Deep Learning for Detecting Mental Health Disorders using Social Media Generated Content

Luke Abela
200588919
Dr.Arkaitz Zubiaga
Artificial Intelligence MSc

***Abstract*— The drastic increase in the use of social media over the last few years has occurred concurrently with increasing attention being given to issues pertaining to mental health disorders. Various social media applications such as Facebook, Twitter, Instagram, and Reddit have all become widely used, and their prominence allows frequent online interaction between various people. The instant gratification of social media applications causes many people to lapse into excessive use, which may lead to symptoms of mental health disorders. There are many mental health disorders with varying presentations – generally characterised by a combination of abnormal thoughts, perceptions, emotions, behaviour, and altered relationships with others. Such disorders include Depression, Bipolar, Schizophrenia and Autism. This work aimed to use social media generated content to develop a method based on a Deep Learning framework to detect whether someone is likely to have a mental health condition based on the text they post. Deep Learning based models have progressed beyond traditional machine learning techniques in Natural Language Processing tasks such as sentiment analysis, text classification, natural language inference, and news categorisation, and are therefore a promising tool to use for online mental health detection applications.**

Keywords—Deep Learning, Natural Language Processing, Mental Health Disorders, social media, Text Classification, Transformers

I. INTRODUCTION

The advent of social media has enabled new forms of socialisation which have become increasingly essential to the daily lives of millions of people. Many people, using various social media platforms have been experienced an unprecedented level of connectivity. Social Networking sites have gained incredible prominence with Facebook registering 2,740,000,000 million users as of January 2021. (statista, 2021)

Social media is fundamentally an interactive technology which allows users to create and share information. This user-generated content primarily consists of text-based posts and/or comments, possibly accompanied with digital photos or videos, and the corresponding metadata. Such broad use has enabled increased communication, allowing people to stay up to date with family and friends, join and promote social causes, and consume content of textual and graphic nature. Social media has also facilitated damaging forms of interaction such as cyber-texting, sexting, and online stalking. Notably, excessive use of these online platforms has been shown to fuel feelings of Anxiety, Depression, isolation, and FOMO (fear of missing out). (Obar & Wildman, 2015).

Mental health awareness has increased significantly in recent years, owing its origins to the mental hygiene movement, initiated in 1908, which was driven by a desire to improve the treatment and quality of life of people with mental disorders. (Bertolote, 2008) With approximately 1 in 4 people in the UK suffering from a mental condition every year, and 1 in 6 suffering from a common disorder such as Anxiety or Depression, such issues have become widespread and prevalent.

It has been noted that people are commonly readily eager to express their views anonymously online rather than in person. (Al-Saggar & Nielsen, 2014) Using various applications, anonymous users are likely to discuss their mental health problems on an online platform. (Hanwen Shen & Rudzicz, 2017) This is a positive trend due to patterns of hidden behaviours exhibited by people when fearing stigmatisation, i.e. people have been more likely to under report mental health issues compared to other health conditions due to the associated stigma. (Bharadwaj & Suziedelyte, 2017) Therefore, the anonymisation of identity, allowed by online interaction, has allowed people to discuss their mental issues without risking social stigma. By discussing their issues openly and without fear of stigmatisation, people have been less likely to under-report and therefore, have been giving more accurate accounts of potential symptoms. People may also engage in discussions but be unaware or not willing to consider that they are potentially suffering from a mental health condition.

It is unrealistic to expect mental health professionals to inspect and review large and ever-increasing amounts of data. An automated unit for the purposes of mental health text classification could alert online users if the text they are posting would be indicative of a mental health disorder and would encourage them to seek professional help.

This project focused on using the social media generated textual data to train a Deep Learning based model with the aim of detecting whether the person posting on social media is likely to be suffering from a mental health disorder based on the content they are posting onto social media. Such a task would fall into the field of Natural Language Processing (NLP). Core NLP techniques were traditionally dominated by machine learning methods using linear methods such as support vector machines or linear regression, trained over very high dimensional yet sparse feature vectors. The field has since found increased success in recent years by making use of non-linear neural network models over dense inputs, a technique known as Neural Networks. (Goldberg, 2015)

Such a technique, specifically when extended to Deep Learning, allows computational models that are composed of multiple processing layers to develop representations of data with various levels of abstraction. Deep Learning models have dramatically improved the state-of-the-art in various domains of application. The strength of these models is their capability to discover intricate structure and nuanced patterns in large datasets. This is accomplished by using the backpropagation algorithm to indicate how a machine should change its learnable weights which are used to compute the representation in each layer from the representation in the

previous layer. (LeCun, et al., 2015) The application of Deep Learning to Natural Language Processing tasks yields several advantages: superior performance at pattern recognition tasks, and the capability of end-to-end training (little or no domain knowledge is needed prior to the system construction).

Deep Learning however is a data hungry process and is hence not suitable for small quantities of data. Its resultant models are typically black box, making them difficult to understand due to the continuing lack of theoretical foundation. Furthermore, the cost of training Deep Learning models is computationally expensive. (Li, 2018)

This work relies on the notion that the text posted by a user when suffering from a mental health condition would contain different, detectable features when compared to the text generated by the same user when not suffering from a mental health condition.

II. Related Work

The diagnosis that a person requires treatment for a mental health condition is a sensitive clinical decision, a decision which encompasses a potential range of symptoms, the quality of life of the patient in relation to those symptoms, and the outcomes of those treatments. (American Psychiatric Association, 2013). It is worthwhile to note that the diagnosis of a disorder is a non-trivial task which can and should only be done by a trained mental health professional. The purpose of this work is to serve as an initial step to generate awareness for this increasingly prevalent issue. It should be noted that the respect of the appropriate use of social media data is of the utmost importance. Working with such data requires that necessary precautions be taken to prevent further potential psychological distress.

