iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** V **Month of publication:** May 2025

DOI: https://doi.org/10.22214/ijraset.2025.71605

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

Personality Traits Prediction Using DISC

Dhruv[1], Mr. Jamkhongam Touthang[2], Dikshant Jajoriya[3], Harsh Umak[4]
Department of Applied Mathematics Delhi Technological University New Delhi, India

Abstract: ***This research introduces a novel method of optimizing the process of cultural fit assessment in corporate recruitment through the combination of natural language processing (NLP) methods and semi-supervised self-training models. Using an unlabeled dataset drawn from tweets, we used preprocessing techniques and word2vec embeddings to determine semantic relationships in the text. K-means clustering was used to find the optimal number of clusters (k=4). In cooperation with psychology professionals, we marked 10% of the dataset based on the DISC model of psychology to facilitate model training. A mixed model integrating Convolutional Neural Networks (CNN) and Long Short-Term Memory networks (LSTM) was created to properly leverage labeled and unlabeled data via semi-supervised learning. This study will simplify cultural fit evaluation during recruitment, making it possible to quickly and accurately evaluate candidates.***

Index Terms: ***Natural Language Processing, K-means cluster- ing, DISC Psychology Model, Convolutional Neural Networks, Long Term Short Memory, Semi-Supervised Learning, Twitter.***

I. INTRODUCTION

The contemporary recruitment process is beset with tremendous challenges owing to the growing number of job applications, which introduce inefficiencies into the conventional cultural fit evaluations made by HR experts. As application figures grow, the constraints of face-to-face interviews become more glaring, leading to hiring bottlenecks. Though past research into personality assessment via social media information has been significantly advanced, it has certain shortcomings. For example, in the paper "Personality Prediction from Twitter and Instagram," feature extraction was presented but depended on random forest regression that did not have the depth needed for extensive personality analysis. Likewise, in "Personality Classification Through Twitter Data Analysis," high accuracy was shown but to the Big Five personality traits alone, limiting its application. Conversely, this research suggests a more complex and holistic approach in resolving these issues. Our study utilizes semi-supervised learning to automate personality evaluations from social media behavior with the DISC psychology model. By embracing cutting-edge methodologies, this method updates the hiring process, leading to more precise and effective candidate screening. In particular, we suggest using machine learning methods to analyze candidates' personality characteristics from their social media usage, providing a real-world solution to the issues of modern recruitment. By using advanced algorithms, this technique has the potential to transform the recruitment process by simplifying the identification of good candidates from large pools of applicants. The research started with the procurement of an unlabeled dataset of tweets, which was preprocessed to improve data quality. By applying Natural Language Processing (NLP) methods, including word2vec, the text data were transformed into numerical vectors in order to extract semantic similarities. K-means clustering was subsequently used to find k=4 as the best number of clusters in the dataset. In order to enhance model precision, psychology professionals labeled 10% of the dataset with the DISC psychological model. This model, created by psychologist William Moulton Marston, categorizes personality traits into four primary dimensions: Dominance, Influence, Steadiness, and Conscientiousness. These dimensions characterize different behavior tendencies and communication patterns and form a solid analytical framework for human behavior. We applied our labeled data to build on this by using a semi-supervised self-training model with a hybrid structure that incorporates Convolutional Neural Networks (CNN) and Long Short-Term Memory networks (LSTM). This hybrid structure successfully leveraged labeled and unlabeled data to create highly robust and accurate personality assessment framework.

Our main aim was to make the process of cultural fit evaluation in corporate recruitment more streamlined, so candidate selection would become easier and more efficient. Our approach is presented here as a more sophisticated substitute for conventional cultural fit tests, which offer a more precise and thorough assessment of personality traits of candidates.

II. RELATED WORKS

A great number of different methods, every one of which has its own strategy and contribution, have emerged in recent years to address this issue. We will begin discussing cutting-edge approaches.

