premanand, and A. Durbha, "A Unique and Dynamic Methodology for Detection Fake News using Neural Network," in IEEE Joint 19th International Symposium on Computational Intelligence and Informatics and 7th International Conference on Recent Achievements in Mechatronics, Automation, Computer Sciences and Robotics, Szeged, Hungary, 2019.
7. M. Razno, "Machine learning text classification model with NLP approach," Computational Linguistics and Intelligent Systems, vol. 2, no. 18-Apr-2019, pp. 71–73, 2019.
8. V. Sabeeh, M. Zohdy, A. Mollah, and R. Al Bashaireh, "Fake News Detection on Social Media using Deep learning and Semantic Knowledge Sources," International Journal of Computer Science and Information Security (IJCSIS), vol. 18, no. 2, pp. 45–68, 2020.
9. J. Tao, X. Fang, and L. Zhou, "Unsupervised deep learning for fake content detection in social media," in Proceedings of the Annual Hawaii International Conference on System Sciences, 2021, pp. 274–283. doi: 10.24251/hicss.2021.032.
10. P. Meel and D. K. Vishwakarma, "Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities," Expert Systems with Applications, vol. 153, p. 112986, 2020, doi: 10.1016/j.eswa.2019.112986.
11. D. Rousidis, P. Koukaras, and C. Tjortjis, "Social media prediction: a literature review," Multimedia Tools and Applications, vol. 79, no. 9–10, pp. 6279–6311, 2020, doi: 10.1007/s11042-019-08291-9.
12. S. Kemp, "Digital 2021: Global Overview Report," Online. Accessed: Oct. 29, 2021. [Online]. Available: https://datareportal.com/reports/digital-2021-global-overview-report
13. F. M. T. Hossain, M. I. Hossain, and S. Nawshin, "Machine learning based class level prediction of restaurant reviews," in 5th IEEE Region 10 Humanitarian Technology Conference 2017, R10-HTC 2017, 2018, pp. 420–423. doi: 10.1109/R10-HTC.2017.8288989.
14. S. Sedhai and A. Sun, "Semi-Supervised Spam Detection in Twitter Stream," IEEE Transactions on Computational Social Systems, vol. 5, no. 1, pp. 169–175, 2018, doi: 10.1109/TCSS.2017.2773581.
15. S. Arefnezhad, S. Samiee, A. Eichberger, and A. Nahvi, "Driver drowsiness detection based on steering wheel data applying adaptive neuro-fuzzy feature selection," Sensors (Switzerland), vol. 19, no. 4, p. 943, 2019, doi: 10.3390/s19040943.
16. B. Dresp-Langley, O. K. Ekseth, J. Fesl, S. Gohshi, M. Kurz, and H. W. Sehring, "Occam's razor for

big data? On detecting quality in large unstructured datasets," Applied Sciences (Switzerland), vol. 9, no. 15, p. 3065, 2019, doi: 10.3390/app9153065.

17. U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: A review," Journal of King Saud University - Computer and Information Sciences, 2019, doi: 10.1016/j.jksuci.2019.06.012.
18. S. Biswas, M. Bordoloi, and B. Purkayastha, "Review on Feature Selection and Classification using Neuro-Fuzzy Approaches," International Journal of Applied Evolutionary Computation, vol. 7, no. 4, pp. 28–44, 2017, doi: 10.4018/ijaec.2016100102.
19. S. Cateni, V. Colla, and M. Vannucci, "A Fuzzy System for Combining Filter Features Selection Methods," International Journal of Fuzzy Systems, vol. 19, no. 4, pp. 1168–1180, 2017, doi: 10.1007/s40815-016-0208-7.
20. Y. B. Wah, N. Ibrahim, H. A. Hamid, S. Abdul-Rahman, and S. Fong, "Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy," Pertanika Journal of Science and Technology, vol. 26, no. 1, pp. 329–340, 2018.
21. R. Guha, M. Ghosh, S. Mutsuddi, R. Sarkar, and S. Mirjalili, "Embedded chaotic whale survival algorithm for filter–wrapper feature selection," Soft Computing, vol. 24, no. 17, pp. 12821–12843, 2020, doi: 10.1007/s00500-020-05183-1.
22. C W. Chen, Y. H. Tsai, F. R. Chang, and W. C. Lin, "Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results," Expert Systems, vol. 37, no. 5, p. e12553, 2020, doi: 10.1111/exsy.12553.
23. H. Roh, Y. J. Oh, M. J. Tahk, K. J. Kwon, and H. H. Kwon, "L1 Penalized Sequential Convex Programming for Fast Trajectory Optimization: With Application to Optimal Missile Guidance," International Journal of Aeronautical and Space Sciences, vol. 21, no. 2, pp. 493–503, 2020, doi: 10.1007/s42405-019-00230-0.
24. J. Chen et al., "A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide," Environment International, vol. 130, p. 104934, 2019, doi: 10.1016/j.envint.2019.104934.
25. R. Mburu, "Comparison of elastic net and random forest in identifying risk factors of stunting in children under five years of age in Kenya," University of Nairobi, 2020.
26. A. Munzel, "Deceptive Opinion Spam," 2016.
27. D. Petrova-Antonova and R. Tancheva, "Data cleaning: A case study with openrefine and trifacta wrangler," in Communications in Computer and Information Science, 2020, pp. 32–40. doi: 10.1007/978-3-030-58793-2_3.
28. I. Pintye, E. Kail, P. Kacsuk, and R. Lovas, "Big data and machine learning framework for clouds and its usage for text classification," Concurrency and Computation: Practice and Experience, vol. 33, no. 19, p. e6164, 2021, doi: 10.1002/cpe.6164.
29. D. Sarkar, "References BT - Lattice: Multivariate Data Visualization with R," Lattice: Multivariate Data Visualization with R, no. references, 2008.
30. R. K. Sarkar, A. R. Rao, P. K. Meher, T. Nepolean, and T. Mohapatra, "Evaluation of random forest n for prediction of breeding value from genomewide SNPs," Journal of Genetics, vol. 94, no. 2, pp. 187–192, 2015, doi: 10.1007/s12041-015-0501-5.
31. H. Yu and J. Wu, "Real Estate Price Prediction with Regression and Classification," Stanford, 2016.
32. S. Sabahi and M. M. Parast, "The impact of entrepreneurship orientation on project performance: A machine learning approach," International Journal of Production Economics, vol. 226, p. 107621, 2020, doi: 10.1016/j.ijpe.2020.107621.
33. L. Wu, Y. Xiao, M. Ghosh, Q. Zhou, and Q. Hao, "Machine Learning Prediction for Bandgaps of Inorganic Materials," ES Materials & Manufacturing, 2020, doi: 10.30919/esmm5f756.

AUTHOR'S DETAILS

1. Nelson B. Wekesa, Department of Computing, Jomo Kenyatta University of Agriculture and Technology, Kenya barasa.nelson@students.jkuat.ac.ke

2. Dr. Kennedy Ogada, Department of Computing, Jomo Kenyatta University of Agriculture and Technology, Kenya kodhiambo@scit.jkuat.ac.ke

3. Dr. Tobias Mwalili, Department of Computing, Jomo Kenyatta University of Agriculture and Technology, Kenya tmwalili@jkuat.ac.ke