A substantial portion of the research carried out in the domain of mental disorder detection has placed an emphasis on feature engineering. Earlier works made use of the Linguistic inquiry word count (LIWC) lexicon, containing more than 30 categories of psychological features. (Pennebaker, et al., 2007). The LIWC functions by reading text and counting the percentage of words which reflect differing emotions, styles of thinking, social concerns, and parts of speech. It was developed with interests in social, clinical, health, and cognitive psychology, hence the language categories were created to capture people's social and psychological states. (liwc.wpengine, n.d.)

In more traditional machine learning approaches, researchers would dedicate time to finding the optimal set of features which feature some correlation to the specific disorder they were attempting to detect. (Rude, et al., 2004) This means that for new applications or innovative approaches to previously studied applications, an exploration of the potential features and possible combinations of features is required to produce the optimal performance. In the work of 'Detecting distressed and non-distressed affect states in short forum texts', the use of first-person pronouns as opposed to the use of second or third-person pronouns correlated well with users susceptible to distress and Depression. (Lehrman, et al., 2012). A common feature used for machine learning in text-based applications is the term frequency – inverse document frequency (TF-IDF). The TF-IDF is a method of text vectorisation and allows the resultant scores to be fed to Naïve Bayes and Support Vector Machines, greatly improving on more basic methods such as word counts. The application of TD-IDF in machine learning was used in Transgender Community Sentimental Analysis from Social Media Data. (Li, et al., 2020)

Further exploration of machine learning features has been accomplished. Mitchell, Hollingshead, and Coppersmith explored features such as Lexicon based approached (LIWC), Latent Dirichlet Allocation (LDA), Brown Clustering, Character n-grams (sequences of words), and Perplexity. This exploration was carried out with the aim of 'Quantifying the Language of Schizophrenia in Social Media'. LDA operates on the principle that a document contains a mixture of topics, where each topic contains words with varying probabilities. LDA infers topics from unlabelled text. Brown clustering follows the notion that words in context provide more meaning than words in isolation – hence methods for grouping together words occurring in similar linguistic constructions. Brown clustering is a greedy hierarchical algorithm which locates clusters of words which maximise the mutual information between adjacent clusters. (Coppersmith, et al., 2015)

There was a division of preference from where datasets for such work were extracted from. Despite the popularity of platforms such as Facebook and Instagram, the main sources of data were Twitter and Reddit. A significant fraction of the work produced in this domain has been done so using Twitter data, however one limitation of that is that the platform limits user text to 280 characters (previously 140). (Coppersmith, et al., 2014). One study attempted to circumvent this limitation by concatenating all tweets (posts) of a Twitter user in a single document. (Orabi, et al., 2018). Social media users may be modelled as a collection of their respective posts. In contrast, Reddit offers a richer source of high-volume data due to the absence of a Twitter style character limit. (Choudhury, 2014) (Cohan, et al., 2018)

Others have sought to apply Deep Learning methods to social media based mental health detection, making use of a range of techniques including custom word embeddings, and convolutional or recurrent neural network intermediate layers. (Orabi, et al., 2018) Sekulic and Strube made use of a Hierarchical Attention Network (HAN) to construct a series of binary classifiers for different mental disorders. (Sekulic & Strube, 2019) The nature of which was to detect different mental health concepts in posts themselves (Rojas-Barahona, et al., 2018), whilst others have sought to infer general information about a user. (Kshirsagar, et al., 2017) In the work of Yates et al (Yates, et al., 2017), a CNN was used on a post-level to extract features, which were then concatenated to get a user representation to be used for self-harm and Depression assessment. The CNN required posts of constant length which put constraints on the data available to the model.

Deep Learning has been shown to exceed machine learning for various tasks including text classification. In attempting to extract psychiatric stressors for suicide from social media using Deep Learning, with a CNN-based approach a resultant F1-measure of 83% was obtained, outperforming support vector machines and extra tree methods. (Du, et al., 2018) The research of Zogan, Razzak, Wang, Jameel, and Xu sought to build on the early success of Deep Learning and tackle the issue of interpretability – previous machine learning methods and their inferred results were obscure due to a lack of explanation. To this end, their work proposed the interpretive Multi-Modal Depression Detection with Hierarchical Attention Network (MDHAN) for detecting depressed users and explaining the model

prediction. The MDHAN functioned by encoding user posts using two levels of attention mechanism applied at the tweet-level and word-level, calculate the importance of the tweet and its words, and captured semantic sequences from user timelines. (Zogan, et al., 2020)

III. Preliminaries

A. Gradient Descent

Gradient descent is a foremost algorithm in popularity for performing optimization, particularly in the field of neural networks. Such an algorithm is a black-box optimizer. Gradient descent is a method to minimise an objective function by updating the parameters in the opposite direction of the gradient of the objective function. The learning rate used controls the magnitude of steps taken to reach the (local) minimum.

There are variations of the gradient descent algorithm: batch gradient descent (works directly with the entire dataset), stochastic gradient descent (performs parameter update for each training example), and mini-batch gradient descent (works with a portion of n training samples). The mini-batch gradient descent reduces the variance of each parameter update, leading to more stable convergence. However, mini-batch learning can be difficult to due several factors:

- Difficulty in selecting an appropriate learning rate – small learning rates lead to unnecessarily slow convergence; excessively large learning rates can prohibit appropriate convergence.
- The same learning rate is applied to all parameter updates, i.e., if the data is spare with distinctive features occurring with different frequencies – it may not be desirable to update all parameters at the same rate. (Ruder, 2016)

B. Deep Averaging Networks

The Deep Averaging Network consists of input embeddings for words, low dimensional vectors in N dimensional space describing a word, and bigrams averaged together, the latter of which are then passed through a feedforward deep neural network (DNN) to produce sentence embeddings. An appropriate composition function is required to obtain a vector space model for sentences or documents, The composition function is the mathematical process of combining multiple words into a single vector. The composition function may be of unordered or syntactic nature. An unordered composition function considers a bag of words approach. The syntactic approach considers sentence structure and word order. Whilst syntactic composition functions typically outperform the unordered composition functions, the syntactic functions are more expensive computationally and require increased computation time.