A. Fusing Social Media Cues

The research "Personality Prediction from Twitter and Instagram" investigates the challenging process of obtaining varied features from social media users in the US, such as image features, language patterns, and user meta-data. Using random forest regression, the research seeks to forecast people's personality traits from these multi-faceted data points. This pioneering method redefines conventional personality testing by tapping into the vast behavioral data found on social media. By bringing together heterogeneous data streams and employing state-of-the-art regression methodologies, the research discovers latent relationships among online behaviors and psychological traits. These findings not only contribute deeper insights into human behavior in cyber spaces but also offer practical value in customized content delivery, precise marketing, and psychological profiling. [2]

B. Naive Bayes and KNN based Classification

The research "Personality Classification Through Twitter Data Analysis" focuses primarily on the Big Five personality traits. Utilizing Multinomial Naive Bayes and K-Nearest Neighbors algorithms, the research attains a remarkable 65% precision in personality trait prediction, proving the efficacy of machine learning methods in social media text analysis. This research emphasizes the potential for extracting useful information from user-generated content from websites like Twitter and indicates possibilities for enhancing personality prediction models.The research shows how social media could be an important source of behavioral data, with important implications for identifying individual tendencies and preferences. Aside from its technical success, the study adds to the emerging discipline of computational psychology, allowing for a more in-depth investigation of human personality through the online digital traces people leave behind. This method fills the gap between conventional psychological tests and contemporary data-driven methods, opening the door to novel applications in personality analysis.

C. Semi Supervised Learning-based Sentiment Analysis

This paper examines personality detection by probing text data taken from online social networks using the Five Factor Model as its theoretical basis. It proposes a hybrid model made up of Convolutional Neural Networks (CNN) to efficiently extract features and Recurrent Neural Networks (RNN) to capture long dependencies in text data.The proposed hybrid CNN-RNN model outperforms the conventional method in accuracy and performance measures, and its suitability for personality detection in any field is proven. Through the power of CNNs to distill out useful features from text and RNNs' prowess in capturing temporal relationships, the architecture presents a more comprehensive and solid evaluation of people's personality characteristics.This break-through not only enhances the accuracy of personality identification but also brings out the greater potential of deep learning methods in psychological profiling and behavioral analysis from social network text. In general, the study constitutes a valuable addition to the enhancement of personality identification methods using novel neural network designs.

D. Research Articles Categorization via NLP

This research explores the classification of scientific texts through the application of state-of-the-art Natural Language Processing (NLP) methods and pre-trained language models. Through the use of SciBERT, a specialized language model specifically trained for scientific texts, coupled with the K-Means clustering algorithm, the research obtains enhanced text classification in academic literature. The results illustrate the efficiency of combining specialized language models with clustering methods to make research literature more organized and accessible. Proper text classification not only simplifies navigation and recommendation systems across scholarly fields but also allows for more accurate and customized research investigation. Such advancements make knowledge sharing efficient and support targeted academic discovery.On the whole, this study emphasizes the role of NLP methods and pre-trained language models to transform text classification in scientific publications into a highly useful tool for enhancing academic studies and encouraging innovation in knowledge management systems.

E. Hybrid CNN-RNN Model for Personality Detection

This research explores personality detection through the analysis of social network texts, using the Five Factor Model as its theoretical framework. It presents a hybrid model that integrates Convolutional Neural Networks (CNN) for efficient feature extraction and Recurrent Neural Networks (RNN) for handling long-term dependencies in text data. The hybrid CNN-RNN model outperforms baseline methods in terms of accuracy and performance measures, proving its effectiveness for personality detection in different fields. By leveraging CNNs' power to derive useful features from text and RNNs' capability to grasp temporal relationships, this structure offers a richer and more stringent evaluation of people's personality traits.

This breakthrough not only enhances the accuracy of personality identification but also identifies the wider possibilities of deep learning methods in psychological profiling and behavioral analysis from social network texts. In total, the study is a major contribution to personality detection methods using state-of-the-art neural network architectures.

