Classifying Text and Image of Social Media Post Using Hybrid Feature Selection Technique

Sumit Jain (Ph.D. Scholor)[#1], Dr. Hare Ram Sah (Professor)[*2]
Department of Computer Science & Engineering, Sage University, Indore, India
[1]sumitjain1679@gmail.com(SKITM)
[2]ramaayu1@gmail.com

Abstract:
Social media is a platform that accumulates a significant amount of user-generated data without any proper control, which poses a potential threat to individuals and communities. This research paper aims to contribute in three main areas: (1) investigating various techniques used for feature selection in text and image-based data analysis for social media, (2) conducting experiments to demonstrate the impact of different feature selection models on classifier performance for text and images, and (3) developing a novel approach to combine features from text and images for social media data classification. To achieve this, we utilized a dataset consisting of Twitter posts and text-based images from Kaggle. We first employed Optical Character Recognition (OCR) to extract text from images on social media and aligned the images with their corresponding text. We then utilized TF-IDF and chi-square tests to identify the features from the combined image and text data. The experimental results demonstrate that our proposed approach outperforms other techniques and provides an acceptable accuracy rate of up to 89%.

Keywords: **Text Feature selection, Image Feature selection, machine learning algorithm, heterogeneous data, social media data.**
DOINumber:10.48047/nq.2022.20.22.NQ10346 **NeuroQuantology2022;20(22):3469-3481**

I. INTRODUCTION

The rapid growth of digital data in social media applications has created opportunities to develop data-driven applications utilizing machine learning (ML) techniques. However, social media data comes in different formats including text, image, video, and audio, which requires specialized techniques for effective analysis. Among these formats, text and image-based data are the most common forms of content on social media platforms like Facebook and Twitter. While there are several techniques available to analyze these data formats individually, few approaches can handle both image and text data. However, these methods are facing potential performance issues, including classification accuracy and computational resource costs. Therefore, researchers and developers need to explore new techniques and strategies to overcome these challenges and develop effective data-driven applications that utilize social media data for social welfare.

The main objective of this paper is to develop a Machine Learning (ML) based model for analyzing social media data. The model is designed to process both text and image data formats. To achieve this objective, the paper first provides a review of various text and image feature selection techniques used in recent social media data analysis. The review helps to identify the strengths and weaknesses of these techniques, providing a comprehensive understanding of their suitability for processing social media data.

In addition to the review, the paper also includes a comparative performance study of some text and image feature selection techniques. This study provides a detailed analysis of the techniques' effectiveness in processing social media data, allowing for an informed choice of the best technique to use.

Finally, the paper proposes a model that provides the classification of both image and text data of social media. The model uses the identified best feature selection technique(s) and ML algorithms to effectively classify social media data. By providing a model that can handle both text and image data, the paper offers a comprehensive approach to social media analysis, enhancing the accuracy and completeness of the analysis results. The entire document has been organized in the following manner:

1. The first section discusses the key aim and objective of the paper, which is to develop an ML-based model for analyzing social media data in both text and image formats.
2. The second section provides a review of recent techniques used in text and image data analysis in social media.
3. The third section discusses an experimental model for comparing text feature selection techniques and image feature selection techniques, and it presents the experimental performance of these techniques.
4. The fourth section provides a model that is used to classify text and images from social media.
5. The next section discusses the experimental analysis of the developed model.
6. Finally, the paper concludes with a summary of the study and a plan for future extensions.

II. LITERATURE REVIEW

In this section, three main components are presented. Firstly, a list of essential abbreviations used in various reviewed articles is provided. Secondly, recent techniques utilized for analyzing text-based social media data are discussed, followed by a description of techniques used for image-based classification tasks.

Abbreviation

The Table 1 shows the essential keywords used in survey.

Table 1 Abbreviation

S. No.	Term	Full Form
1	GLCM	Gray Level Co-occurrence Matrix
2	LBP	Local Binary Pattern
3	KNN	K- Nearest Neighbor
4	NB	Naïve Bays
5	SVM	Support Vector Machine
6	AFS	Active Feature Selection
7	RLSD	Regional Latent Semantic Dependencies model
8	HIS	Hyper Spectral Image
9	ALO	Ant Lion Optimizer
10	CNN	Convolutional Neural Networks
11	RNN	Recurrent Neural Networks
12	WSVM	Wavelet Support Vector Machine
13	NNLMs	Neural Network Language Models
14	LDA	Linear Discriminant Analysis
15	RF	Random Forest
16	VSC	Visual Sentiment Classifier

A. Classifiers Used

Machine learning classifiers are an essential component in the analysis of both text and image

data in social media. These classifiers can be used to classify data into different categories, such as positive or negative sentiment, or to identify different objects or features in images. To develop effective techniques for text and image analysis, it is important to identify the most appropriate classifiers for each task.