The deep unordered model typically obtains reliable performance on sentence and document level applications. This is achieved by taking the average of the embedding vectors associated with an input sequence of tokens. The resultant average is passed through one or more feed forward layers. Finally, a linear classification is performed on the final layer's representation with a cross entropy loss-function. (Iyyer, et al., 2015)

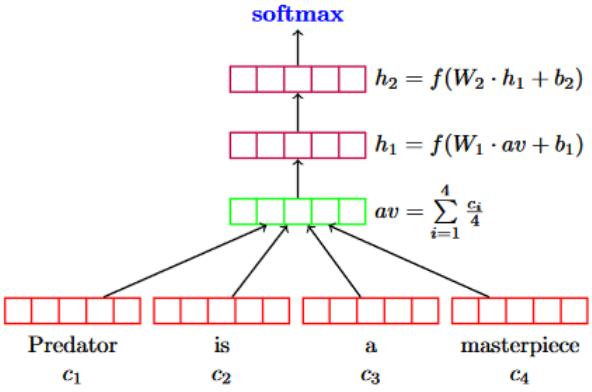

Figure 1: Deep Averaging Network (Iyyer, et al., 2015)

C. (Bidirectional) Recurrent Neural Networks

The Recurrent Neural Network (RNN) provides a means of working with (time) sequential data which contains correlations between data points which occur in proximity in sequence. (Schuster & Paliwal, 1997) The RNN typically suffers from the vanishing gradient problem – the inability of the network to propagate valuable information from the model output to the intermediate model layers. Such an issue is typically overcome using a Long-Term Short-Term Memory module (LSTM) and/or Gated Recurrent Unit (GRU).

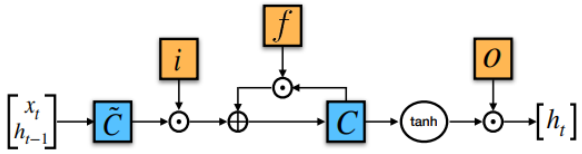

Figure 2: LSTM Architecture (Chung, et al., 2014)

The bidirectional RNN attempts to overcome certain limitations of the standard RNN. Its structure is such that it splits the state neurons of the standard RNN into a section which contains the positive time direction, and a section for the negative time direction. In this manner, both past and future information can be evaluated.

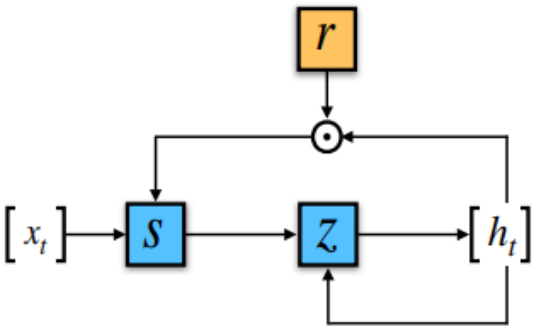

Figure 3: Bidirectional Architecture (Chung, et al., 2014)

D. Convolutional Neural Networks

The Convolutional Neural Network is similar to the standard Neural Network but makes the assumption that the input is an image. The forward function's increased efficiency emerges due to the reduced number of parameters in the network. Each neuron in a layer will only be connected to a

small region of neurons within the layer before it, as opposed to all the neurons of the previous layer in the style of a Fully Connected Layer.

The Convolutional Neural Network is a sequence of layers, making use of a Convolutional Layer, a Pooling Layer, and a Fully Connected Layer, with each layer accepting an input 3D volume and transforming it into another 3D output volume through some differentiable function. (Albawi, et al., 2017)

It makes use of a local receptive field and parameter sharing. The former representing the region of input data which affects a neuron, known as a filter. The latter refers to the sharing of parameters within filters during the training phase. The benefit is such that training parameters used in a feature detector at one section of the input data may be used in other sections.

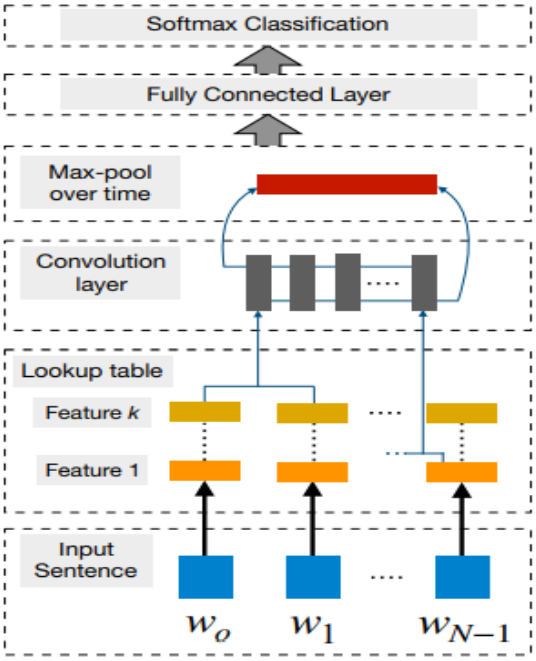

Figure 4: Convolutional Neural Network Framework for word wise class prediction (Collobert & Weston, 2008)

E. *Word Embeddings*

Word embeddings, a form of word representations which allows words of similar meaning to have similar representation, are based on the unsupervised training of distributed representations. The distributed representation is based on the context in which the words occur. Each word is represented by a vector in a predefined vector space. The distributed representation is learnt based on the usage of words – words which are used in similar ways should therefore have similar representations, capturing their meaning.

Embeddings may be developed using the Word2Vec method, a statistical procedure for efficiently developing a word embedding from a text corpus. The Word2Vec approach consists of one of two different models: Continuous Bag-of-Words (CBOW) and Continuous Skip-Gram Model. The CBOW model learns embeddings by predicting the current word based on context, whilst the continuous skip-gram develops embeddings by predicting the surrounding words given a current word. Both models aim to learn the word vector based on the local usage context of the word, where the word's context is defined as a window of neighbouring words.

