CHAPTER 1

INTRODUCTION

This chapter includes the following subtopics, namely: (1) Rationale; (2) statement of the problem; (3) Hypothesis; (4) Assumption; (5) Scope and Delimitation; and (6) Significance of the Study.

1.1 Rationale

In the digital era, social media has become an integral aspect of human life, facilitating the active sharing of thoughts, feelings, and daily activities among individuals. This data presents an opportunity for the automated analysis of individual personalities through artificial intelligence technology. The Big Five Personality Traits model is a widely recognized framework for analyzing personality [1] The Big Five Personality is classified into five different categories: Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (OCEAN). The Openness to Experience dimension characterizes individuals who exhibit a propensity for engaging in novel activities. The contrasting characteristic to this dimension is individuals who experience anxiety when confronted with new challenges [2][3]. Conscientiousness represents a trait characterized by a heightened level of alertness. These individuals exhibit a disciplined and reliable disposition. Extraversion is a personality dimension characterized by a preference for and comfort in engaging with others. The Agreeableness dimension is characterized by a tendency to avoid conflict [2][3]. Neuroticism refers to individuals who can manage their emotions, including stress and pressure [3]. Analyzing user personality traits via text analysis is classified as an information classification or processing task [4].

Conventional methods for personality detection, including individual machine learning algorithms, exhibit constraints regarding accuracy and efficiency. Over the last ten years, deep learning neural networks have gained prominence as effective instruments for analyzing complex data, including social media texts. Techniques including Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Bidirectional LSTM (BiLSTM) provide various methodologies for information extraction from data with varying structures. Nonetheless, single deep learning methods frequently struggle to adequately represent the intricate relationships present in personality data. Hybrid approaches have been developed to address these limitations by integrating the strengths of multiple deep learning models to enhance performance.

Recent research on the use of CNN and RNN [5] indicates an accuracy of 90 and a f1-score of 51. This work identifies an issue wherein the system fails to capture information in lengthy sentences, thus proposing a method capable of extracting information features

to recognize extended sequences. The RNN approach is constrained in its ability to retain long-distance information, but LSTM and BiLSTM are capable of capturing extensive informational properties.

Deep learning neural networks exhibit varied methodologies for data processing and demonstrate differing levels of performance. Long Short Term Memory (LSTM) and Bidirectional Long Short Term Memory (BiLSTM) are enhancements on the Recurrent Neural Network approach [3]. The input layer differentiates LSTM from BiLSTM, as LSTM retrieves information alone in a forward direction, while BiLSTM retrieves information in both forward and backward orientations. BiLSTM necessitates the capacity to extract local feature information through patterns and characteristics of the utilized data, such as word embeddings, and to collect long-term information, which can be augmented by a hybrid approach using neural networks in the form of CNN.

This research project selected the combination of CNN and BiLSTM to augment CNN's capacity for spatial feature extraction alongside BiLSTM's skill in comprehending the temporal context of the data. The present research examines the efficacy of the CNN-BiLSTM hybrid model in identifying personality traits based on Big Five data, in comparison to the standalone CNN and BiLSTM techniques. The motivation was to enhance the system to effectively capture elements of relatively lengthy sentences, with the expectation that the accuracy of both machine learning and deep learning methodologies would improve, particularly in the realm of personality analysis. A hybrid methodology is expected to address the shortcomings of each model by amalgamating the advantages of CNN and BiLSTM. This research employs the Big Five personality model, providing a robust psychological framework for delineating personality and enabling additional investigation in this study.

The rest of this paper will be analyzed through multiple subsections. Chapter 2 presents an overview of the literature review related to pertinent studies and the current research landscape in the field. Chapter 3 will outline the algorithm and the integration of the two methods, commencing with data preprocessing. Chapter 4 presents the results and discussion, outlining the findings of the test outcomes. Chapter 5 outlines the conclusions derived from the research findings.

