**INTRODUCTION**

Depression as a common mental health disorder has long been defined as a single disease with a set of diagnostic criteria. It often co-occurs with anxiety or other psychological and physical disorders; and has an impact on feelings and behavior of the affected individuals [1]. According to the WHO study, there are 322 million people estimated to suffer from depression, equivalent to 4.4% of the global population. Nearly half of the in-risk individuals live in the South-East Asia (27%) and Western Pacific region (27%) including China and India. In many countries depression is still under-diagnosed and left without any adequate treatment which can lead into a serious self-perception and at its worst, to suicide [2]. In addition, the social stigma surrounding depression prevents many affected individuals from seeking an appropriate professional assistance.

As a result, they turn to less formal resources such as social media. With the development of Internet usage, people have started to share their experiences and challenges with mental health disorders through online forums, micro-blogs or tweets. Their online activities inspired many researchers to introduce new forms of potential health care solutions and methods for early depression detection systems. Using different Natural Language Processing (NLP) techniques and text classification approaches, they tried to succeed in a higher performance improvement. Some studies use single set features, such as bag of words (BOW) [3], [4], N-grams [5], LIWC [6] or LDA [7], [8] to identify depression in their posts. Some other papers compare the performance of individual features with various machine learning classifiers [9] [12]. Recent studies examine the power of single features and their combinations such as N-grams+LIWC [13] or BOW+LDA and TF-IDF+LDA [14] to improve the accuracy results. They experiment with a smarter text pre-processing, and introduce different substitute words depending on the nature of the original string. For instance, Tyshchenko et al. [14] suggested categorizing the stop words and adding LIWC-like word categories as an extra feature to an already designed method (BOW+TFIDF+LIWC). In addition, he applied multiple feature combinations to increase the perfor- 1mance using Convolutional Neural Networks (CNN) which consist of neurons with learnable weights and differ in terms of their layers. CNNs are very similar to simple feed-forward neural networks and state of the art method in the text and sentence classification tasks.

A meta-analysis by Guntuku et al. [15] summarizes several iterations of depression detection tasks in computational linguistics. Another interesting review for mental health support and intervention in social media is written by Calvo et al. [16] who reviewed the taxonomy of data sources, NLP techniques and computational methods to detect various mental health applications. Even with this significant progress, challenges still remain. This paper aims to search for a solution to a performance increase through a proper features selection and their multiple feature combinations. First, we choose the most beneficial linguistic features applied for depression identification to characterize the content of the posts.

Second, we analyze the correlation significance, hidden topics and word frequency extracted from the text. Regarding the correlation, we focus on the LIWC dictionary and its three feature types (linguistic dimensions, psychological processes and personal concerns). For the topic examination, we choose the LDA method as one of the successful features. For the word frequency, we use unigrams and bigrams by leveraging the vectors based on TF-IDF scheme. Finally, we set five text classifying techniques and conduct their execution using the extracted data to detect depression. We compare the performance results based on three single feature sets and their multiple feature combinations. In our experiment, we use data collected from the Reddit social media platform. It was chosen as the

data source as it allows longer posts. Targeting technical approaches towards detection tasks, our paper follows the lines of Calvo et al. research [17].

Our study has four specific contributions: first, to examine the relationship between depression and user’s language usage; second, to design three LIWC features for our specific research problem; third, to evaluate the power of N-grams probabilities, LIWC and LDA as single features for performance accuracy; fourth, to show the predictive power of both single and combined features with proposed classification approaches to achieve a higher performancein depression identification tasks.

The rest of the paper is organized as follows. In section II, we discuss related work in depression detection. In section III, we define the properties of the Reddit dataset. In section IV, we introduce the methodology and conduct data preprocessing followed by feature extraction. In section V, we compare and analyze the feature sets and examine the results as well as the most powerful machine learning technique for depression detection. We conclude our study and provide a direction for future work in section VI.