The salient benefit of the word embeddings approach is that high quality word embeddings be developed efficiently, with low space and time complexity. Larger embeddings (higher dimensional spaces) may be learnt from larger corpora of text. Previously words were treated as independent units, with no concept of similarity between words, as they were represented as indices in a vocabulary – a simplistic but robust approach. (Mikolov, et al., 2013)

F. *Transformers*

Dominant sequence transduction models are based on complicated convolutional or recurrent neural networks which make use of an encoder and a decoder. The encoder-decoder architecture originates in the sequence-to-sequence model, an end-to-end approach making use of multi-layered LSTMs to map input sequences to a vector of pre-determined dimensionality (encoding), and then a subsequent LSTM to decode the target sequence from the vector (decoding). Such an architecture typically finds application in the task of language translation. (Sutskever, et al., 2014)

The best performing models use an attention mechanism to connect the encoder and decoder, known as the Transformer. Within the context of Deep Learning, attention may be considered as a vector of importance weights. When predicting an element such as a word in a sentence, the calculated attention vector is used to determine how closely the predicted word correlates with other elements. This is accomplished by taking the sum of their values, weighted by the attention vector, as an approximation of the target. (Vaswani, et al., 2017)

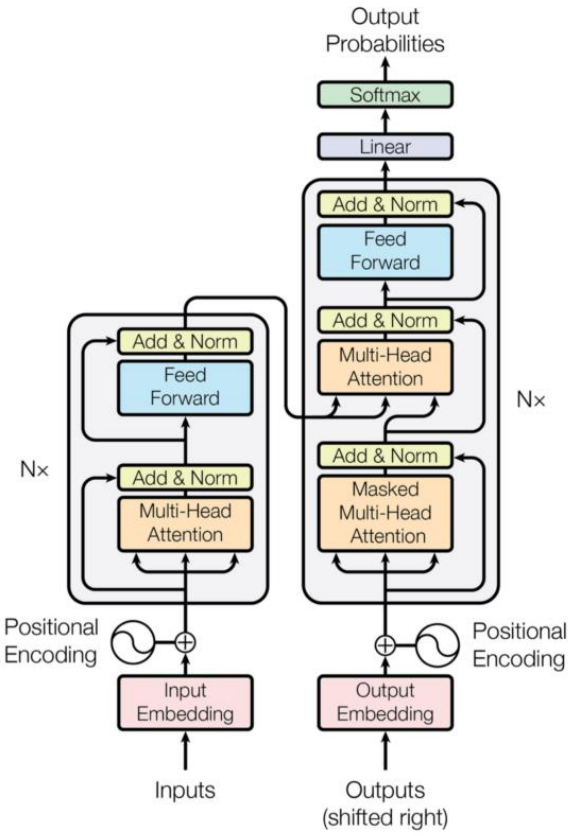

Figure 5: Transformer Architecture

Transformers were extended to Bidirectional Encoder Representations from Transformers (BERT), designed to pre-train deep bidirectional representations from unlabeled text by conditioning on both left and right context in layers. The pre-trained BERT model can be fine-tuned with an added final layer to create high performance models for various applications, such as text classification with task specific architecture modifications. (Devlin, et al., 2018)

Language model pretraining has shown significant performance improvements. The model 'A Robustly Optimised BERT Pretraining Approach' (RoBERTa) improves on the performance of 'Pre-training of Deep Bidirectional Transformers for Language Understanding' (BERT). (Liu, et al., 2019) (Devlin, et al., 2018)

RoBERTa is a large model trained for high performance. Pretrained models such as RoBERTa have had increasingly prevalence in Natural Language Processing, however, operating such large models requires significant computational training. A model known as DistilBERT, model of the same nature as BERT and operating at 97% of BERT's language understanding whilst simultaneously being 60% faster and 40% smaller in size. (Sanh, et al., 2019)

IV. TEXT CLASSIFICATION MODELS

A. Model Selection

The selected approach was to create an individual binary classifier for each disorder class. Theoretically one could consider a multiclass classifier or a multi-label classifier. However, these are currently more difficult to implement. A multiclass classifier would make the inherent assumption that any single person could only be experiencing one disorder at a time, which is not the case. Multi-label classification overcomes this limitation; however, this would require a dataset with text indicating any combination of mental disorders, single disorders, and no disorders. Such a dataset would take significant work to put together, including but not limited to the ethical and privacy hurdles which must be overcome to assemble such data. Unfortunately, sub-Reddits for the combinations of disorders do not exist.

B. Custom Embeddings Based Models

The Keras API Embedding layer turns positive integers into dense vectors of a fixed size. This layer allows the creation of a custom set of embeddings based on the input corpus dataset provided. These custom embeddings can then be used as a means of conversion from words to a dense numerical vector representation, an appropriate form for digital computation, and gradient descent training for model training. Such embeddings may be used for standard neural networks, convolutional, LSTM and Bidirectional LSTM neural networks.

C. Pre-trained Transformer Embeddings Based Models

For the models used, the pretrained RoBERTa text classifier was used as a baseline to measure RoBERTa performance. Furthermore, the pretrained RoBERTa word embeddings were used for deep averaging, 1-dimensional convolution, and LSTM and Bidirectional LSTM models.

For the models used, the pretrained DistilBERT text classifier was used as a control to measure DistilBERT performance. Furthermore, the pretrained DistilBERT word embeddings were used for deep averaging, 1-dimensional convolution, and LSTM and Bidirectional LSTM models.