III. METHODOLOGY

In the methodology section, we will discuss in detail the steps and procedures followed to conduct this study.

A. Dataset and Data Preprocessing

In the process of data preprocessing, we routinely sanitized the raw data of the tweets to prepare them for analysis. First, punctuation, URLs, and shortened words were deleted for clarity and easy readability. Finally, text standardization was accomplished by transforming everything to lowercases for consistency throughout. To concentrate on substantive content, stopwords like "and," "the," and "is" were removed; for instance, "The sky is blue" was shortened to "sky blue." Lemmatization [7] was also used to minimize words to their root forms to enhance coherence and consistency—e.g., "running" was converted into "run." These preprocessing procedures played a key role in preparing the dataset for proper analysis of personality traits and cultural fit. The Sentiment140 dataset utilized in this research consists of 1,600,000 tweets accessed using the Twitter API, presenting a solid ground for further examination.

B. Model Architecture

With Natural Language Processing (NLP) [8] methods, we processed a vast, unlabeled dataset obtained from social media, prioritizing word embeddings to deepen semantic meaning. From many possibilities, Word2Vec proved to be the best choice, infusing the dataset with a meaningful level of contextual insight. To meet the challenge of working with unlabeled data, we used K-Means clustering, training the model on 10% of the entire dataset. This sub-sample provided a basis to work with the psychologists, who carefully labeled the clusters with a high degree of precision. We introduced a combined model that joined Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN) together in a semi-supervised setup. This design applied iterative self-training, and this was a shift in paradigms to our methodology. Dynamically embedding feedback from prior iterations, by selection criteria, our model learned to improve predictively incrementally. The incorporation of Long Short-Term Memory (LSTM) networks and CNN as supervised modules further enhanced the model's ability for sophisticated analysis, providing a strong framework for personality testing and cultural fit evaluation.

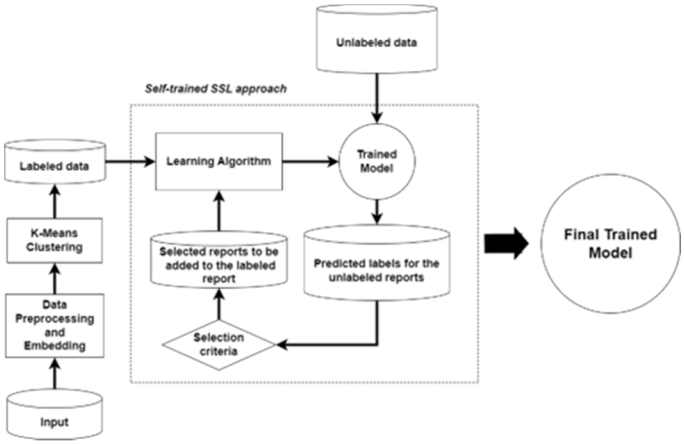


Fig. 1. Model Architecture.

a) *Embedding Layer* : Word2Vec [9], a basic NLP algorithm, maps words into dense vectors preserving subtle semantic patterns. By employing two architectures—CBOW (Continuous Bag of Words) and Skip-gram—it learns context-sensitive representations greatly enhancing language understanding in machine learning tasks. An interesting application of Word2Vec's proficiency is the analogy "king – man + woman," which returns a vector close to "queen," showing its aptness for comprehending semantic associations. This robust representation facilitates applications such as sentiment analysis, in which contextually rich subtleties are vital, to achieve more accurate and insightful language processing results.

b) *Labelling Layer:* K-Means, a basic clustering algorithm, partitions data into different groups by trying to minimize the sum of squared distances between the data points and their cluster centers. Celebrated for its scalability and simplicity, K-Means is a cornerstone of unsupervised machine learning, which is extensively used in pattern recognition and data discovery in various disciplines. To further improve efficiency, we reduced the high-dimensional vectors of Word2Vec using dimensionality reduction algorithms like PCA and t-SNE to achieve enhanced performance and readability. Then, 10% of the data was subjected to K-Means clustering. The Elbow Method was used to identify the best number of clusters (k), with the "elbow" point indicating a point of decreasing return in reduction of distortion.In our analysis, k=4 was found to be the optimal value, showing a sharp fall in distortion after this point. This validated the use of four clusters for our data, increasing the accuracy and interpretability of our K-Means analysis.