Table 2 provides a list of popular classifiers and their use in different research work. The table shows that several classifiers are commonly used in both text and image-based analysis tasks, including Support Vector Machines (SVMs), k-Nearest Neighbor (k-NN), and Neural Networks (NNs). These classifiers are used to perform tasks such as sentiment analysis, topic classification, object recognition, and image segmentation.

By analyzing the list of classifiers in Table 2, it becomes clear that similar classifiers can be used to process both text and image data. For example, SVMs have been used for both sentiment analysis of text data and object recognition in images. Similarly, k-NN has been used for both topic classification of text data and image segmentation. These findings suggest that the same classifiers can be used to process both types of data, making it easier to develop unified techniques for analyzing social media data.

Table 2 Classifiers used in Text classification and Image classification

Classifiers	**Image classification**	**Text classification**
KNN	[6] [29]	[7]
NB	[6]	[7]
SVM	[6] [12] [18] [29]	[7] [8] [12] [13] [19] [25]
Deep CNN	[9] [24] [28] [30]	[26]
RNN	[15] [20]	[23]
VSC	[20]	
WSVM	[21]	
NNLMs	[22]	
LDA	[29]	
RF	[29]	[19]
ZeroR	[29]	
Image Processing Pipeline	[9]	
K-Means		[7] [16] [17]
LSTM		[14]

B. Features, Dataset Used and Results

In developing techniques for analyzing text and images, experimental datasets are a crucial component. Table 3 illustrates the various datasets that have been used in different experimental studies, along with the feature selection techniques used to obtain features from these datasets. The table also presents the results obtained from the combination of datasets and feature selection techniques. The review findings indicate that the same classifier can be used for classifying different data formats and their features. However, it is important to treat the process of identifying features from text or images separately. Moreover, the classification performance can be influenced by the combination of different feature selection techniques and classifiers.

III. COMPARING FEATURE SELECTION TECHNIQUES

The objective of this study is to experimentally investigate the influence of various feature selection

techniques on the performance of classification algorithms. To achieve this objective, we have developed a model that compares the performance of feature selection methods for two types of social media data: text and image. Figure 1 displays the model we have created to carry out this comparison, and the various components of the model are explained in this discussion.

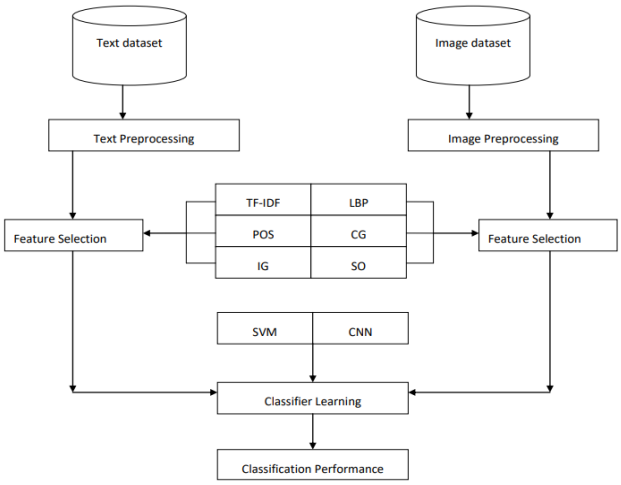


Figure 1 Implemented Model for comparing feature selection techniques and classifiers

In order to investigate the effects of different feature selection techniques on the performance of classification algorithms, a model was developed. The model, depicted in Figure 1, comprises various components that will be discussed in detail. The first components of the model are the datasets, which consist of two different sets: a text dataset based on Twitter social media posts and a plant leaf image dataset from Kaggle. These datasets are utilized as input and undergo distinct preprocessing procedures. The image dataset is preprocessed using Equation (1):

$$pI = \frac{I}{255} \ldots\ldots\ldots (1)$$

Table 3 Literature summary for Text and Image analysis techniques

Author	Features	Dataset
Image Classification		
[6]	GLCM, LBP, Gabor filter, K-means	Kelberg dataset, Brodatz-1, Brodatz-2
[11]	-	-
[9]	-	Social media images
[10]	De-duplication and relevancy filtering	Twitter Images during, Typhoon Ruby, Nepal Earthquake, Ecuador Earthquake, Hurricane
[12]	AFS	EmoDB, SAVEE and EMOVO
[15]	RLSD	VOC PASCAL 2007, Microsoft COCO and NUS-WIDE
[18]	Gain ratio	Aerial vehicle images and ISPRS
[20]	-	More than 3 million tweets (text and images)