D. Dataset

The dataset used, the Reddit Mental Health Dataset, was downloaded through using the 'push shift' API by Low, et al. The dataset contained posts and text features from 28 mental health and non-mental health sub-Reddits: 15 mental health support groups including Depression and Anxiety, 2 broad mental health sub-Reddits, and 11 non-mental health sub-Reddits including divorce and fitness. The dataset contained posts from 826,961 unique users from 2018 to 2020. (Low, et al., 2020) For each disorder, a list of text posts, the LIWC and TFIDF features were included within the dataset. The LIWC and TFIDF features were discarded as they were not necessary for the Deep Learning approach employed. The data used was simply the text posted to Reddit by the various users, and the label of the sub-Reddit to which the text was posted (i.e., the mental health disorder class).

For the purposes of this work, the Depression, Anxiety, schizophrenia, Bipolar, borderline personality disorder (BPD), and Autism sub-Reddits were used. Mental Health, a broad mental health sub-Reddit, was also included. The mental health sub-Reddit specifically was included to function as a form of control group, as it contained broad discussion and was not dedicated to any single specific disorder.

Table 1: Dataset Breakdown

Sub-Reddit Title	Number of Posts	Reddit Description
Anxiety	19976	Discussion and support for sufferers and loved ones of any anxiety disorder.
Autism	4576	Autism news, information, and support. Please feel free to submit articles to enhance the knowledge, acceptance, understanding and research of Autism and ASD.
Bipolar	2720	A safe haven for Bipolar related issues. We are a community here not just a help page. Be a part of something that cares about who you are.
BPD	11007	A place for those who have BPD (also known as EUPD), their family members and friends, and anyone else who is interested in learning more about the mental illness.
Depression	21209	Peer support for anyone struggling with a depressive disorder.
Schizophrenia	42481	N/A – private community
Mental Health	18925	The Mental Health sub-Reddit is the central forum

		to discuss, vent, support and share information about mental health, illness, and wellness.

E. Data Augmentation

Data augmentation methods have garnered increased interest in Natural Language Processing due to work being conducted with low-resource domains, innovative tasks, and the usage of neural networks, which require substantial amounts of data. To this end, there have been many data augmentation techniques explored. Rule-based methods such as easy data augmentation are typically sufficient for unsupervised data. Alternative methods include model-based translation, such as sequence to sequence neural translation methods (backtranslation). The relationship from language to language does not have an absolute, direct one-to-one relationship. Hence, translating a translated sentence back to its language of origin is likely to yield a differing sentence. For this reason, backtranslation is typically used as a method to generate sentences to supplement a dataset of phrases and sentences. The attention mechanism is often used to overcome the non-direct relationship between individual words in the source sentence and the translated sentence.

The selected method for data augmentation was EDA – easy data augmentation techniques. EDA were designed to boost performance for text classification tasks. It consists of four straightforward but effective operations, synonym replacement, random insertion, random swap, and random deletion:

- Synonym replacements randomly selects n words from the document, which are not pre-defined stop words, and substitutes these words.
- Random Insertion obtains a random synonym of a random word from the document which is not a stop word. The synonym is placed at a random location with the document.
- Random Swap selects two words in the document and swaps their position.
- Random deletion randomly removes each word in the document with some probability p.

EDA has been shown to boost performance on various text classification tasks. (Zou & Wei, 2019)

F. Preprocessing

For Custom Embeddings Based Models, pre-processing removed all uppercase letters, and all stop words, defined in the NLTK set. Furthermore, all punctuation, alphanumeric characters, and extra whitespaces were stripped from the tokenized documents. For Pre-trained Transformer Based Models, the relevant transformer tokenizer was used.

A statistical analysis of the text indicates that following the pre-processing, the mean length of a document was 183 words, with a standard deviation of 184 words. The length of the document at the 5th percentile was 28 words. The length of the document at the 95th percentile was 511 words.

Table 2: Data Observations

Disorder	Mean Length	STD Length	Upper Bound Percentile	Lower Bound Percentile
Autism	162	162	446	26
Anxiety	168	153	435	31
Schizo	168	190	502	22
Depression	172	187	492	24
Bipolar	173	185	471	29
BPD	188	183	514	30
Mental Health	217	209	608	33

G. Data Preparation

For each disorder, data is partitioned into two sets, positive cases of the disorder from the relevant dataset, and negative cases of the disorder randomly sampled from all order datasets, this is repeated until the number of negative cases is equal to the number of positive cases, ensuring a balanced dataset. The prepared data for the binary classifier is segmented into training, development, and testing sets in a ratio of 0.6:0.2:0.2.

A vocabulary (word index dictionary) was created by storing all unique words occurring the training dataset in a dictionary, with entries for padding and out of vocabulary words. The train, development, and test tokens were then converted from word tokens to numerical index form. Words in the development and test set which did not occur in the train were labelled as out of vocabulary words.

Model development requires data sequences of uniform length; hence each list of word tokens was pre-padded with zeros.

H. Model Training

For model training, the Keras TensorFlow API was used.

- For custom trained models, a custom embedding layer was implemented. For pre-trained models, the pre-trained word embeddings were loaded.
- For a 1-dimensional convolution model, a convolutional layer of 200 filters was used with kernels of size 3. This was followed by a layer of global average pooling.
- For BiLSTM and LSTM models, a total of 50 units were used.

Finally, a dense layer of 112 neurons was followed by an output classification layer with soft-max activation. The result for the soft-max function is the ratio of the exponential of the parameter and the sum of exponential parameter. The ADAM optimizer was used with binary cross entropy loss. Training was conducted over a maximum of 30 epochs with batch sizes of 100. Early stopping was accomplished by monitoring for limited validation loss change over 5 epochs.