1.2 Statement of the Problem

Deep learning and machine learning are often linked to research aimed at personality identification via social media. Numerous investigations on personality recognition implementations have encountered the challenge of the system's inability to obtain word information in lengthy sequences.The length of a comment varies, and we cannot regulate its brevity or extent. Consequently, sentences are categorized into two types: those with fewer than 30 words are classified as short sentences, while those with more are deemed long sentences. In this circumstance, the BiLSTM technique exhibits greater potential due to its capacity

to retain information bidirectionally; yet, to enhance the extraction of local features, it necessitates support from CNNs.

1.3 Objective and Hypotheses

This research seeks to create a system that can capture remote word information regarding sentence patterns of different durations. It is anticipated that assessment metrics, such as accuracy and precision, will enhance in the realm of personality recognition through a Hybrid CNN + BiLSTM methodology.

The length or complexity of sentences influences the model's efficacy in identifying text-based personality traits. CNN-based models are more adept at identifying local patterns in short texts, but BiLSTM excels at capturing long-range dependencies in lengthy texts. Consequently, the integration of Hybrid CNN and BiLSTM is presumed to more effectively address the limitations inherent in each approach, with CNN managing spatial elements in text and BiLSTM ideally capturing long-term word associations.

The Hybrid CNN + BiLSTM technique is anticipated to outperform individual methods (Single CNN or Single BiLSTM) in the detection of text-based personality, as measured by accuracy and precision evaluations. This is predicated on the premise that the hybrid model can more efficiently integrate the strengths of CNN in identifying short-range word patterns with the benefits of BiLSTM in comprehending extended sentence structures bidirectionally.

1.4 Assumption

This study assumes that the datasets possess equivalent properties across labels, avoiding both excessive divergence and excessive similarity. The dataset is structured chronologically to enable the verification of patterns over an extended sequence. Furthermore, the dataset comprises user comments that have not undergone a pre-processing phase, which may influence the analysis outcomes.

The sentence structures derived from user tweets are presumed to reflect individual personalities according to the Big Five Personality Traits. Sequential patterns in textual data possess temporal or contextual linkages that the model might leverage to enhance classification outcomes. The model is presumed to perform effectively on training and test data with analogous properties, hence allowing for the generalization of experimental results.

Moreover, it is presumed that the annotation process or data labeling relies on user comment data and pertains to the Big Five inventory questionnaire.

This study posits that external elements, such alterations in language style, social setting, or emotions while composing tweets, do not affect individuals' personalities in the

short term. The created approach exclusively analyzes text-based data and does not use multimodal aspects such as photos, videos, or sounds in its assessment of personality.

1.5 Scope and Delimitation

This research possesses restrictions that may hinder its execution as originally intended. The problem's constraints arise from the dataset being composed of tweets, which lack a time series that may facilitate the management of extended sequences. The dataset has significant overlap and lacks balance. Under some conditions, attention is exclusively directed towards the two primary labels (Openness and Agreeableness). The technology cannot distinguish between typographical errors and slang.

1.6 Significance of the Study

This research adds to the development of deep learning for text-based personality classification with varied sentence lengths from social media, by stressing the Hybrid CNN + BiLSTM technique. The originality of this model resides in the architecture built to optimize the capabilities of CNN in catching local patterns as well as BiLSTM in comprehending long-term word associations, resulting in more accurate personality diagnosis.

This research employs a methodology that addresses the issue of variability in text length for personality classification. The model's adaptability to both short and long texts is enhanced through the use of Word2Vec's CBOW-based word embedding. This research underscores the significance of selecting and calibrating parameters, including batch size, units, and maximum length, to guarantee optimal model performance. Hyperparameter optimization through GridSearch is utilized to enhance efficiency and accuracy.

This research presents a superior solution to singular methods, achieving a balance between local feature extraction and long-term contextual comprehension, thereby establishing a foundation for the advancement of social media text-based personality analysis to accommodate diverse sentence lengths.