[21]	ALO	Botswana, KSC dataset, Indian Pines
[22]	-	Finnish, Arabic, Swedish, English
[24]	-	Labeled Faces Wild (LFW)
[28]	Bag of visual features	1000 web pages
[29]	Levy Flight-Based Grey Wolf Optimization	BOSSbase ver 1.01 image dataset
[30]	-	Interior images of Caffe Nero and Arabica Coffee
Text Classification		
[7]	TF-IDF, IG, MI, χ2, AM, TS, TF-RF, SFS	Reuters-21578, 20-Newsgroups and 4 University dataset
[8]	Tax2vec, Taxonomy-based features	PAN 2017 (Gender), MBTI (Myers-Briggs personality type), PAN 2016 (Age)
[12]	AFS	EmoDB, SAVEE and EMOVO
[13]	DF, IG, Gini index, Cross-entropy, Class discriminate, χ2, Odds ratio	Pang & Lee1, Pang & Lee 2, Imdb, Farm-ads, Spam, 20-newsgroup 1, 20-newsgroup 2, Cade 1, Cade 2, Reuter8
[14]	Word embedding	PTT posts
[16]	H-FSPSOTC	20Newsgroups
[17]	FSPSOTC	Reuters−21578, 20Newsgroups
[19]	PoS tagged	48 healthy controls and 48 impaired subjects.
[23]	-	SCHOLAT and Fudan University document
[25]	Improved Chi-square	Dataset of 5070 Arabic documents
[26]	-	Kaggle's competition regarding Wikipedia's talk
[27]	Information retrieved	Examples of approaches for real-world challenges

Furthermore, in order to preprocess the text data, we have applied various steps such as the removal of stop words and special characters. After the preprocessing step, both the text and image datasets are utilized with feature selection techniques. A provision has been developed to select either a single feature selection technique or a combination of multiple techniques. Finally, two popular machine learning classifiers are used for the training and validation of the model.

There are two experimental scenarios are implemented:

1. The first scenario involves demonstrating the performance of individual feature selection algorithms.
2. The second scenario involves testing different combinations of feature selection techniques.

Table 4 displays the performance of various feature selection methods using two machine learning models. Table 5 shows the performance of feature selection method combinations, where model performance is evaluated in terms of accuracy and training time. Specifically, Table 4 showcases the accuracy and training time for text and image feature selection techniques utilizing CNN and SVM classifiers. Results demonstrate that for text features, CNN outperforms SVM. Moreover, the IG-based technique performs better for both classifiers in comparison to TF-IDF and POS-based features. Therefore, CNN provides efficient results for text classification scenarios and IG requires less training time. In the case of image data, CNN consumes less training time than SVM, and color grid movement is also faster. Based on these results, TF-IDF-based features are recommended for text classification, while SO-based features are recommended for image classification, as they provide better results compared to other feature selection methods.

We found that the CNN classifier performs better than SVM for both datasets. Therefore, we used the CNN classifier to evaluate the combinations of feature selection techniques in Table 5, which showcases the performance in terms of accuracy and training time. Our results indicate that the combinations of feature

selection techniques outperform individual feature selection techniques in both image and text classification scenarios based on classification accuracy.

Table 4 Classification outcomes for text and image datasets

		Text Features			Image Features		
		TF-IDF	IG	POS	CG	LBP	SO
Accuracy (%)							
1	SVM	75.31	79.032	71.242	58.2	59.6	69.7
2	CNN	84.26	86.74	75.53	70.5	54.8	76.4
Training Time (Sec)							
3	SVM	268.95	234.97	279.05	543.87	876.8	826.96
4	CNN	82.434	76.987	82.903	432.88	478.91	454.95

However, as the size of feature vectors increases, the time requirements also increase. Our findings show that for image analysis, the combination of CG, LBP, and SO yields higher classification accuracy, but with higher time consumption. Alternatively, the combination of CG and SO demonstrates similar accuracy, but with comparatively lower time consumption. Thus, color and edge features are more significant than other combinations. Similarly, for text classification, the combination of TF-IDF, POS, and IG yields higher accuracy but with higher time consumption. However, by reducing dimensions, we can manage time consumption. Therefore, hybrid features are more advantageous than individual features.

Table 5 Performance of combined feature classification using CNN

Combinations of Text Features				
	TF-IDF + IG	TF-IDF + POS	IG + POS	IG + POS + TF-IDF
Accuracy	89.773	82.549	87.48	92.976
Training Time	189.27	159.81	175.47	265.64
Combinations of Image Features				
	CG + LBP	CG + SO	LBP + SO	LBP + SO + CG
Accuracy	65.87	83.27	78.53	85.74
Training time	562.97	559.76	762.91	887.28

For this experiment, we considered three text feature selection techniques and three image-based feature selection techniques. Furthermore, we used SVM and CNN classifiers to classify the features. Based on the experimental results we obtained, the following conclusions were drawn:

1. Hybrid features are more effective than individual features.
2. For text classification, Information gain yields higher accuracy while Sobel filter yields higher accuracy for image data.
3. The combination of IG, POS, and TF-IDF yields higher accuracy but requires more time.
4. Regarding image feature selection, two combinations, namely (LBP, SO and CG) and (CG and SO), provide similar accuracy, but the combination of LBP, SO and CG is more computationally expensive.

IV. PROPOSED SOCIAL MEDIA POST CLASSIFICATION

Based on the literature review, various approaches are available for analyzing text and image data. However, most authors have developed separate techniques for each type of data since the features extracted from text and images are completely different. Nonetheless, there are some techniques available that combine features from both text and images. Motivated by this concept, the proposed work aims to develop an approach that can work on social media text and image data using a common technique. The goal is to overcome the limitations of previous methods that require separate techniques for text and image data analysis, and instead

develop a single, integrated approach that can handle both types of data. By doing so, the proposed approach can provide a more comprehensive and accurate analysis of social media data, which can be valuable for various applications such as sentiment analysis, opinion mining, and content recommendation. In this context the required model is demonstrated in figure 2.

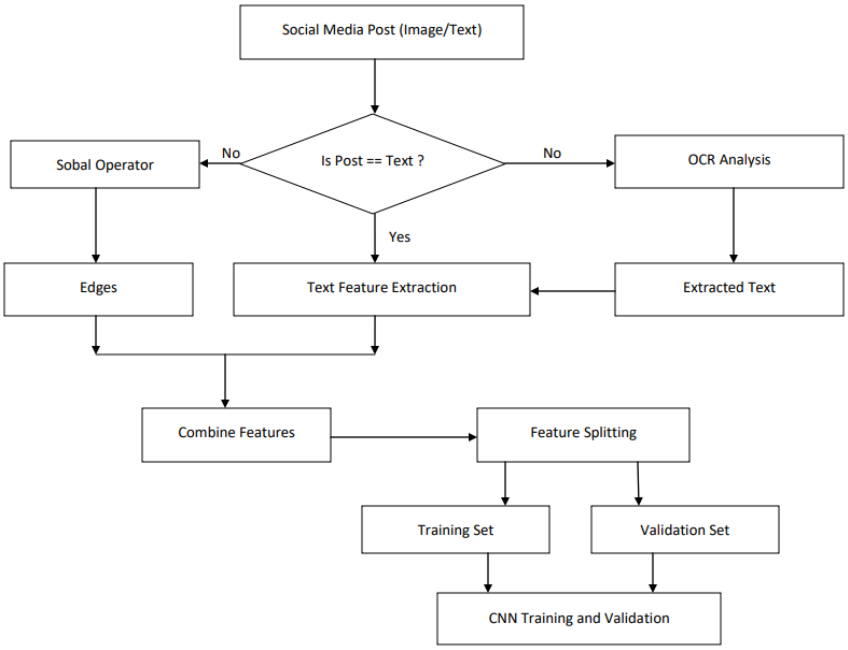


Figure 2 Proposed System for Classifying Social Media Image and Text Post

The model depicted in Figure 2 shows that the first component is the dataset. For this study, two datasets were utilized, both obtained from Kaggle datasets. The first dataset is called "Sentiment analysis of OCR text!!" and contains 239 images sourced from social media [44].. The images are divided into three classes: positive, negative, and random, and contain text, objects, or a combination of both [45]. Examples of the dataset's images can be seen in Figure 3. The second dataset used in this study is intended for entry-level sentiment analysis tasks and also contains three classes: positive, negative, and random. Both datasets were used in the proposed model.

Figure 3 Sample Dataset Images

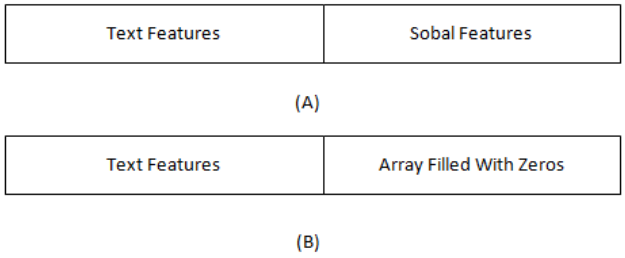
The system checks the posts in the dataset, and if the post contains text, simple text features are extracted. However, if the post contains an image, the system processes the image to extract both the