c) *Semi-supervised Learning:* Semi-supervised learning [6] integrates labeled and unlabeled data for model training, minimizing reliance on large sets of labeled data by taking advantage of a small quantity of labeled data and a large set of unlabeled data to improve accuracy and robustness. In this research, Convolutional Neural Networks (CNNs) [10] successfully detect local structures in sequential text data, including word arrangements and syntax, using filters and pooling layers for reducing dimensions. augmented by Recurrent Neural Networks (RNNs) [11], specifically Long Short-Term Memory (LSTM) networks [12], which perform well with sequential data through the handling of temporal dependencies and context, critical in persona classification and behavior prediction. In combination, CNNs identify local features and RNNs focus on sequential relationships, enhancing the ability of the model to analyze user behavior and classify personas. The model architecture incorporates a Conv1D layer for extracting features, a MaxPooling1D layer for reducing dimensions, an LSTM layer for identifying temporal dependencies, and dense layers for transforming features and classifying. Training uses labeled data initially, and then later supervised learning with pseudo-labelling confident predictions of the unlabeled data and they are fed back into further iterations. These iterative cycles refine the predictions and generalize better. The model is trained with the Adam optimizer and categorical cross-entropy loss function and has high accuracy in persona classification and behavior prediction on the Twitter data.

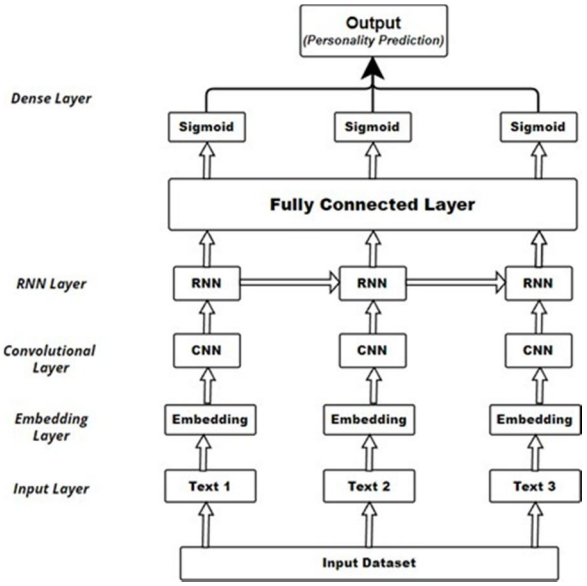


Fig. 2. Semi-Supervised Model Architecture

C. Metrics

In assessing K-Means clustering, three pivotal measures enhance the analysis of clustering performance: the Elbow Method, Silhouette Score, and Davies-Bouldin Method. The Elbow Method determines the ideal number of clusters by locating the "elbow" point, which strikes a balance between variance capture and avoiding overfitting. The Silhouette Score assesses cluster quality based on cohesion and separation, giving insights into cluster distinguishability. The Davies-Bouldin method provides a full evaluation by striking a balance between intra-cluster tightness and inter-cluster distance. Further, categorical cross-entropy loss function plays a central part in multi-class classification problems as it computes the discrepancy between class probability predictions and ground truth labels. In our project, the loss function measures the difference between predicted probabilities of personality traits and ground truth labels to allow the model to fine-tune its parameters for effective classification results.