V. RESULTS

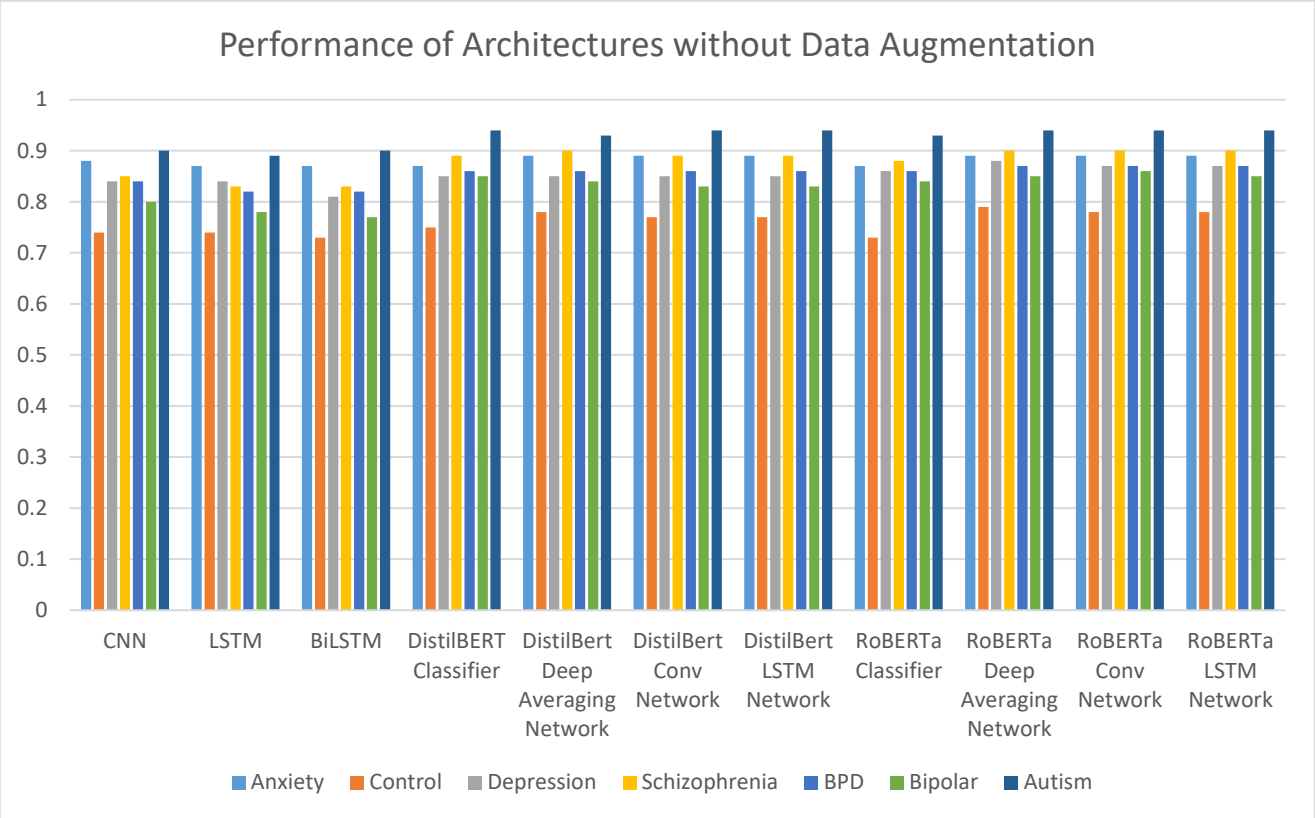

Figure 6: Performance of Architectures without Data Augmentation

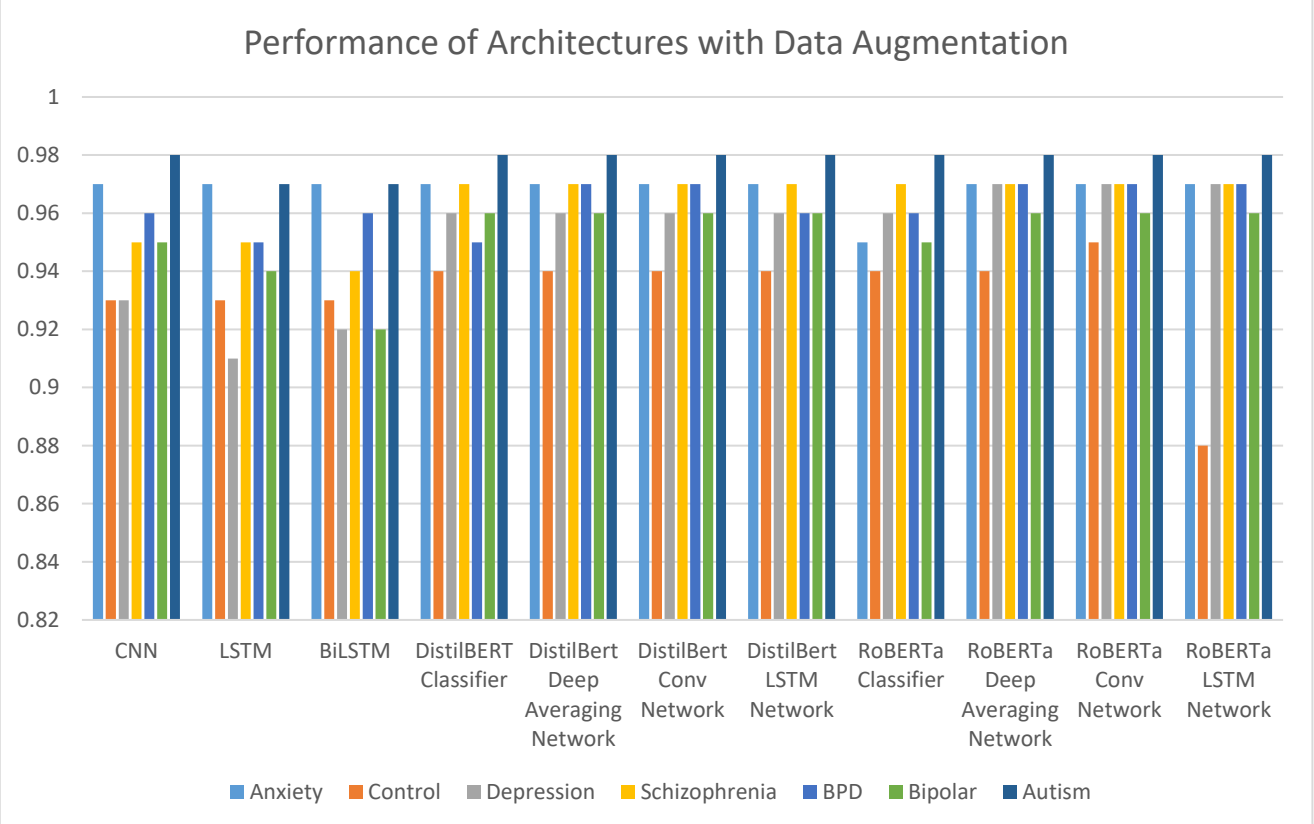

Figure 7: Performance of Architectures with Data Augmentation

VI. Discussion

A comparison of different standard architectures with different embeddings was conducted. The performance of these architectures was then improved further using data augmentation methods. In general, LSTMs, CNNs and Deep Averaging Networks performed at comparatively similar levels. Closer inspection indicated that the CNN architectures typically minimally outperformed their LSTM counter parts. The CNN results were matched but not exceeded by the Deep Averaging Network. The pre-loaded DistilBERT and RoBERTa classifiers were used as a performance comparison, to determine whether there would be any benefit to the inclusion of the CNN, LSTM, and Deep Averaging architectures. Furthermore, the custom-trained embeddings-based LSTM and CNN models without data augmentation were used as a benchmark to which all other models could be compared.

A previous analysis of the dataset showed that the average Reddit post was not excessively long, with the mean length of a document approximately 183 words. 183 words, though sufficient to constitute a short essay, is not a significant amount, this indicated the sequential properties of the LSTM did not seem to encompass the context of the sentence better than the feature extraction of the CNNs to determine the class of the observation.

Another interesting facet was the variation in performance for the same architecture between disorders. Typically, the 'Mental Health' Reddit data performed poorest. This was possibly due to the broad nature and intentions of the sub-Reddit, as opposed to the more focused natures of the other sub-Reddits. In contrast, the best performing dataset across the architectures was Autism, which was also the second smallest class. The mean length of the Autism observations was the smallest from all datasets; thus, the brevity of the sentences could have been a contributing factor to the success. Similarly, the 'Mental health' dataset sentences had the greatest mean length and performed poorly. Depression, Schizophrenia, and Anxiety also performed well in comparison to datasets with longer average sentence lengths. This trend could indicate that for such a text classification task, brevity of sentence was beneficial, and would explain why the LSTM based architectures did not perform as well as its counterparts.

The pre-trained embeddings of DistilBERT and RoBERTa served to improve the performance, exceeding the performance of the custom trained embedding set. DistilBERT and RoBERTa were trained on a very large collections of text, making use of techniques such as byte-pair encoding, as opposed to a more traditional word encoding mechanism. Such sub-word learning would represent words as sub-word tokens with start and end tokens, which would provide superior potential recognition of previously unseen words by treating sub-word tokens as opposed to whole words. Overall, the highest performing model was the RoBERTa Convolutional Network. RoBERTa was a higher performance model than DistilBERT but also larger and more computationally expensive. DistilBERT provided a significant increase over the baseline whilst being smaller than RoBERTa. The selection between such models would be dependent on computational resources available and ultimately the importance placed upon maximising the performance of the text classifiers.