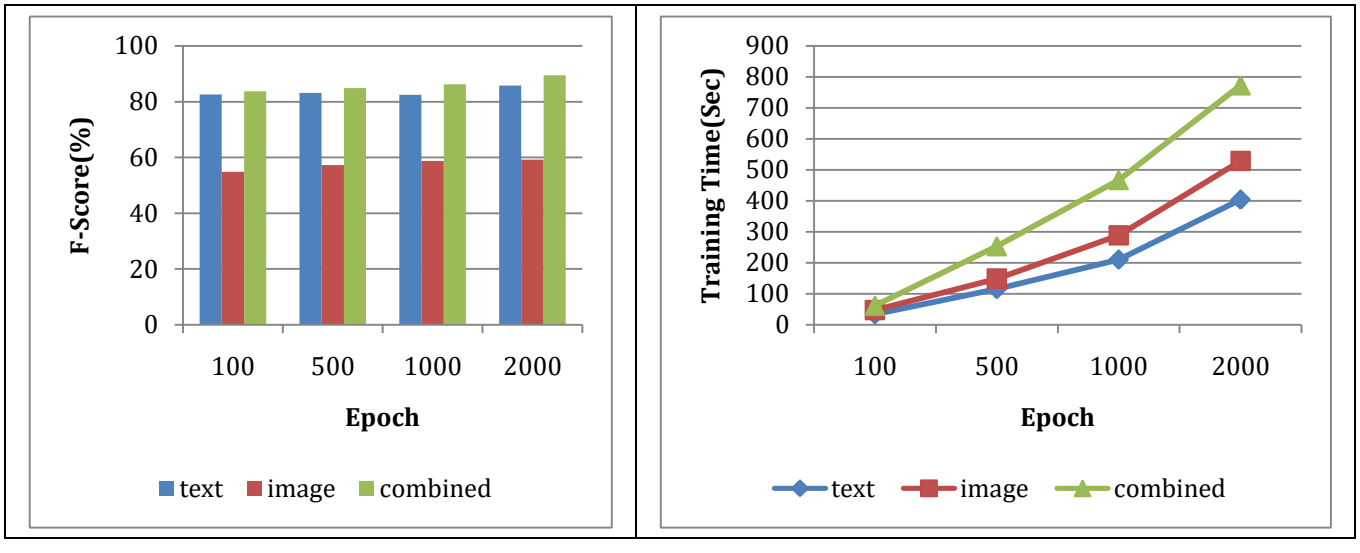
text and edges. To extract text from the images, OCR techniques have been utilized.

The text data is processed using TF-IDF based features and a chi-square test is performed to select essential features from the text. TF-IDF is a common feature extraction technique used in natural language processing to reflect the importance of a word to a document in a collection. It works by assigning a weight to each term in a document based on its frequency within that document and across the collection of documents. The chi-square test is used to determine the statistical significance of the association between each term and the class labels. By performing this test, we can identify the most important terms that are most likely to distinguish between different classes of posts. Now based on feature extracted we have two types of data and three features from the single post, which can be recognized as:

1. **Text post:** features extracted from TF-IDF
2. **Image post:**
   a. OCR based text and TF-IDF features
   b. Edge Feature using sobal Operator

In order to integrate the features extracted from text and image posts, two separate vectors were created. One vector contained the text features extracted using TF-IDF, while the other vector contained the image features, including OCR-based text and edge features using Sobel Operator. These two vectors were then combined to create a single feature vector that incorporated both the text and image features. Figure 4 illustrates an example of how the features were combined to create the final feature vector. The combined feature vector provides a more comprehensive representation of the post, which can improve the accuracy of sentiment analysis and other
tasks.





(A)	(B)
Figure 5 Shows The Performance Of The Proposed Social Media Text Classification Model In Terms Of (A) F-Score And (B) Training Time	

Figure 4 demonstrate the combined features (A) for the image and text features and (B) for the text features

The features extracted from social media posts were combined and used to develop a Convolutional Neural Network (CNN) for classification purposes. The CNN was designed with an input layer, followed by a Convolutional layer with ReLu activation function and a MaxPooling layer. This configuration was repeated with reduced neuron size. Then, a fully connected layer was implemented with two layers, the first using ReLu activation function and the output layer using softmax activation function. The neural network was trained on 75% of randomized samples, and the remaining 25% were used to validate the model. The performance of the model was evaluated and reported in the subsequent section.

V. RESULTS ANALYSIS

In this section, we evaluate the proposed model for classifying text and image social media posts. To perform the evaluation, we have considered the following three experimental scenarios:

1. Type of features: We considered three types of features for the classifier training: only text, only image features, and combined features (text and image).
2. Epoch: We experimented with different numbers of epoch cycles during classifier training.
3. Parameters: We evaluated the performance of the classifiers based on the f-score and training time.

In this section, we will discuss the performance evaluation of the proposed text and image social media post classification model. To evaluate the performance of the model, we have conducted three experimental scenarios.