a) *Elbow Method:* The elbow method is a visualization technique used for finding the right number of clusters (k) in K-Means clustering. It includes graphing the Within-Cluster Sum of Squares (WSS) with different values of k and locating the "elbow" point where the decline in WSS significantly decreases. This is an equilibrium between the reduction of variance within the cluster and preventing overfitting. Through giving a visual guideline for choosing k, this technique helps in making effective and well-informed decisions about clustering in various applications. The equation to calculate the WSS for finding the optimal number of clusters (k) is:

$$\text{WSS} = \sum_{i=1}^{k} \sum_{j=1}^{n} d(x_j, c_i)^2 \qquad (1)$$

where:
WSS is Within-Cluster Sum of Squares,
k is the number of clusters in K-Means clustering, n is the total number of data points,
$d(x_j,c_i)$ represents distance between the data point x_j and centroid c_i

b) *Silhouette Absolute Score:* A measure known as the Silhouette Score determines the cohesiveness and the distance of a cluster. It measures how well-separated clusters are, with arange of -1 to 1. More clearly defined clusters are signaled by a higher score. The formula used to compute the score is:

$$S = \frac{b - a}{\max(a, b)} \qquad (2)$$

where:

S represents the Silhouette Score, a measure of cluster quality,
a average distance from a data point to other points within same cluster,
b average distance from a data point to the nearest neighboring cluster.

c) *Davies-Bouldin Method:* Davies-Bouldin Index is an index used to measure the quality of clustering in terms of both compactness (how close together the points in a cluster are) and separation (how far apart each cluster is from the others). Lower values of Davies-Bouldin Index mean better clustering because it implies well-separated clusters with tight cohesion. The formula for Davies-Bouldin Index is:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right) \qquad (3)$$

where:
DB is the Davies-Bouldin index, k is the number of clusters,
i iterates over each cluster,
j iterates over each cluster s.t. j ,
σ_i and σ_j are the average distance within clusters i and j,
$d(c_i, c_j)$ distance between the centroids of clusters i and j

d) *Categorical Cross-entropy:* A typical loss function in multiclass classification issues is categorical cross-entropy. It measures how different the actual class distributions are from the expected probability. When working with several classes and categorical data, this function is especially useful.The Categorical Cross-entropy Method formula is given by:

$$L(y, \hat{y}) = -\sum_{i} y_i \cdot \log(\hat{y}_i) \qquad (4)$$

where:
L is categorical cross entropy loss,
y denotes the true probability of class i ,
ŷ denotes the predicted probability of class y,

IV. RESULTS

Our experiment proves the efficiency of the model in predicting candidate personality. With visualizations and comparative analysis, we have better insights into the performance of the model, indicating its strengths and weaknesses. These visualization tools present a clear view of how accurately the model is capturing and predicting personality traits, allowing for better assessment of its accuracy and reliability.

A. Results & Observations: K-Means

As demonstrated in the findings in Fig. 4, we effectively identified 4 clusters, as we had anticipated using the k=4 obtained via the Elbow Method. The Silhouette Absolute Score from the K-Means clustering analysis was 0.25335, reflecting moderate separation and cohesion between the groups. Additionally, the Davies-Bouldin Method, which takes into account both compactness and separation, confirmed the clustering quality with a score of 1.2404. The Elbow Method, as seen in Fig. 3, helped us determine that k=4 was the optimal number of clusters, providing a solid foundation for understanding the dataset. This careful and systematic approach ensures a comprehensive and accurate interpretation of the data.

