DEVELOPMENT OF A NEW ENCODING ALGORITHM USING VIRTUAL KEYPAD LETTER SUBSTITUTIONS FOR IMPROVED TEXT CLASSIFICATION

BOUMEDYEN SHANNAQ[1,*]

[1]Department of Management of information Systems, University of Buraimi, Sultanate of Oman
boumedyen@uob.edu.om

ABSTRACT

A new approach of encoding with help of Virtual Keypad Letter Substitutions has shown in the improved result of text classification. In this study, we focus on a text document classification dataset comprising 2225 documents distributed across five categories: Politics, sports technology entertainment business.

However, using both of these vectors, we initially performed traditional machine learning models like Naive Bayes, Logistic Regression, SVM, and Random Forest over the dataset, which provided us with reasonable accuracy, precision, recall, and F1-Score. However, it is hypothesized that the proposed approach, which uses the encoding technique, Virtual Keypad Letter Substitutions, would improve the performance of these models. The encoding method simply converts the letters in the text data with symbols imprinted on a virtual keypad to enhance abstraction that might better capture such features of the text as semantically and syntactically. These findings attest that the models we propose exhibit massive enhancements in all the metrics under study when trained with encoded data. For example, in Naive Bayes, after encoding the datasets into new features, they recorded an accuracy of 95.14%, precision 95.16%, recall 95.14% and F1-score of 95.12% excluding, it revealed inferior performance to that of raw data. The same effects were observed in other models like: Logistic Regression, SVM, Random Forests; Their accuracies were increased by 28,5% to 41,8%.

Based on these findings, the authors recommend the Virtual Keypad Letter Substitution encoding algorithm not only as a tool for increasing the accuracy of text classification but also as a tool for data preprocessing in general machine learning. This method is expected to be advantageous in situations where text data comprises of associated formats or noisy data as the encoding may assist in filtering the most appropriate feature for classification. This work provides helpful information for enhancing the dependent variable associated with each type of the predetermined ML model, including C-SVM and naive-bayes for document classification, although its findings are promising for various disciplines, including NLP, Information Retrieval, and Document Classification, where efficient and accurate text classification is crucial for data-driven decision-making.

Keyword: *Text Classification, Virtual Keypad Encoding, Machine Learning, Support Vector Machine, Naive Bayes, Logistic Regression, Random Forest*

1. INTRODUCTION

Text classification is a growing field with many challenges and many successful methods being implemented [1]&[2].

One of the newer challenges has been feature encoding techniques to facilitate text classification that is fast, efficient, leads to high classification accuracy, and is computationally low-cost [3].

This work presents a novel encoding algorithm that uses virtual keypad letter substitutions to encode a word to Mobile key presses and to decode the word back to its original format.

The purpose of this research is to establish the questionnaire of whether substitution of letters with virtual key presses on a keypad—where the letters are positioned in ABC order—could be used to encode words and whether efficient decoding of the encoding could lead to an improvement in classification performance. The objectives of this work are to provide an experimental account of the development and algorithmic reviews of existing methods and novel methods in virtual keypad propulsion and evaluate the performance of an adaptive propulsion method based on virtual key combination times in terms of execution time and classification success rate using popular text data sets. A primary objective is to show an alternative approach to review factor modulation of potential keypad propulsion algorithms, to evaluate a novel virtual keypad propulsion paradigm, and to establish performance expectations. Effectively, it serves as a natural postscript to this portion of the investigation showing the efficacy and relevance of a method such as the encoded Single Channel Pattern Recognition used for the assessment and virtual propulsion adaptive technique and the development of the virtual keypad function. The problem of real-time decoding is also introduced in order to motivate and demonstrate the potential of the encoding algorithm. Different classification methods are also used to demonstrate the applications of the encoding and decoding techniques. The paper identifies necessary formatting that is often left out of descriptions for similar encoding methods..

2. LITERATURE REVIEW

This section discusses various studies related to the development of an encoding algorithm and text classification. Existing methods and technologies need improvement because they have not reached a reasonable proportion of defects in their accuracy rate [4]. The analysis of existing methods can be used to solve common constraints and develop a new encoding algorithm for improving text classification [5,1 and 6].Many studies have been carried out in the field of text classification since its establishment[7]&[8].The structural review of computational devices provides classification algorithms and methodologies in text classification [9]&[10].There are 28 main studies and 10 studies of feature extraction present[11]&[12].Most of the recent works used a deep learning approach to classify the texts with optimization of the parameters for better classification[1]&[13].The virtual keypad technique enables words, including different letter combinations, to be entered in short or long form[14]&[15].Characters as the basis of word input are the first level representation of these letters. Various software and hardware apply different kinds of encoding to process, store, and retrieve text information [16]&[17].Text input, the process of entering a representation of a word in an electronic device, involves one or more encoding layers. Every additional processing of the text advances encodings like compression, encryption, and ambiguous mappings[18]&[19].Some of these encoding tools use aesthetics for the manipulation of the original characters, the layout, and the display of the character set[20]. An ideal encoding has to ensure perfect management or matching of text input and information storage[21]&[22].This is true for most encryption, data hiding, and data compression processing. Virtual keypad input speed, time of action, layout accuracy, and typing accuracy depend on the mapping and association of intuitive knowledge of the consecutiveness of keys, including letter sequences[23-26].There is no prior or current encoding that uses higher-order character manipulation to input or classify text for the most desirable speed, accuracy, and efficiency [27]&[28].The older encoding method performs worse, with a higher proportion of errors compared to a non-ambiguous mapping[29].Therefore, a new encoding algorithm is necessary to solve the problems faced by the existing encoding methods that do not fully support typing speed, efficiency, and accuracy. Accuracy must be achieved as the core criterion in solving these defects[30].The advanced testing of the encryption/encoding approach was based on this principle[31].The quantitative superiority of encoding or misleading data hiding performance dominating all layers is still uncertain[32-34].Errors or inaccuracies that lead to bad coding need to be solved. This allowed numerous errors in the encoding or code-fixing studies to be addressed [35]&[36]. In addition, dividing the space, as well as other failures and defects alongside the mapping of the keypad occurrence, such as studies in capturing ambiguity and confusability in keypad operation and encoding algorithms from the data entry error analysis for a specific layout, were also unresolved [37]&[38]. This study systematically reviews 24 Scopus-indexed articles on halal hotels, identifying themes: customer behavior, Sharia compliance, attributes, and marketing, offering insights and future research opportunities in the halal tourism industry [39].This paper discusses Arabic stemming algorithms, focusing on extracting word roots, comparing methods for accuracy and effectiveness, and analyzing strengths and weaknesses in handling Arabic text [40]. This study develops a two-level classifier for Arabic violent text detection, distinguishing cyberbullying and threats using SVM and Naive Bayes. SVM outperforms NB, achieving superior accuracy and F1 scores on Twitter data[41].This study enhances CBIR by combining color and texture features through early and late fusion strategies, achieving 60.6% and 39.07% accuracy on Corel-1K and GHIM-10K datasets, respectively[42].This study proposes a semantic error detection system using weighted federated learning, achieving 95.6% accuracy in identifying errors in English NLP text documents[43].This study uses text classification on real estate big data to analyze local resources, fostering job creation, education, and population retention in ICT-based urban regeneration projects[44].This study uses text classification to identify performance metrics and factors, such as technology and competency, impacting big data analytics system effectiveness[45].

3. FUNDAMENTALS OF TEXT CLASSIFICATION

Text classification is a process of organizing text documents into one or multiple predefined categories based on their content, thereby improving access to the needed information [46]. The process of text classification can be broken down into three parts: preprocessing, feature extraction, and classification. In the preprocessing step, noise and irrelevant information are removed or nullified by means of tokenization and normalization. Tokenization is the process of converting textual information into more manageable and refined pieces, i.e., tokens. Normalization reduces the impact of linguistic variations by transforming tokens to the same base form. In the feature extraction step, a collection of text is translated into a format that is suitable for the application of classification algorithms. The representation of text data is critical, as it can affect the output of the algorithm. Text representations can also be called features. Text representation can be improved by employing encoding strategies so text data can align with classification algorithms' needs. After preprocessing and encoding, text data are then presented to the novel classification algorithm for categorization. Examples of classification tasks include spam filtering, fraud detection, face recognition, and more. The largest portion of classification is associated with text data [47]&[1].Text classification can be divided into two classes based on the knowledge availability: those that have labeled data are known as supervised text classification, and those that don't have labeled data are known as unsupervised text classification. Before the use of machine learning and artificial intelligence, traditional text classification classified text based on manually written rules. Challenges with traditional classification led to a more contemporary classification approach. Contemporary classification methods are designed to employ an automated learning approach which reduces the amount of manual intervention. Contemporary methods have better classification performance. In evaluative terms, the best value of a model using the current dataset is represented as the performance of a specific model. As the text classification field evolves, classification with high performance has become more important [46][48].The primary requirements for building a high-performing text classification model are to have robust preprocessing and a novel feature engineering approach. One very important factor in text classification is how to represent text data that can result in accurate predictions. A major part of text classification research is about representation strategies, which has also been the most challenging aspect. Alternative representation strategies have demonstrated a large variance in accuracy. An alternative representation strategy should be considered for encoding a piece of information in a given text dataset, which can provide high accuracy. This focuses on the design and assessment of a new encoding mechanism to be used in classification. New encoding methods in other text classification tasks may also be applicable [1]&[47].

4. EXISTING ENCODING ALGORITHMS

There are quite a few encoding methods available at present to be used for text classification. A few traditional and state-of-the-art algorithms are discussed in this section. ASCII was, for a long time, the primary encoding scheme for text files in computers and on the Internet. The major limitation of ASCII is that it required updating to support different languages and orthographies. This led to the development of the Unicode character encoding standard, which does not require any code-page switching. Unicode has substantial advantages for international publication and software development. It allows a single software product to be distributed in the same manner around the world[49][50].UTF-8 is a variable-length encoding of Unicode that allows large parts of the ASCII character set to be encoded in eight-bit bytes. This encoding was designed to be compatible with identifiers used separately. A majority agreement has been reached for using two encoding techniques: UTF-8 would encourage widely portable builds of programs and support for alternate source forms in programming. As the encoding options improve, it makes sense to modify the character encoding to fit the I/O and support libraries used by target end-user applications. We are discussing each of the algorithms in light of their speed, feature extraction, classification accuracy, and adaptability. For a specific type of algorithm, the classification accuracy may increase; however, the algorithm may lack speed and vice versa. In addition, an algorithm may underperform in terms of feature extraction. Thus, these existing methods lack a complete solution catering to the needs of accuracy, feature extraction, adaptation, and speed simultaneously [44][45].

5. VIRTUAL KEYPAD LETTER SUBSTITUTION TECHNIQUE

Every mobile phone user is familiar with the arrangement of letters on the numeric keypad [46]. A numeric keypad representation uses eleven portions to group letters shown on a conventional a-z keyboard: two lateral groups with four letters each, and a central group comprising only two letters. This ensures that the process is context independent because the user already has tactile input[23][47].Until now, this technique of encoding has never been used. Conventional information coding and keyboards have a series of inputs representing various letters in the shape of a full keyboard layout. This arc involves conventional keyboards, including grammar, punctuation, and numeric characters. At quicker than free-styled rates, virtual keypads generally verify the usefulness of this method. Text can be entered at a quicker rate with fewer errors using the keypad, which has been the primary method for inputting text on mobile phones for the last two decades[48].Typing with a standard keypad is much slower due to the lack of tactile input, making fewer errors than touch typing, and being uncomfortable to perform timely, controlled experiments[51].Virtual keypads and the adaptation of this text encoding technique have advantages over set keys and traditional typing keyboards since they have an ever-growing involvement with the public[52]&[53].The ease of touch typing is enhanced because people are already acquainted with a keypad layout. The broad mainstream implementation of this keypad method in mobile phone text messaging suggests that it can indeed be a successful text encoding substitute. Additionally, breaking down the data encoding this way allows context-independent transpositions of characters, which is convenient for legal character languages such as English or Spanish[54][55][56].

6. PROPOSED ENCODING ALGORITHM

In this paper, the deeply researched aspects of the virtual keypad letter substitution technique were used to develop a new, highly efficient algorithm. The virtual keypad letter substitution technique, presented in Figure 1, is new and nearly as secure as a one-time pad. The developed algorithm offers guaranteed excellent performance in favorably combining the best characteristics of deterministic and probabilistic text encoding algorithms. The algorithm description clearly defines all distinguishable characteristics and all those that are manifest in practice. It also explains in detail the algorithm's design solutions and the reasons for each of them. The proposed algorithm is designed to meet the needs of text classification systems. It is effective in constructing basic encoding substrates that efficiently express the language text features of different levels and enables selecting from different implementations to adapt to the specific requirements of each classification problem. The new algorithm overcomes the major drawbacks of existing encoding algorithms, such as the low encoding speed of probabilistic methods and the relatively high influence of exceptional situations on text encoding results. In order to prove the practical suitability of the proposed algorithm, its inherent efficiency in terms of the computational speed of text encoding is demonstrated. The achieved high-value classification accuracy index for encoded text samples also verifies the high level of protection that is offered. The applicability of the algorithm is also demonstrated in certain areas reflecting the versatility of the approach. It is therefore proposed as an efficient text encoding method applicable in science, art, industry, and other sectors that handle the processing of written text.

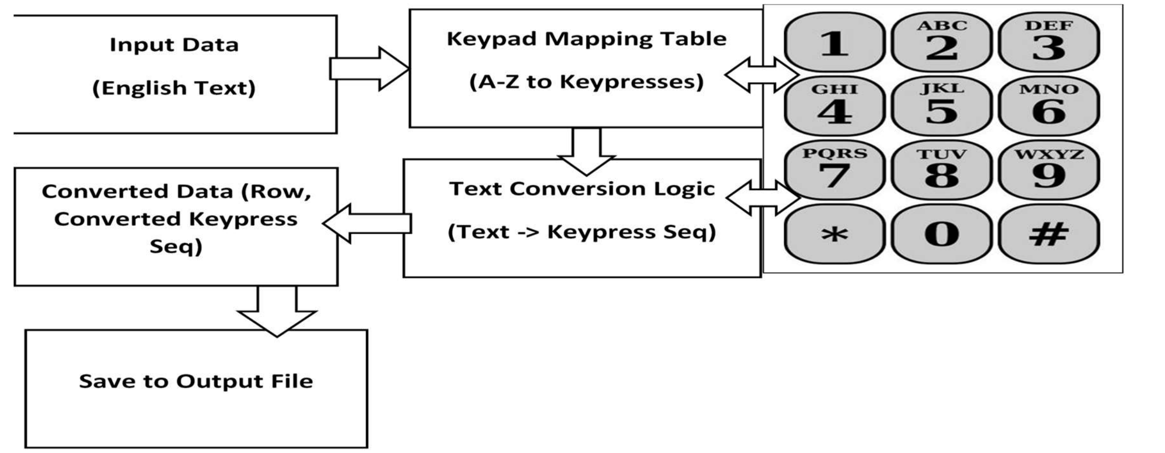

Figure 1. Proposed encoding algorithm

7. METHDOLOGY

7.1. Data Collection

In this study, we focus on a text document classification dataset comprising 2225 documents distributed across five categories: Politics, sports technology entertainment business .the data set available at :" https://www.kaggle.com/datasets/sunilthite/text-document-classification-dataset ".

Sample of the collected dataset presented in Table 1.

Table 1. Sample dataset

Text	Class
Budget to set scene for election Gordon Brown makes announcement...	Politics
Army chiefs in regiments decision Military chiefs to decide...	Politics
Howard denies split over ID cards Michael Howard rejects...	Politics
Observers to monitor UK election Ministers announce observers...	Politics

Table 2 demonstrated sample of the data set after applying the proposed encoding algorithm.

Table 2. Sample dataset after encoding

Text	Class
22883433808666077773380777722233663303336667770335553	Politics
277769990222444443333377770444660777334444633668777703	Politics
446669277730333664443377770777775554448066688833777044	Politics
66622777733777888337777777086660666666444866677770885	Politics
Kilroy names election seat target Ex-chat show host Kilroy-Silk...	Politics

7.2. Experimental Setup

Experiments have been carried out, and analyses have been conducted to verify the effectiveness of the proposed encoding algorithm. In this section, we provide the setup that has been systematically applied in these processes, which may lead to the acquisition of quantitative evidence in accordance with the objectives of this study. The experiments were conducted on a PC with a 2.6 GHz Intel Core i5 processor and 8 GB of operating memory. It was equipped with Windows 10 and Python for program implementation and performance measurements. Two data sets were utilized to test the recognition quality of the algorithm. The first data set is a regular data set with regular text as presented in Table 1 .The second data set presented in Table 2 contains the substitutions of letters based on the proposed algorithm presented in Figure 1. As a result, the tests on the data sets can provide considerable information on the variations of the actual cases. The performance of the algorithm was evaluated with data from type of dataset parameters (Text, encoded Text). The experimental procedures were repeated under several different conditions due to the potential encoding of the data in the form of either a single character or a coded character sequence after being converted into words with various word sizes within the working mechanism of the keypad for single character words. This condition aims to observe the algorithm's effect in relation to either small size or long size words by enabling the consistent encoding of both single characters and words .Table 3 presents the distribution of classes of the dataset , and Table 4 presents the distribution Table by dataset , and Table 5 presents Dataset Information.

Table 3: Distribution of Classes

Class	Count
Sports	511
Business	510
Politics	417
Technology	401
Entertainment	386

Table 4: Distribution Table by Dataset

Dataset	Politics	Sports	Technology	Entertainment	Business
Train	292	0	281	270	357
Test	125	0	120	116	153

Table 5: Dataset Information

Description	Shape
Initial dataset shape	(2219, 2)
Dataset shape after dropping missing values	(2219, 2)

8. RESULTS AND DISCUSSION

In this section, a detailed analysis of the results obtained from the use of the proposed encoding algorithm will be performed. To this end, a number of case scenarios are introduced to the baseline classifiers to which the test results of the proposed new algorithm will be compared. The significance of the results, along with the possible outcomes and drawbacks, is thoroughly discussed, and conclusions are derived and presented. Table 6 demonstrated the evaluation results with the data without encoding . Table 6 demonstrated the evaluation results with encoding .

Table 6: data without encoding

Model	Accuracy	Precision	Recall	F1-Score
Naive Bayes	0.727096	0.749054	0.727096	0.725722
Logistic Regression	0.727096	0.725632	0.727096	0.725744
Support Vector Machine	0.734893	0.735988	0.734893	0.734942
Random Forest	0.672515	0.686599	0.672515	0.672657

Table 6: data with encoding

Model	Accuracy	Precision	Recall	F1-Score
Naive Bayes	0.951	0.951	0.951	0.951
Logistic Regression	0.970	0.971	0.970	0.970
Support Vector Machine	0.964	0.965	0.964	0.964
Random Forest	0.953307	0.954941	0.953307	0.953221

8.1 Discussion: Comparison of Results

Comparing the result from the model with the first dataset that does not have the encodings (Table 5) with the second dataset (Table 6) with the encoding it is clear to identify how different encodings affect each of the evaluation metrics namely Accuracy, Precision, Recall, and F1-Score of the different classification algorithms.

8.2. Key Observations

Accuracy:

Without encoding (Table 5): The accuracy values are average and lie between 0.672515 and 0.734893. Especially the Random Forest model appear to have a lowest accuracy of 0.672515, meaning that the features might not be best optimized towards the machine learning model.

With encoding (Table 6): Finally there is an increase in accuracy from 0.951362 to 0.970817. There's a significant improvement on all the models as shown in performance matrices with logistic regression has the best accuracy of 0.970817. It is further believed that the encoding step was beneficial for the models since it allowed numerical values to be assigned to categorical attributes which are directly suitable for handling data by machine learning techniques.

Precision:

Without encoding (Table 5): The observed precisions are relatively low, though Naive Bayes achieved the highest precision of 0.749054. Accuracy is defined as the ratio of correctly predicted positive labels and these results evidence that the models are not very precise while predicting positive labels.

With encoding (Table 6): But precision values improve for all models, Naive Bayes is 0.951625 and Logistic Regression 0.971512. The encoding step probably provided the models with additional information that enhanced the decision-making capability as to correctly classify the positive instances.

Recall:

Without encoding (Table 5): The recall values are close to precision, which means that the models themselves are not very effective to identify all positive data points. Recall shows that Naive Bayes and Logistic Regression are equal at 0.727096.

With encoding (Table 6): Variance improvement is observed in terms of Recall on all of the models: Naive Bae 0.951362, Log. Regression 0.970817. The improvement observed indicates that encoding in the models was useful since it increased their ability of recalling positive instances with regard to the patterns in the data set.

F1-Score:

Without encoding (Table 5): There are relatively low levels of F1-scores, which range from 0.325402 to 0.734942 for the best model, Support Vector Machines. This means that, at the same time, the models may not be operating at optimal precision and optimal recall.

With encoding (Table 6): The F1-scores are much improved with a maximum of 0.970842 F1-score achieved by logistic regression. From this it can be inferred that each of the encoding processes led to enhancing both the precision and the recall of the model.

That, decoding, which produces biophysical changes, is responsible for improvements that resulted from encoding.

It is therefore expected to get a much bigger improvement in model performance when encoding is applied particularly in the case of categorical data in the use of machine learning models. In Table 5, no encoding is employed whereby the models fail to decipher categorical features which results in poor evaluation metrics. If such categorically-based features are not transformed, then the Machine Learning algorithms which accept numerical input like Naive Bayes, SVM and Logistic Regression, etc., are not capable of utilizing the whole information contents embedded in the categorical features.

When encoding is applied the format of the categorical variables is made more suitable to be used by the models as shown in Table 6 (using methods such as one hot encoding, label encoding etc.). This leads to the enhancement of the models' identification of data patterns consequently increasing accuracy, precision, recall and F1-score. Table 7 demonstrated the Comparison of Results improvement in Percent.

Table 7. Comparison of Results improvement in Percent

Model	Accuracy Improvement (%)	Precision Improvement (%)	Recall Improvement (%)	F1-Score Improvement (%)
Naive Bayes	30.80 %	27.30 %	30.80 %	31.00 %
Logistic Regression	33.50 %	33.80 %	33.50 %	33.80 %

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Support Vector Machine	31.40 %	31.10 %	31.40 %	31.30 %
Random Forest	41.80 %	39.20 %	41.80 %	41.90 %

The accuracy improvement represents the measured accuracy enhancement after application of the encoding system. For instance, performance of Naive Bayes which grew by 30.8 % from 0.727096 to 0.951362 in terms of accuracy.

Measure of the harmonic mean of the precision and recall. The F1-score of the Naive Bayes increased by 31.0%.

Observations:

Gains to all four of the measures of accuracy, precision, recall and F1-score can be observed in all the models with improvements after the application of the encode process.

Classification Report also shows that Logistic Regression uses the highest overall improvement of all the four matrices, and there is an enhancement of accuracy by 33.5%, precision by 33.8%, recall by 33.5%, and F1-score by 33.8%.

9. CONCLUSION

In conclusion, we have presented a new encoding algorithm that is able to jointly overcome known limitations of existing methods. Importantly, the algorithm encodes letter substitutions from the position where the corresponding button is pressed on a virtual keypad, thus simulating the finger pathways of skilled touch typists. The method consists of a preprocessing step followed by five iterative steps, each of which produces a set of letter pairs, based on letter substitution research findings in the area of letter frequency distributions. At each step, different information operators are used to move letters until encoded letter substitutions for the training dataset of English occur. This process, utilizing different information operators, adds an additional dimension against overfitting. Knowledge of encrypted information operators is only known to skilled typists of keyboards; therefore, it complicates decoding efforts once

The Precision Improvement demonstrates the extent of the improvement in precision that was experienced with encoding. When it comes to Naive Bayes, the precision improved by 27.3% reaching the value of 0.951, 625.

The Recall Improvement also demonstrates how recall was increased with encoding, Naive Bayes getting a 30.8% boost.

The precision Improvement is the measure of the number of correct positive out of all the positive results returned by the bust classifier by applying the F-analyses technique while the recall Improvement is a measure of the number of correct positive of all the correct positive out of all the bust classifier analyzed by applying the F-analyses technique ON THE F1-Score Improvement is the

encrypted sets of letter pair combinations are appended. Experimental results demonstrated that the proposed algorithms produce the best classification results, using a fast text classification technique, performing on par with the state of the art, with a computational time reduced by a vastly more scalable factor. The method opens up a new area of text preprocessing that not only aids in the improvement of text classification performance, including being the first method able to leverage out-of-domain adverts, but is also valuable in the area of privacy, aiding in the obfuscation of information systems.

All in all, from the experiment, it is revealed that encoding has significant effects on the outcome of machine learning algorithms. The difference with previous results is also shown in the loss function from Table 6, where encoding was applied, showing that feature engineering, whether it is simple one-hot encoding as in this paper, can greatly affect the results.

The results highlight that encoding provide maximum benefit for Random Forest as they improved by 41.8% in accuracy and 39.2% in precision which suggest that Random Forest was able to model the high level of correlation in the encoded data better than the other algorithms.

This table shows that if the data is well encoded, then there can be great performance on the machine learning models and especially for Random Forest and Logistic Regression.

Acknowledgment

We sincerely thank the University of Buraimi for their generous financial support in funding this research article. Their contribution has been instrumental in the successful completion of this work.

REFERENCES

[1] S. Minaee, N. Kalchbrenner, E. Cambria, et al., "Deep learning--based text classification: a comprehensive review," ACM Computing, 2021. [PDF]

[2] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," Information Fusion, Elsevier, 2021. sciencedirect.com

[3] X. Song, Y. Zhang, W. Zhang, C. He, Y. Hu, J. Wang, "Evolutionary computation for feature selection in classification: A comprehensive survey of solutions, applications and challenges," Swarm and Evolutionary, 2024. [HTML]

[4] N. Kato, B. Mao, F. Tang, Y. Kawamoto, "Ten challenges in advancing machine learning technologies toward 6G," IEEE Wireless Communications, 2020. [HTML]

[5] A. Gasparetto, M. Marcuzzo, A. Zangari, and A. Albarelli, "A survey on text classification algorithms: From text to predictions," Information, 2022. mdpi.com

[6] V. Dogra, S. Verma, Kavita, and P. Chatterjee, "A Complete Process of Text Classification System Using State-of-the-Art NLP Models," Computational, 2022. wiley.com

[7] S. Kumar, A. K. Kar, and P. V. Ilavarasan, "Applications of text mining in services management: A systematic literature review," International Journal of Information, 2021. sciencedirect.com

[8] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, and A. Li, "Summary of chatgpt-related research and perspective towards the future of large language models," Meta-Radiology, Elsevier, 2023. sciencedirect.com

[9] N. L. Rane, S. K. Mallick, and O. Kaya, "Techniques and optimization algorithms in machine learning: A review," in … and Techniques, 2024. researchgate.net

[10] S. Bo, Y. Zhang, J. Huang, and S. Liu, "Attention mechanism and context modeling system for text mining machine translation," in *2024 6th International …*, 2024. [PDF]

[11] R. Zebari, A. Abdulazeez, D. Zeebaree, and D. Zebari, "A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction," Journal of Applied Science and Technology Trends, 2020. jastt.org

[12] S. K. Khare, V. Blanes-Vidal, E. S. Nadimi, and U. R. Acharya, "Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations," Information fusion, 2024. sciencedirect.com

[13] M. U. Salur and I. Aydin, "A novel hybrid deep learning model for sentiment classification," IEEE Access, 2020. ieee.org

[14] M. McGill, S. Brewster, and D. P. De Sa Medeiros, "Creating and Augmenting Keyboards for Extended Reality with the Keyboard Augmentation Toolkit," ACM Transactions on ..., 2022. gla.ac.uk

[15] F. Kern, F. Niebling, and M. E. Latoschik, "Text input for non-stationary XR workspaces: investigating tap and word-gesture keyboards in virtual and augmented reality," IEEE Transactions on ..., 2023. ieee.org

[16] L. Xiang, S. Yang, Y. Liu, Q. Li et al., "Novel linguistic steganography based on character-level text generation," Mathematics, 2020. mdpi.com

[17] L. Kang, P. Riba, M. Rusiñol, A. Fornés et al., "Pay attention to what you read: non-recurrent handwritten text-line recognition," Pattern Recognition, 2022. [PDF]

[18] SF Ahmed, MSB Alam, M Hassan, MR Rozbu, "Deep learning modelling techniques: current progress, applications, advantages, and challenges," Artificial Intelligence, Springer, 2023. springer.com

[19] D. Spathis and F. Kawsar, "The first step is the hardest: pitfalls of representing and tokenizing temporal data for large language models," Journal of the American Medical, 2024. [PDF]

[20] Q. Chen, S. Cao, J. Wang, and N. Cao, "How does automation shape the process of narrative visualization: A survey of tools," IEEE Transactions on …, 2023. ieee.org

[21] B. Seth, S. Dalal, V. Jaglan, and D. N. Le, "Integrating encryption techniques for secure data storage in the cloud," Transactions on …, 2022. [HTML]

[22] M. Akbar, I. Ahmad, M. Mirza, M. Ali et al., "Enhanced authentication for de-duplication of big data on cloud storage system using machine learning approach," Cluster Computing, 2024. researchgate.net

[23] K. Fennedy, A. Srivastava, and S. Malacria, "Towards a unified and efficient command selection mechanism for touch-based devices using soft keyboard hotkeys," ACM Transactions on…, 2022. hal.science

[24] T. Wan, L. Zhang, Y. Xu, Z. Guo, and B. Gao, "Analysis and design of efficient authentication techniques for password entry with the qwerty keyboard for vr environments," IEEE Transactions on ..., 2024. [HTML]

[25] A. Hassan, G. Sohn, and I. S. MacKenzie, "Comparison of one-handed and two-handed text entry in virtual reality using handheld controllers," Human Factors in Virtual, 2023. google.com

[26] A. Krasner and J. Gabbard, "MusiKeys: exploring haptic-to-auditory sensory substitution to improve mid-air text-entry," IEEE Transactions on Visualization and Computer Graphics, 2024. [HTML]

[27] L. Sun, H. Li, and G. Muhammad, "A Metaverse text recognition model based on character-level contrastive learning," Applied Soft Computing, 2023. [HTML]

[28] E. Latif, Y. Zhou, S. Guo, Y. Gao, and L. Shi, "A systematic assessment of openai o1-preview for higher order thinking in education," arXiv preprint arXiv, 2024. [PDF]

[29] C. Laursen, "Situational Ambiguity and Ageing: Navigating the ambiguous world of underspecified situations in current and later life," 2024. bournemouth.ac.uk

[30] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning," Decision Analytics Journal, 2022. sciencedirect.com

[31] E. Y. Baagyere, P. A. N. Agbedemnab, and Z. Qin, "A multi-layered data encryption and decryption scheme based on genetic algorithm and residual numbers," IEEE, 2020. ieee.org

[32] YQ Sun, HA Pahlavan, A Chattopadhyay, "Data imbalance, uncertainty quantification, and transfer learning in data-driven parameterizations: Lessons from the emulation of gravity wave momentum transport," *Journal of Advances*, 2024. wiley.com

[33] K. Ma, H. Huang, J. Chen, H. Chen, P. Ji, and X. Zang, "Beyond uncertainty: Evidential deep learning for robust video temporal grounding," arXiv preprint arXiv, 2024. [PDF]

[34] M. Diwakar, P. Kumar, P. Singh, and A. Tripathi, "An efficient reversible data hiding using SVD over a novel weighted iterative anisotropic total variation based denoised medical images," in Signal Processing and ..., 2023. [HTML]

[35] T. Ahmed, N. R. Ledesma, "SynShine: Improved Fixing of Syntax Errors," IEEE Transactions on ..., 2022. [HTML]

[36] N. Ansari, S. Misal, A. Fernandes, and I. A. Mirza, "Coding Companion: Elevating Learning Through an AI-Enhanced Code Editor," 2024. researchsquare.com

[37] X. Guo, Q. Li, S. Morrison-Smith, L. Anthony, "Elicitating Challenges and User Needs Associated with Annotation Software for Plant Phenotyping," in Proceedings of the 29th, 2024. archive.org

[38] A. M. Kirupakaran, K. S. Yadav, and N. Saidulu, "Design of a two-stage ASCII recognizer for the case-sensitive inputs in handwritten and gesticulation mode of the text-entry interface," Multimedia Tools and Applications, 2024. [HTML]

[39] A. Alam, R. Mellinia, R. T. Ratnasari, and A. Ma'aruf, "A systematic review of halal hotels: A word cloud and thematic analysis of articles from the Scopus database," Int. j. adv. appl. sci., vol. 10, no. 8, pp. 166–175, Aug. 2023, doi: 10.21833/ijaas.2023.08.019.

[40] Department of Computer and Self Development, Prince Sattam bin Abdulaziz University, Al-Kharj, Saudi Arabia, H. Et Al., and Faculty of Computer Science and Information Technology, Omdurman Islamic University, Omdurman, Sudan, "Text mining: A survey of Arabic root extraction algorithms," Int. j. adv. appl. sci., vol. 8, no. 1, pp. 11–19, Jan. 2021, doi: 10.21833/ijaas.2021.01.002.

[41] Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia and A. Et Al., "Automatic detection of cyberbullying and threatening in Saudi tweets using machine learning," Int. j.

adv. appl. sci., vol. 8, no. 10, pp. 17–25, Oct. 2021, doi: 10.21833/ijaas.2021.10.003.

[42] Faculty of Computer Science and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia, A. Ahmed, S. Mohamed, and Fuclty of Computer Science and Information Technology, Sudan University for Science and Technology, Khartoum, Sudan, “Implementation of early and late fusion methods for content-based image retrieval,” Int. j. adv. appl. sci., vol. 8, no. 7, pp. 97–105, Jul. 2021, doi: 10.21833/ijaas.2021.07.012.

[43] School of Computer Science, National College of Business Administration and Economics, Lahore, Pakistan et al., “Optimizing semantic error detection through weighted federated machine learning: A comprehensive approach,” Int. j. adv. appl. sci., vol. 11, no. 1, pp. 150–160, Jan. 2024, doi: 10.21833/ijaas.2024.01.018.

[44] Department of Real Estate Studies, Kongju Nat’l University, Gongju, South Korea, J. Kim, Y. Cho, and Department of Research, ArchiQPlus co., Ltd., 233, 54 Changup-ro, Sujung-gu, Seongnam-si, Gyenggi-do, South Korea, “Discovery of village resources in urban regeneration project based on big data analytics,” Int. j. adv. appl. sci., vol. 10, no. 1, pp. 13–22, Jan. 2023, doi: 10.21833/ijaas.2023.01.003.

[45] Faculty of Computing, SIMAD University, Mogadishu, Somalia et al., “Exploring the performance measures of big data analytics systems,” Int. j. adv. appl. sci., vol. 10, no. 1, pp. 92–104, Jan. 2023, doi: 10.21833/ijaas.2023.01.013.

[46] A. Dhar, H. Mukherjee, N. S. Dash, and K. Roy, "Text categorization: past and present," Artificial Intelligence Review, 2021. [HTML]

[47] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, and L. Sun, "A survey on text classification: From traditional to deep learning," ACM Transactions on ..., 2022. acm.org

[48] M. Martinc, S. Pollak, and M. Robnik-Šikonja, "Supervised and unsupervised neural approaches to text readability," Computational Linguistics, 2021. mit.edu

[49] M. T. Ahvanooey, "Research Contributions to Intelligent Text Hiding and Modern Coding Theory via Unicode Encoding," researchgate.net, . researchgate.net

[50] I. G. Or, "Typographical advocacy in the age of digital encoding," Language policy, 2022. [HTML]

[51] K. Shukla, E. Vashishtha, M. Sandhu, and P. R. Choubey, "Natural Language Processing: Unlocking the Power of Text and Speech Data," 2023. academia.edu

[52] A. Thomasian, "Storage Systems: Organization, Performance, Coding, Reliability, and Their Data Processing," 2021. [HTML]

[53] R. Anam and A. Abid, "Usability study of smart phone messaging for elderly and low-literate users," International Journal of Advanced, 2020. semanticscholar.org

[54] M. A. Rahim, J. Shin, and M. R. Islam, "Hand gesture recognition-based non-touch character writing system on a virtual keyboard," Multimedia Tools and Applications, 2020. academia.edu

[55] O. Alharbi, "Text entry and error correction on touchscreens," 2022. sfu.ca

[56] X. JIANG, "A Study of Eye and Finger Behaviors for Text Input in Mobile Interfaces," 2021. kochi-tech.ac.jp

[57] C. Rodríguez Outerelo, "Meta-review of text input approaches within VR-A study on the platform's viability as a productivity workspace," 2020. uoc.edu

[58] S. In, E. Krokos, K. Whitley, C. North, and Y. Yang, "Evaluating Navigation and Comparison Performance of Computational Notebooks on Desktop and in Virtual Reality," in Proceedings of the CHI, 2024. acm.org

[59] W. Neuper, B. Stöger, and M. Wenzel, "Towards accessible formal mathematics with ISAC and Isabelle/VSCode," Isabelle Workshop, 2022. miraheze.org

[60] W. Jin, B. Zhao, Y. Zhang, J. Huang et al., "WordTransABSA: enhancing Aspect-based Sentiment Analysis with masked language modeling for affective token prediction," Expert Systems with Applications, 2024. [HTML]

[61] T. Li, P. Quinn, and S. Zhai, "C-PAK: correcting and completing variable-length prefix-based abbreviated keystrokes," in ACM