The performance of the model in terms of f-score is presented in figure 5(A) in the form of a bar graph. The x-axis of the graph represents the number of epoch cycles, while the y-axis represents the f-score in terms of percentage (%). Additionally, figure 5(B) shows the training time of the model in the considered experimental scenarios. The x-axis of the graph contains the epochs, while the y-axis represents the training time in seconds (Sec).

Based on the experimental results, the performance of the text and image-based classification is found to be better than that of using a single feature. The hybridization of text and image features has led to more reliable and accurate classification results. However, the training time of the classification model has increased as compared to using individual features.

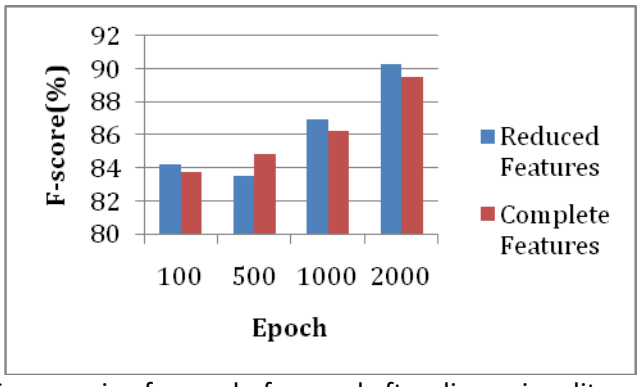


Figure 6 comparing f-score before and after dimensionality reduction

We aimed to address the issue of the long training time of our proposed text and image social media post classification model. To achieve this, we introduced a chi-square test to reduce the dimensions of the combined features. We compared the performance of the classification model with and without the dimensionality reduction technique and observed improvements in both the f-score and

training time. The results are presented in figures 6 and 7.

Figure 6 shows a comparison of the performance of the model with and without the chi-square test. The X-axis represents the number of epochs, while the Y-axis represents the f-score in percentage. The graph indicates that the proposed technique with dimensionality reduction has a better f-score than the technique without dimensionality reduction. Moreover, the proposed technique requires less training time than the previous technique.

Figure 7 demonstrates the comparison of training time for the proposed technique with and without dimensionality reduction. The X-axis represents the number of epochs, while the Y-axis shows the training time in seconds. The graph shows that the training time of the proposed technique with dimensionality reduction is significantly less than the previous technique.

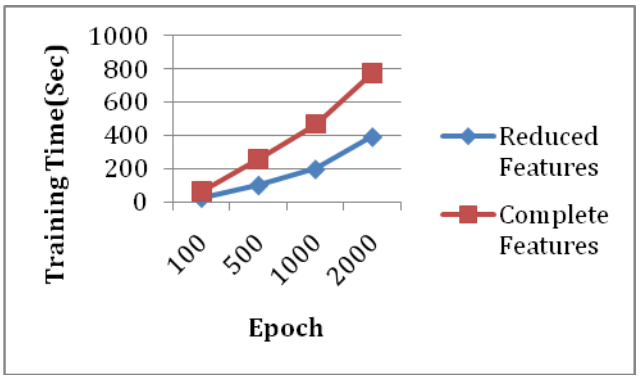


Figure 7 comparing Training Time before and after dimensionality reduction

According to the obtained results the dimensionality reduction can help to improve accuracy as well as training time.

VI. CONCLUSION AND FUTURE WORK

Social media has become an integral part of our daily life and provides various services such as disaster management, marketing, and advertising. Therefore, it is essential to analyze social media data. In this study, we aim to explore various techniques for analyzing social media data, particularly focusing on text and image formats. We conducted a review of existing techniques for analyzing text and images on social media, identifying various classifiers, feature selection techniques, and dataset sources. To understand the impact of feature selection techniques on image and text data, we conducted a comparative performance study on different text and image features and evaluated their impact on classification performance. This study provides insights into the effectiveness of various techniques for analyzing social media data and can aid in developing more accurate and efficient social media data analysis methods.

After identifying suitable techniques, we implemented a novel approach for classifying social media posts containing text and images. Our technique involved extracting TF-IDF-based features and Sobel operators. We combined both feature vectors to train a CNN model configured to classify posts as positive, negative, or random in both text and image formats. However, during experimentation, we found that our model had a lengthy running time. To reduce the dimensionality, we conducted a chi-square test. We evaluated the performance of combined features with and without dimensionality reduction and found that it not only improved the running time of the classifier but also improved classification accuracy. Moreover, our approach enabled the classification of both types of social media posts using a single feature vector set.

REFERENCES

[1] A. Oussous, F. Z. Benjelloun, A. A. Lahcen, S. Belfkih, “Big Data technologies: A survey”, Contents lists available at ScienceDirect Journal of King Saud University – Computer and Information Sciences, 30, 431–448, 2018

[2] S. Ahmad, M. Z. Asghar, F. M. Alotaibi, I. Awan, "Detection and classification of social media-based extremist affiliations using sentiment analysis techniques", Hum. Cent. Comput. Inf. Sci. 9, 24, 2019

[3] H. Zhang, J. Pany, "CASM: A Deep-Learning Approach For Identifying Collectives Action Events with Text and Image Data from Social Media", Sociological Methodology, Vol. 49(1), 1–57 American Sociological Association 2019

[4] X. Bai, B. Shi, C. Zhang, X. Cai, L. Qi, "Text/non-text image classification in the wild with convolutional neural networks", Pattern Recognition, 66, 437–446, 2017

[5] L. Zhou, S. Pan, J. Wang, A. V. Vasilakos, "Machine learning on big data: Opportunities and challenges", Neurocomputing 237 (2017) 350–361

[6] A. F. H. Alharan, H. K. Fatlawi, N. S. Ali, "A cluster-based feature selection method for image texture classification", Indonesian Journal of Electrical Engineering and Computer Science, Vol. 14, No. 3, pp. 1433~1442, June 2019

[7] B. S. Harish, M. B. Revanasiddappa, "A Comprehensive Survey on various Feature Selection Methods to Categorize Text Documents", International Journal of Computer Applications (0975 - 8887), Volume 164 - No.8, April 2017

[8] B. Skrlj, M. Martinc, J. Kralj, N. Lavrac, S. Pollak, "tax2vec: Constructing Interpretable Features from Taxonomies for Short Text Classification", Computer Speech & Language, 65, 101104, 2021

[9] D. T. Nguyen, F. Alam, F. Ofli, M. Imran, "Automatic Image Filtering on Social Networks Using Deep Learning and Perceptual Hashing During Crises", Social Media Studies, Proceedings of the 14th ISCRAM Conference – Albi, France, May 2017

[10] D. T. Nguyen, F. Ofli, M. Imran, P. Mitra, "Damage Assessment from Social Media Imagery Data During Disasters", ASONAM '17, Sydney, Australia, Association for Computing Machinery, July 31 - August 03, 2017

[11] D. Kamin, "Mid-Century Visions, Programmed Affinities: The Enduring Challenges of Image Classification", journal of visual culture, Vol 16(3): 310–336

[12] F. Haider, S. Pollak, P. Albert, S. Luz, "Emotion recognition in low-resource settings: An evaluation of automatic feature selection methods", Computer Speech & Language, 65, 101119, 2021

[13] G. Kou, P. Yang, Y. Peng, F. Xiao, Y. Chen, F. E. Alsaadi, "Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods", Applied Soft Computing Journal, 86, 105836, 2020

[14] J. H. Wang, T. W. Liu, X. Luo, L. Wang, "An LSTM Approach to Short Text Sentiment Classification with Word Embeddings", Conference on Computational Linguistics and Speech Processing ROCLING 2018, pp. 214-223

[15] J. Zhang, Q. Wu, C. Shen, J. Zhang, J. Lu, "Multi-label Image Classification with Regional Latent Semantic Dependencies", arXiv:1612.01082v3 [cs.CV] 12 Mar 2017

[16] L. M. Abualigah, A. T. Khader, "Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering", J Supercomput.

[17] L. M. Abualigah, A. T. Khader, E. S. Hanandeh, "A new feature selection method to improve the document clustering using particle swarm optimization algorithm", Journal of Computational Science, 2017

[18] L. Ma, M. Li, Y. Gao, T. Chen, X. Ma, L. Qu, "A Novel Wrapper Approach for Feature Selection in Object-Based Image Classification Using Polygon-Based Cross-Validation", IEEE Geo-science And Remote Sensing Letters, VOL. 14, NO. 3, Mar. 2017

[19] L. Calza, G. Gagliardi, R. R. Favretti, F. Tamburini, "Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia", Computer Speech & Language, 65, 101113, 2021

[20] L. Vadicamo, F. Carrara, A. Cimino, S. Cresci, F. D. Orletta, F. Falchi, M. Tesconi, "Cross-Media Learning for Image Sentiment Analysis in the Wild", IEEE International Conference on Computer Vision Workshops, 22-29 Oct. 2017

[21] M. Wang, C. Wu, L. Wang, D. Xiang, X. Huang, "A feature selection approach for hyperspectral image based on modified ant lion optimizer", Knowledge-Based Systems, 168, 39–48, 2019

[22] P. Smit, S. Virpioja, M. Kurimo, "Advances in subword-based HMM-DNN speech recognition across languages", Computer Speech & Language 66, 101158, 2021

[23] R. Lin, C. Fu, C. Mao, J. Wei, J. Li, "Academic News Text Classification Model Based on Attention Mechanism and RCNN", Springer Nature Singapore Pte Ltd., Chinese CSCW, CCIS 917, pp. 507–516, 2019.

[24] R. G. Babu, K D. kumar, R. Sharma, R, Krishnamoorthy, "A Survey of Machine Learning Techniques using for Image Classification in Home Security", IOP Conf. Series: Materials Science and Engineering, 1055, 012088, 2021

[25] S. Bahassine, A. Madani, M. Al-Sarem, M. Kissi, "Feature selection using an improved Chi-square for Arabic text classification", Journal of King Saud University – Computer and Information Sciences, 32, 225–231, 2020

[26] S. V. Georgakopoulos, S. K. Tasoulis, A. G. Vrahatis, V. P. Plagianakos, "Convolutional Neural Networks for Toxic Comment Classification", arXiv:1802.09957v1 [cs.CL] 27 Feb 2018

[27] T. Toivonen, V. Heikinheimo, C. Fink, A. Hausmann, T. Hiippala, O. Järv, H. Tenkanen, E. D. Minin, "Social media data for conservation science: A methodological overview", Biological Conservation, 233, 298–315, 2019

[28] V. P. Fralenko, R. E. Suvorov and I. A. Tikhomirov, "Automatic Image Classification for Web Content Filtering: New Dataset Evaluation", Recent Developments and the New Direction in Soft-Computing Foundations and Applications, Studies in Fuzziness and Soft Computing 361

[29] Y. Pathak, K. V. Arya, S. Tiwari, "Feature selection for image steganalysis using levy flight-based grey wolf optimization", Multimed Tools Appl, Springer Science+Business Media, LLC, part of Springer Nature, 6155-6, 2018

[30] Y. Han, H. Lee, "A Deep Learning Approach For Brand Store Image And Positioning", Anthropocene, Proceedings of the 25th International Conference of the Association for Computer-Aided, Architectural Design Research in Asia, Volume 2, 689-696, 2020

[31] P. Tao, Z. Sun, Z. Sun, "An Improved Intrusion Detection Algorithm Based on GA and SVM", VOLUME 6, IEEE Access, 2018

[32] A. K. Ojo, T. O. Idowu, "Improved Model for Facial Expression Classification for Fear and Sadness Using Local Binary Pattern Histogram", Journal of Advances in Mathematics and Computer Science 35(5): 22-33, Article no.JAMCS.59130, 2020

[33] H. Fang, D. Zhang, Y. Shu, G. Guo, "Deep Learning for Sequential Recommendation: Algorithms, Influential Factors, and Evaluations", ACM Transactions on Information Systems, Vol. 1, No. 1, Article 1. Publication date: January 2020

[34] A. F. Al-daour, M. O. Al-shawwa, S. S. Abu-Naser, "Banana Classification Using Deep Learning", International Journal of Academic Information Systems Research, Vol. 3 Issue 12, Pages: 6-11, Dec. – 2019

[35] https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-

neural-networks-the-eli5-way-3bd2b1164a53

[36] M. N. Fekri; H. Patel; K. Grolinger; V. Sharma, "Deep learning for load forecasting with smart meter data: Online Adaptive Recurrent Neural Network", Electrical and Computer Engineering Publications. 181, 2020

[37] R. N. Waykole, A. D. Thakare, "A Review of Feature Extraction Methods for Text Classification", International Journal of Advance Engineering and Research Development Volume 5, Issue 04, April - 2018

[38] P. Bafna, D. Pramod, A. Vaidya, "Document Clustering: TF-IDF approach", International Conference on Electrical, Electronics, and Optimization Techniques, IEEE, 2016

[39] T. Kenter, M. de Rijke, "Short Text Similarity with Word Embeddings", CIKM'15, Melbourne, Australia, ACM, Oct. 19–23, 2015,

[40] B S. Kumar, E B. Varma, "A Different Type of Feature Selection Methods for Text Categorization on Imbalanced Data", International Journal of Advanced Research in Computer and Communication Engineering, ISO 3297:2007 Certified, Vol. 5, Issue 9, September 2016

[41] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, H. Liu, "Feature Selection: A Data Perspective", ACM Computing Surveys, Vol. 9, No. 4, Article 39, Publication date: March 2010

[42] S. Sasikala, S. Appavu alias Balamurugan, S. Geetha, "Multi Filtration Feature Selection (MFFS) toimprove discriminatory ability in clinical data set", Applied Computing and Informatics, 12, 117–127, 2016

[43] Mengle, S. S. R., & Goharian, N., "Ambiguity measure feature-selection algorithm". Journal of the American Society for Information Science and Technology, 60, 1037– 1050, 2009

[44] https://www.kaggle.com/datasets/somnath796/sentiment-analysis-of-ocr-text

[45] https://www.kaggle.com/datasets/jp797498e/twitter-entity-sentiment-analysis