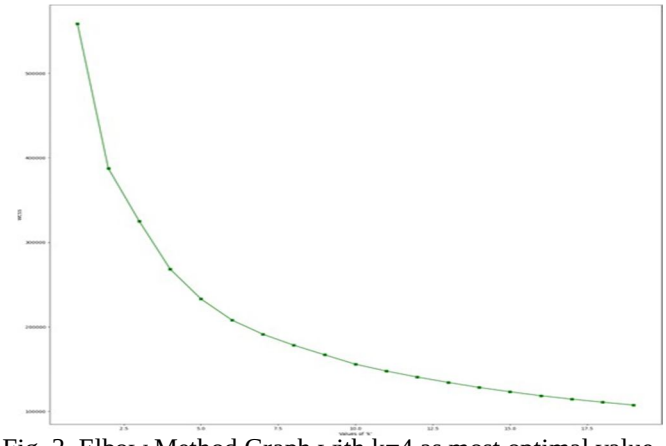


Fig. 3. Elbow Method Graph with k=4 as most optimal value

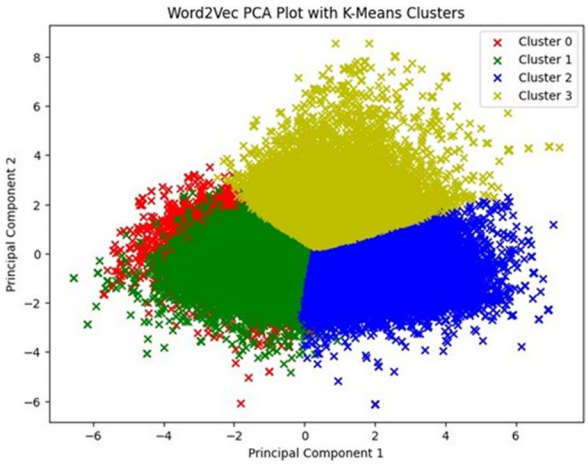


Fig. 4. Word2Vec PCA Plot with K-Means Clustering showing 4 clusters K-M

Metrics	*Value*
Silhouette Absolute Score	0.25335
Davies-Bouldin Method	1.2404

TABLE I
eans Clustering Evaluation Results

The comparison of training and testing accuracy emphasizes the improved performance of the hybrid CNN & RNN model over the standalone RNN model. The hybrid model outperformed the RNN model in both testing (97.51% vs. 91.26%) and training (99.36% vs. 99.13%). Both models employed the Adam optimizer and the categorical cross-entropy loss function. These findings highlight the efficacy of the hybrid architecture in the case of the provided Twitter dataset, showing that the integration of CNN and RNN components improves accuracy and holds great promise for more accurate personality prediction.

The performance assessment of the hybrid RNN-CNN model, based on accuracy and categorical

TABLE II
RNN V/S HYBRID CNN & RNN ACCURACY RESULTS AT EPOCH=20

Model	Training Accuracy	Testing Accuracy
RNN	99.13%	91.26%
Hybrid CNN & RNN	99.36%	97.51%

cross-entropy loss metrics, brought out some interesting points. Throughout epochs 1 to 15, the model convergence was changing, as shown by varying loss values. Accuracy showed a non-saturating behavior, with oscillations and a maximum at epoch 11, as seen in Fig. 5. Furthermore, the difference in categorical cross-entropy loss values between epochs is illustrated in the Loss vs. Epochs plot presented in Fig. 6, providing insights into the learning dynamics of the model during training.