The quantity and quality of training such embeddings clearly exceeded the custom trained embeddings. The strength of BERT model embeddings is remarkable and could conceivably become a standard in various text applications with exceptions being made for tasks which require highly unusual and/or technical terminology such as medical journals due to the frequent use of Latin words.

Furthermore, data augmentation also showed great improvements to performance at test time. The use of EDA data augmentation methods was well suited to text originating from social media. Text originating from articles published by reputable sources, such as BBC, a corporate publication, and/or professional writers is typically proofread and continuously edited to ensure no spelling or grammatical errors are left. Social Media text, in contrast, is much more likely to contain spelling errors, incorrect grammar, accidental repeated words, random missing words, and randomly swapped words. These errors occur due to social media users not typically proof reading or editing their once positing. Thus, by effectively introducing similar modifications of synonym replacement, random insertion, random word swap, and random word deletion through EDA, a social media dataset can be augmented by creating sentences which will not necessarily be grammatically perfect, like text predominantly on social media.

VII. Conclusiojns & Future Work

The results achieved in this study indicate that there is significant potential for the detection of possible mental health disorder symptoms via text classification of social media. Deep Learning continues to achieve state-of-the-art results in various works, and this field is no exception. There do exist short comings in this work and others. Significant effort has been placed into developing the text classifiers and developing individual datasets from various domains such as Twitter and Reddit. Further work should delve into unifying these various text classifiers to develop a cross-platform classifier which would be applicable to many social media platforms including Facebook, Twitter, Reddit.

Building on such work, an individual user could theoretically have their various posts across all social media sites concatenated to assess an individual more comprehensively. Along with text, other types of data are posted by users on social media. Photos, gifs, videos, music, links with corresponding content are all posted, and supplement the intent behind the text, if not replacing the text entirely. A comprehensive multi-modal approach would serve to better assess online users.

Furthermore, issues remain with data scraping and collection from social media. The Reddit sub-Reddits provide a good starting point for this task, however precise data collection for such a task remains a challenging task. Users may be struggling with multiple disorders, for varying degrees of time, and similarly may recover and/or relapse. Developing a precisely annotated dataset for the ideal multi-label classifier spanning across multiple social media sites considering multiple forms of media would serve as a significant challenge but would represent a significant breakthrough in this field.

VIII. Acknowledgment

I would like to express my gratitude to my supervisor, Dr Arkaitz Zubiaga, for his guidance throughout this project. Furthermore, I would like to thank Endeavour Malta for their sponsorship of my studies.

IX. References

Albawi, S., Mohammed, T. A. & Al-Zawi, S., 2017. *Understanding of a Convolutional Neural Network.* s.l., International Conference on Engineering and Technology (ICET).

Al-Saggar, Y. & Nielsen, S., 2014. Self-disclosure on Facebook among female users and its relationship to feelings of loneliness. *Computers in Human Behaviour*, Volume 36, pp. 460-468.

American Psychiatric Association, 2013. *Diagnostic and statistical manuel of mental disorders.* 5th ed. Washington: American Psychiatric Publishing.

Bertolote, J., 2008. The roots of the concept of mental Health. *World Psychiatry*, 7(2), pp. 113-116.

Bharadwaj, P. M. M. & Suziedelyte, S., 2017. Mental Health Stigma. *Economic Letters*, Volume 159, pp. 57-60.

Choudhury, M., 2014. *Mental health discourse on reddit: Self-disclosure, social support, and anonymity.* s.l., In Eighth International AAAI Conference on Weblogs and Social Media..

Chung, J., Gulcehre, C., Cho, K. C. & Bengio, Y., 2014. *Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling.* s.l.:arXiv.

Cohan, A. et al., 2018. *SMHD: A Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions.* s.l.:s.n.

Collobert, R. & Weston, J., 2008. A unified architecture for natural language processing: Deep neural networks with multitask. *Proceedings of the 25th international conference on Machine learning.*, pp. 160-167.

Coppersmith, G., Dredze, M. & Harman, C., 2014. *Quantifying Mental Health Signals in Twitter.* s.l., Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality.

Coppersmith, G., Mitchell, M. & Hollingshead, K., 2015. *Quantifying the Language of Schizophrenia in Social Media.* s.l., s.n.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K., 2018. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* s.l.:s.n.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K., 2018. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* s.l.:s.n.

Du, J. et al., 2018. Extracting psychiatric stressors for suicide from social media using deep learning. *BMC Medical Informatics and Decision Making*, 18(43).

Goldberg, Y., 2015. *A Primer on Neural Network Models.* [Online]
Available at: https://u.cs.biu.ac.il/~yogo/nnlp.pdf
[Accessed 5 October 2021].

Hanwen Shen, J. & Rudzicz, F., 2017. *Detecting Anxiety through Reddit.* s.l., Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology –- From Linguistic Signal to Clinical Reality.

Iyyer, M., Manjunatha, V., Boyd-Graber, J. & Daume III, H., 2015. *Deep Unordered Composition Rivals Syntactic Methods for Text Classification.* s.l., Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).

Kshirsagar, R., Morris, R. & Bowman, S., 2017. *Detecting and Explaining Crisis.* s.l., Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology –- From Linguistic Signal to Clinical Reality.

LeCun, Y., Bengio, Y. & Hinton, G., 2015. Deep Learning. *Nature*, Volume 521, pp. 436-444.

Lehrman, M. T., Alm, O. A. C. & Proano, R. A., 2012. *Detecting distressed and non-distressed affect states in short forum texts.* s.l., Proceedings of the Second Workshop on Language in Social Media.

Li, H., 2018. Deep learning for natural language processing: advantages. *National Science Review*, 5(1), pp. 24-26.

Li, M., Wang, Y., Zhao, Y. & Li, Z., 2020. *Transgender Community Sentiment Analysis from Social Media Data: A Natural Language Processing Approach.* s.l.:s.n.

Liu, Y. et al., 2019. *RoBERTa: A Robustly Optimized BERT Pretraining Approach.* s.l.:s.n.

liwc.wpengine, n.d. *HOW IT WORKS.* [Online]
Available at: http://liwc.wpengine.com/
[Accessed 17 06 2021].

Low, D. M. et al., 2020. Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study. *Journal of medical Internet research*, 22(10).

Mikolov, T., Corrado, G., Chen, K. & Dean, J., 2013. *Efficient Estimation of Word Representations in Vector Space.* s.l., Proceedings of the International Conference on Learning Representations.

Obar, J. & Wildman, S., 2015. Social media definition and the governance challenge: An introduction to the special issue. *Telecommunications Policy*, 39(9), pp. 745-750.

Orabi, A. H., B. P. & Orabi, M. H. I. D., 2018. *Deep Learning for Depression Detection of Twitter Users.* New Orleans, LA, Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic.

Pennebaker, J. W. et al., 2007. *The Development and Psychometric Properties of LIWC2007.* s.l.:s.n.

Rojas-Barahona, L. M. et al., 2018. *Deep learning for language understanding of mental health concepts derived from Cognitive Behavioural Therapy.* s.l.:s.n.

Ruder, S., 2016. *An Overview of Gradient Descent Optimization Algorithms.* s.l.:s.n.

Rude, S., Gortner, E.-M. & Pennebaker, J. W., 2004. Language Use of Depressed and Depression-Vulnerable College Students. *Cognition and Emotion*, 18(8), pp. 1121-1133.

Sanh, V., Debut, L., Chaumond, J. & Wolf, T., 2019. *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.* s.l.:s.n.

Schuster, M. & Paliwal, K., 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11), pp. 2673-2681.

Sekulic, I. & Strube, M., 2019. *Adapting Deep Learning Methods for Mental Health Prediction on Social Media.* s.l.,

Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019).

statista, 2021. *Global social networks ranked by number of users 2021*. [Online]
Available at: https://www.statista.com/
[Accessed 09 06 2021].

Sutskever, I., Vinyals, O. & Le, Q. V., 2014. *Sequence tp Sequence Learning with Neural Networks.* s.l., Advances in Neural Information Processing Systems 4.

Vaswani, A. et al., 2017. *Attention Is All You Need.* s.l.:Google.

Yates, A., Cohan, A. & Goharian, N., 2017. *Depression and Self-Harm Risk Assessment in Online Forums.* s.l., Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.

Zogan, H., Wang, X., Jameel, S. & Xu, G., 2020. *Depression Detection with Multi-Modalities Using a Hybrid Deep Learning Model on Social Media.* s.l.:s.n.

Zou, K. & Wei, J. W., 2019. *EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks.* s.l.:s.n.