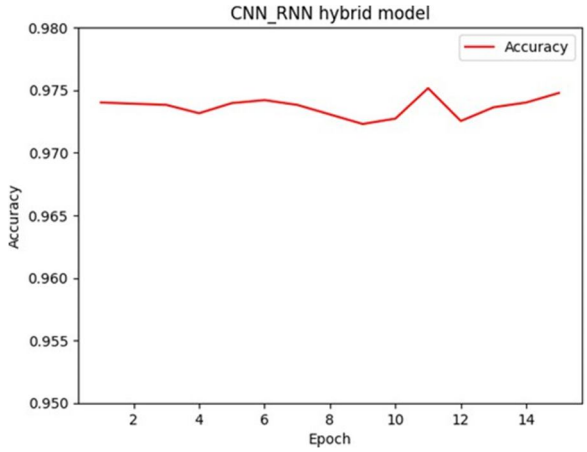


Fig. 5. Graph showing Accuracy v/s Epochs with Adam Optimizer

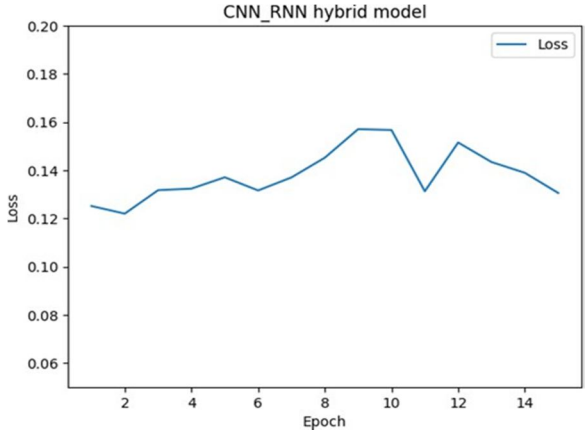


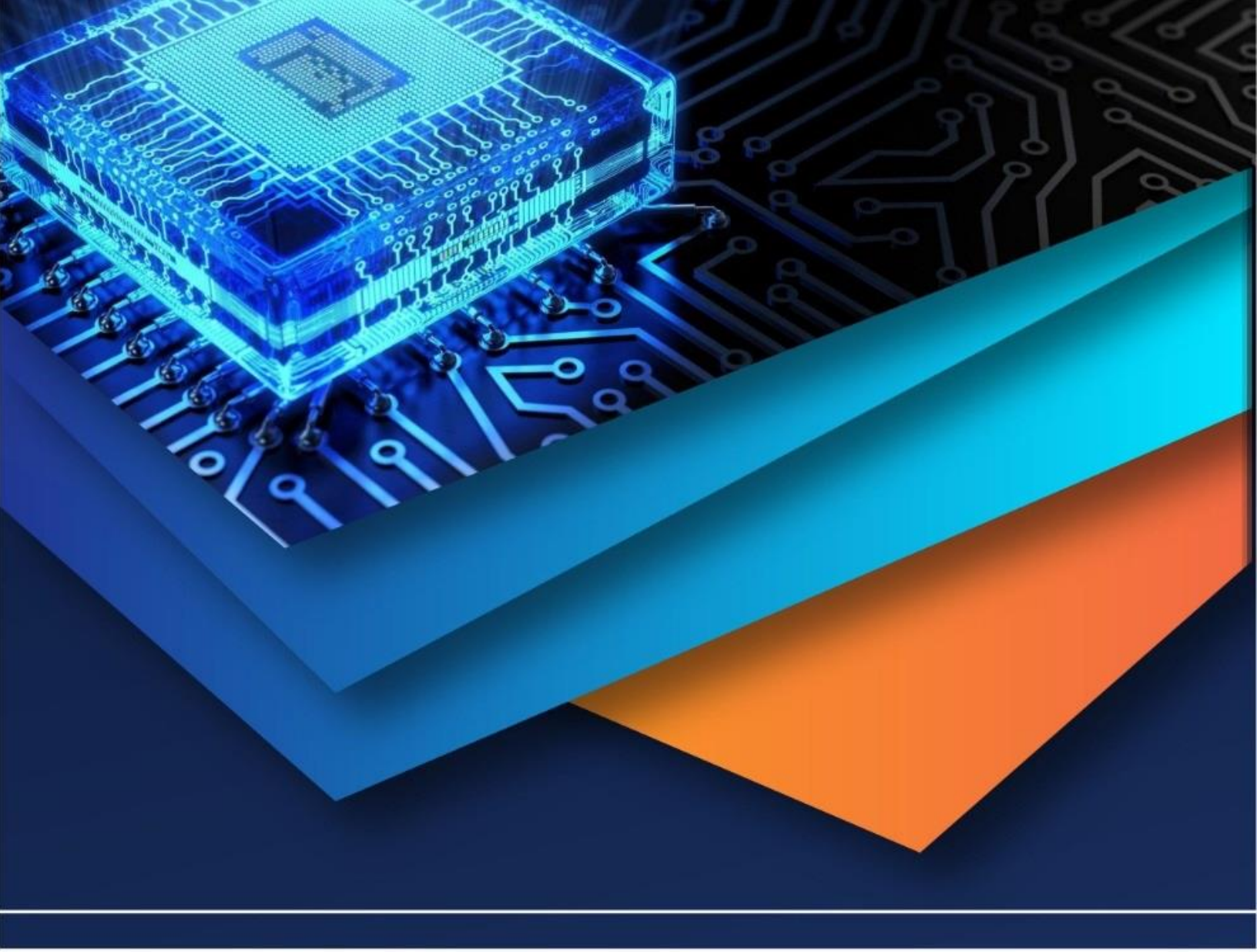
Fig. 6. Graph showing Loss v/s Epochs with Categorical Cross Entropy

V. CONCLUSION

This research illustrates the seamless integration of computational rigor, language analysis, psychological understanding, and sophisticated methodologies. From the early clustering with K-means and NLP to building a sound hybrid model integrating CNNs and RNNs, every phase was carefully designed to reveal the complex patterns of Twitter user behavior. The results highlight the revolutionary potential of multidisciplinary methods in massive-scale social media analysis, opening doors to future breakthroughs in understanding and forecasting online personalities. Future studies can investigate the seamless fusion of text and image data, crossing platform boundaries to record users' holistic online trace. Multimodal neural networks can potentially improve prediction accuracy and sentiment analysis. In addition, placing emphasis on privacy-preserving methods and adaptable fusion mechanisms will be crucial for realizing accurate model applicability in various digital environments.

REFERENCES

[1] Ema Utami, Anggit Dwi Hartanto, Sumarni Adi, Irwan Oyong, and Suwanto Raharjo. Profiling analysis of disc personality traits based on twitter posts in bahasa indonesia. Journal of King Saud University Computer and Information Sciences, 34(2):264–269, 2022.
[2] M.Skowron, M. Tkalc˘ic˘, and B. Ferwerda. Fusing social media cues: Personality prediction from twitter and instagram. 2016.
[3] Bayu Yudha Pratama and Riyanarto Sarno. Personality classification based on twitter text using naive bayes, knn and svm. In 2015 International Conference on Data and Software Engineering (ICoDSE), pages 170–174, 2015. 37
[4] Yuhao Pan, Zhiqun Chen, Yoshimi Suzuki, Fumiyo Fukumoto, and Hiromitsu Nishizaki. Sentiment analysis using semi supervised learning with few labeled data. In 2020 International Conference on Cyberworlds (CW), pages 231 234, 2020.
[5] Rosanna Turrisi. Beyond original research articles categorization via nlp. In Workshop on Human-in-the-Loop Applied Machine Learning (HITLAML), September 04-06, 2023-Belval, Luxembourg, 2023.
[6] Dr. Jayasudha J.S Princy Sathyadas. Hybrid cnn-rnn model for per sonality detection. International Journal for Research in Engineering Application Management (IJREAM), 6:37–44, 2020.
[7] Stanford University. Stemming and lemmatization- stanford nlp. https://nlp.stanford.edu/IR book/html/htmledition/stemming-and lemmatization1.html. [Online]. Accessed on: August 27, 2023.
[8] IBM. Natural language processing at ibm. IBM. Accessed: August 5, 2023.
[9] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies, pages 142–150, 2011.
[10] Maite Gim´enez , Roberto Paredes, and Paolo Rosso, Personality Recognition Using Convolutional Neural Networks, Springer Nature Switzerland, CICLing 2017, LNCS 10762, pp. 313–323, 2018.
[11] Rachma Indira and Warih Maharani. Personality detection on social media twitter using long short-term memory with word2vec. In 2021 IEEE International Conference on Communication, Networks and Satel lite (COMNETSAT), pages 64–69, 2021.
[12] Hochreiter, S., & Schmidhuber, J. Long short term memory. Neural computation,9(8), 17351780, 1997



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY