COMP47200 DISSERTATION

Safiya Rehmat, under the supervision of Dr. Derek Greene

University College Dublin, Belfield, Dublin 4, Ireland
safiya.rehmat@ucdconnect.ie

Abstract. The rapid development of social media provides a platform to share one's thoughts and emotions. This presents a rich opportunity to understand the author's state of mind. As such, it could potentially be used to diagnose mental illnesses in users of such platforms, and could be a beneficial tool for mental health clinicians worldwide. In this thesis we aim to investigate linguistic latent features in Reddit posts, which will be used to classify those users into two categories: depressed and non-depressed, based on machine learning algorithms.

Keywords: Social Media, Depression, Classification, Machine Learning.

1 Introduction

Depressive mental health issues are an area of growing concern in recent years. The WHO's Commission on Social Determinants of Health stated that the leading cause of medical diseases in developed countries by 2030 will be mental disorders like depression and PTSD [1]. According to a new health study by Eurostat, Ireland has the highest share of its population reporting chronic depression (12%)[1]. This makes the early detection of mental health problems an important area of research.

With usage of social media sites becoming more popular than any other form of media, millions of individuals utilize these platforms to convey their thoughts, personal experiences, and social ideals, as well as to gain support and advice from their fellow peers. The aim of this research is to explore the potential of social media platforms to detect depression amongst its users.

1.1 Motivation: How is social media useful for depression detection?

Given the challenges of diagnosing mental illnesses like depression, we want to examine the potential of using social media posts to predict such disorders in individuals. With the growing popularity of microblogging sites like Twitter and Reddit, people post about their everyday lives, emotions, thoughts, and moods online on a regular basis. These posts have the potential to explain their behavior patterns, changes in emotions and moods. The work of Park et al. [2] showed that a healthy individual perceives Twitter as a tool for information sharing and consumption. However, depressed

[1]https://ec.europa.eu/eurostat/web/products-eurostat-news/-/EDN-20181010

individuals perceive it as a social awareness and emotional interaction tool. Social media provides a platform where individuals feel free to broadcast their thoughts, fears or anxieties, which they may not feel comfortable sharing with each other in person. Also, many individuals post about how they recovered from problems like depression, in order to raise awareness and decrease the stigma attached with mental health. Additionally, social activity can be an indicator of behavioral changes in an individual, like decreased frequency of posts, increased social activity during the nights or withdrawal from social media altogether [3]. These are powerful indicators that a person might be suffering from depression. Thus, social media platforms provide a unique quantifiable perspective on human behavior, making it a powerful tool for mental health scientists.

Another useful application of this research is to enhance the data available to mental health clinicians and researchers, thus enriching the mental health field. Mental health research lacks the quantifiable data available to many physical health disciplines due to the complexity of underlying causes of mental illness and the societal stigma attached with it. This lack of data hampers mental health research in terms of better diagnosis and more effective treatment of patients [4]. Since social media is present in abundance and provides access to a diverse set of population, exploiting it to better understand human behavior and trends in patients of mental health problems can be very beneficial [5] [6].

1.2 Why choose the Reddit platform over Twitter?

Twitter has been the most popular choice of social platform for most of the research being conducted in the area of medical illnesses and social media because data is available in large volumes. However, twitter feeds have a length restriction of 140 characters which limits the information contained per tweet, data about linguistic style or the emotion and sentiment of the user as well. On the other hand, Reddit allows posts up-to 40,000 characters per comment and thus offers relatively rich bodies of text from users. It has 330 million active users every month that contribute to more than 130K communities. The subreddits relating to depression have more than 496k members who post personal stories and grievances seeking or giving advice. De Choudhary [7] showed that users often share their experiences with mental illnesses and the impact it has their life, work and relationships. Many people also use this platform to seek diagnosis and treatment as well. Thus, Reddit provides an ideal platform to study the user posts and use them to detect illnesses like depression on social media.

2 Background Literature

There have been many studies conducted to understand the relationship between mental health and language usage, which have provide insights into their detection based on writings of the author. For instance, Stirman and Pennebaker [8] compared the word usage of 300 poems written by 9 suicidal and 9 non-suicidal poets, and demonstrated that there is increased use of first-person pronouns (I, me, we) in suicidal writers. Rude et al. [9] examined the linguistic patters in essays written by college students who are

currently depressed, were previously depressed, or had never been depressed. According to this research, depressed students use less positive emotion words and negative valanced words tend to dominate in their writings.

With the development of social media and the internet, studies about mental health illnesses have found new opportunities and challenges. People are increasingly using social networking sites to express their feelings and moods, help members of their community, and create an online support system. This user-generated content (UGC) can reveal many underlying characteristic details about the users and give insights into their thoughts, mood, communication, activities, and socialization. The emotion and language of the content may indicate feelings like loneliness, hatred, guilt or worthlessness which characterize depression. Also, variations in frequency of social activity can indicate withdrawal from normal behavior. There has been considerable research in using UGC to detect the onset of mental health diseases like depression, bipolar disorder and Post-Traumatic Stress Disorder (PTSD). Reece et al. [10] showed that we can detect early stage depression with an accuracy of 87% months before its onset. Another work by Reece et al. [11] used machine learning to analyze image features signaling depression from Instagram posts.

2.1 Feature Extraction

Many behavioral and linguistic features have been engineered to gain information from social media posts. These features attempt to capture the sentiment, linguistic style, length, and other morphological features of the posts, and then use these features to distinguish depressed users from the control group.

The work of De Chaudhary et al. [3] employs social engagement measures like volume of posts, proportion of reply posts and retweets, and an "insomnia index" defined as the normalized difference in the number of posts made between the day and night. They also consider 4 measures of emotion: Positive Affect (PA), Negative Affect (NA), Activation, and Dominance. Activation refers to the degree of physical intensity in an emotion, while Dominance refers to the degree of control in an emotion. PA and NA measurements were computed using the psycholinguistic resource LIWC[2] whereas Activation and Dominance were computed using the ANEW lexicon [12]. Other features like linguistic style, words from a depression lexicon, and antidepressant usage were also included for each user. Feature vectors were constructed for every user by representing the features above as a time series, containing four values per measure: the mean, variance, mean-momentum, and entropy. The feature vectors are standardized to have zero mean and unit variance.

Coppersmith [5] uses LIWC, language models, and measures of an individual's life pattern as features. These are subsequently used in machine learning models. LIWC provides clinicians with a tool for gathering quantitative data regarding the state of a patient from the patient's writing [13]. In addition to using LIWC to find 'positive affect' and 'negative affect', it is used to create features using LIWC categories like Swear, Anger, Anxiety and combine pronoun classes by linguistic form. Another

[2] http://www.liwc.net

method to create features for the classifiers is using Language Models. These models are used to estimate how likely a given sequence of words is to appear. Here two language models are employed for each class (diagnosed group and control group). The first model is a traditional 1-gram ULM which estimates the probability of each whole word. The second model is a character 5-gram model that estimates the probability of a sequence of characters. Once the models are built, each tweet is scored by computing these probabilities and classifying it according to which model has a higher probability. Pattern of life features measure behavioral trends for a user, like tweet rate, insomnia (measured as the number of tweets a user makes between midnight and 4am in their local time zone), exercise-related data, and presence and valence of sentiment words from a sentiment lexicon [14]. The proportion of these features is used in the subsequent machine learning and analysis.

In addition to linguistic features (Tf-idf, n-grams and bigrams), [15] introduces the use of stylometric and morphological features. The research studies the merits of data represented using bag-of-words, bigrams and embedding models for the classification of depressed and non-risk users. The bag of words feature is assessed by the setting Tf-idf values for all words that appear more than once in the dataset. The intuition to use this feature is that depressed and non-depressed users use different words with different frequencies and capturing this trend can be a useful signal for distinguishing between the two classes. Another source of features is based on word embeddings. A 100-dimensional word embedding trained on twitter messages (Pennington et al. [16]) is used to represent the user posts and the 850 most useful words are then averaged out per user. Bigrams are used as a feature in combination with stylometric features that include the number of words in a message, the number of sentences in a message ,and the number of words in a sentence. The motivation given for using these stylometric features is that people in the depressed class usually write longer sentences than the non-depressed class. Morphological features used for this task consist of parts-of-speech usage proportions based on the idea that the two classes of users use different parts of speech in different proportions in their posts.

An interesting approach to generate features from social media texts has been topic modelling, typically using Latent Dirichlet Allocation (LDA). LDA is a generative technique used for discovering general topics from bodies of texts (Blei et al. [17]). In this model, each document is viewed as a blend of latent topics, and each topic is characterized as a cluster of similar words. Hanwen [18] uses this technique for detecting anxiety amongst Reddit users by generating two LDA models, for the anxiety and control class each, and generating latent topics for each class. Michael et al. [19] use LDA for detecting depression in Reddit posts. They report that the model works best on the validation set when it is limited to 70 topics. For topic selection, they consider only those words that appear in more than 10 documents, where each document is a single post that has been tokenized and stemmed. All stop words are removed prior to the topic modelling process and the LDA implementation they use is provided by the Mallet toolkit [20]. The results of both these papers show that LIWC outperforms LDA when used as a single feature for prediction, but both features combined give superior classification accuracy.

2.2 Deep Learning Approaches

In this section we explore some recent work done to leverage deep learning techniques to distinguish the depressed group from the control group. The majority of the work involved in applying the classical machine learning approaches discussed above is focused on creating useful features that can distinguish a depressed user based on their behavioral, linguistic or morphological styles. As can be seen from the previous section, the complexity involved in defining, extracting for each user, and verifying the utility of each individual feature is substantial. Thus, it is potentially useful to apply deep learning models, which can interpret these latent features within the network itself and give better classification results.

The work done in this area is relatively recent. Orabi et al. [21] evaluates the effectiveness of four deep learning models, in which the first three use Convolutional Neural Networks (CNNs) and the fourth uses a bi-directional Long Short Term Memory (LSTM) RNN model. These models are built on top of a word embedding representation with a dropout layer in between. The first model uses a 1-dimensional convolution operation with 250 filters and a kernel size of 3. This is followed by a global max-pooling layer. The second CNN model is called a multi-channel CNN as it applies 3 convolutions of 128 features and kernel length of 3,4 and 5. This is followed by a 1d convolution operation and a max-pooling layer to extract abstract information. The third CNN model extends the previous one by applying 2 max-pooling layers of 2 and 5. The LSTM model uses bidirectional LSTM layer with 100 units which receives a sequence of tokens as inputs. It captures the temporal and abstract information in forward and backward directions. Some of the deep learning models have not been very useful in this area of research as the datasets we have concerning depressed users are generally small in size.

2.3 Machine Learning Models

This section presents an overview of the features and machine learning models used for the prediction task from various data sources in recent research. Table 1 summarizes these features and models along with the reported performance in each paper.

Table 1. Overview of classification models used to detect mental illnesses in recent studies.

Model	Features	Data Source	Performance	Ref.
SVM	Tf-idf + stylometric + morphology features	Reddit (CLEF eRisk 2017)	F1 - 63% Acc-90.7%	[15]
	Word2vec, n-grams	Reddit (Anxiety)	Acc – 91%	[18]
	LIWC + LDA + bi-grams	Reddit [22]	F1 – 91% Acc – 90%	[19]
Random Forest	TfIdf + morphology features	Reddit (CLEF eRisk 2017)	F1 – 62.79% Acc-92.01%	[15]
	Lexicon, tweet timing, frequency, polarity	Twitter API	Acc – 81.04%	

MLP	N-Grams + LIWC	Reddit (Anxiety)	Acc – 98%	[18]
	LIWC + LDA + bi-gram	Reddit [22]	F1 – 93% Acc – 91%	[19]
CNN with Max Pooling	Optimized Word Embedding (combining skip-gram and CBOW models)	Twitter (CLPsych 2015)	F1 – 86.96% Acc – 95.1%	[21]
BiLSTM			F1 – 80.04% Acc – 80.52%	

3 Experimental Setup

3.1 Problem Statement

This research project will seek to explore and investigate the following tasks:

3.1.1 Automatic early detection of depression in individuals based on user-generated content from online communities.

3.1.2 Identification of key characteristic features from the data extracted for individuals identified as belonging to a risk case of depression.

3.1.3 Comparison and Analysis of the performance of traditional classifiers with feature engineering and deep learning models where the features are learnt within the model itself.

3.2 Dataset

Our research uses an existing dataset which is publicly available and was previously used as part of an open academic research task[3]. The data was collected by Losada & Crestani [23] specifically for the purpose of early detection of depression. The dataset consists of public posts on the popular Reddit platform using the freely-available Reddit API. All posts in this dataset were publicly-posted and do not involve private or protected messages. Since no new data is collected for the research, an application for exemption from full ethical review was made for the same to the UCD School of Computer Science Taught Masters Research Ethics Committee[4].

The CLEF Early Risk Prediction dataset (2017) consists of posts from 887 Reddit users, of which 135 are labeled as "depressed" [23]. Users were manually categorized as risk or non-risk cases of depression. Each document in the CLEF dataset contains all the Reddit posts of a user within a certain time period. The time duration between the first and the last message varies for every user, and so do the number of posts per user.

[3] http://early.irlab.org/2017/task.html

[4] TMREC Reference Number: SCSe19_1_Rehmat_Greene

The data may contain as few as 10 messages or as many as 2000 messages per user. Overall, 15% of the examples are positive (i.e. they belong to the depression risk case) and 85% are in the non-risk control group. Details of the dataset are given in Table 2.

Table 2. Statistics of the pilot CLEF eRisk 2017 dataset.

	Training Data	**Testing Data**
Number of Users	486	401
Number of Posts	294,817	236,371
Non-Risk users	403	349
Risk users	83	52
Non-risk posts	263,966	217,665
Risk posts	30,8511	18,706

3.3 Workflow

In our work, we start with the Reddit dataset mentioned above and preprocess the data by using standard Natural Language Processing (NLP) techniques like tokenization, stemming, and stop word removal. This processed data can then be used to generate features for each user to characterize their social media posts. Once the features are generated, we can use them to classify each user into the risk/non-risk category by passing them to machine learning models: random forests, decision trees, SVMs, and neural networks. The performance of each model is then evaluated using the validation dataset and the utility of each feature in achieving the accuracy of the model is studied.

3.4 Evaluation

In a binary classification task, the terms ‘’positive’’ and ‘’negative’’ refer to the classifier’s prediction. The standard structure of a confusion matrix for a binary classifier is shown in the table below.

Table 3. Structure of a confusion matrix for a binary classification task.

	Actual class (Reference)	
Predicted class	TP (True positive)	FP (False Positive)
	FN (False Negative)	TN (True Negative)

The metrics for evaluating the performance of a classifier are derived from the values of the confusion matrix and are described below.

Precision: The ratio of observations predicted as positive which are in-fact positive.

$$Precision = \frac{TP}{TP + FP}$$

Recall: The ratio of all positive observations which were predicted by the classifier.

$$Recall = \frac{TP}{TP + FN}$$

F1-score: The harmonic mean of precision and recall, where each is given equal weighting.

$$f1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

The values for all of these metrics are calculated separately for each class.

Metrics significance in case of unbalanced classes

If both the classes are balanced and equally important, the values are averaged out to get metrics for the classifier. However, in case when one class is under-represented in the dataset or we are interested in identifying observations of a single class, it is better to use the metrics for the specific class instead of an average metric.

In cases of an imbalanced dataset, an important metric to measure the accuracy of the classifier is the *Balanced Accuracy Score*. It measures how accurately the classifier can identify observations for each class individually.

$$Balanced\ Accuracy = \frac{\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right)}{2}$$

The detection of users from the depressed class is a more useful and informative task than the detection of control group. In such cases, where both classes are not equally important, it becomes crucial to estimate the cost associated with false positives and false negatives. Since this tool is likely to be used as a screening tool for diagnostic purposes, detecting people who belong to the control group as depressed users is less costly compared to missing out the detection of depressed users. Thus, models having a good balanced accuracy score and a good recall for the positive class are ideal for the purpose of the data that we are evaluating.

4 Experiments

In this section we describe the set of experiments that were performed using different classification models and data feature vectors, and we report their performance on the depression detection task. Experiment 1 sets a baseline model using the data and explores the challenges present in the data which are then addressed in the further experiments. All the experiments are performed on the data described in section 3.2 for training and testing the models. The experiments are designed to answer the problems stated previously in Section 3.1:

1. Experiments 1 and 2 aim to answer the first problem statement (Section 3.1.1) via the automatic classification of users into two classes: depressed users, and the control group.
2. Experiments 3,4,5 and 6 are designed to create appropriate feature sets to represent the user generated content as described in the second problem statement (Section 3.1.2).
3. Experiment 7 focuses on the third part of the problem statement (Section 3.1.3) which relates to the use of deep learning models, and performing a comparative analysis of its performance with respect to the other commonly-used models.

4.1 Experiment 1: Baseline Model with Bag-of-Words

Experiment 1 is designed to quantify the performance of the dataset on a classical machine learning model, and establish a baseline for further experiments. We also highlight a number of important challenges and characteristics of the dataset which need to be addressed to improve performance further.

The dataset under consideration contains multiple posts for each user, for both the depressed group and the control group. In this experiment, the user data is extracted from the XML files and the posts for each user are combined into a single document per user. The label for each user is also recorded for the purposes of training and evaluation. The documents for each user are then pre-processed using steps like tokenization, stop word removal, stemming, and lemmatization. The processed text is the converted into feature vectors using the bag-of-word model, so that it can be passed to the learning algorithm. The vectorizer used in the experiment keeps the top 1000 most frequent words. We exclude any stop words and words that appears in more than 95% of the documents.

Once the feature vectors are generated for each user, they are split into training and test sets randomly - 80% of the data is used for training whereas the remaining 20% of the data is used for evaluation of the models. The following classification models are used in this experiment: Decision Trees, Random Forests, Naïve Bayes, Support Vector Machines, and *K*-Nearest Neighbors.

The best performing classifier was the random forest classifier which had a F1 score of 80.42% and a balanced accuracy score of 52.41%. The figure below shows the confusion matrix for the validation set on the Random Forest model.

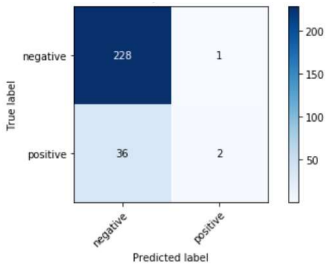

Fig. 1. Confusion matrix for experiment 1 using a Random Forest model.

As we can see from the figure 1, the Random Forest model predicts only the negative class (the control group) with maximum probability. Thus, scores for both classes need to analyzed to understand the classification efficiency.

Table 4. Classification report for experiment 1 using Random Forest

Class	Precision	Recall	f1 score	Support
Negative	0.86	1.00	0.92	229
Positive	0.67	0.05	0.10	38

The positive class, which refers to the depressed users that the algorithm aims to detect, has a recall of 0.05% as only 2 of the 38 users are classified correctly.

Model bias towards the majority class

Since the data is distributed in such a way that one class a dominant group (the control group, which constitutes 85% of the dataset), we will get a high accuracy score if the model always predicts this class. But in this case, the model does not learn a model which fits the depressed users, and is unable to detect them. The dominance of one class over the other not only introduces a model bias, the feature vectors created are also dominated by terms in the text belonging to users of this class, which exaggerates the situation further.

This challenge of majority classes biasing a classification model is present in other spheres of research, such as spam filtering, cancer diagnosis, and anomaly detection, where the class of interest is in the minority. For our task, the problem could be solved if we could obtain more data about the depressed users, so that the distribution is more balanced. However, in practice getting data about users is an expensive and complicated task. Also, the current proportion in the data reflects the real-world scenario, where most users are not depressed and only a relatively small percentage of the population belongs to the depressed class.

4.2 Experiment 2: Imbalanced Data - Under Sampling and Over Sampling

The challenge of imbalanced dataset in machine learning problems can sometimes be resolved by resampling the data so that both the classes in the dataset are approximately balanced. This can be done by *under sampling* the majority class. That is, if one class is over-represented in the dataset, we remove some instances of this majority class. The dataset can also be balanced by duplicating examples from the minority class, if it is under-represented in the data. This technique is known as *over sampling*. Both these techniques produce datasets which maybe significantly smaller or significantly larger than the original dataset, in the case of over sampling and under sampling respectively.

Over sampling can be achieved by applying various algorithms, such as Random Naive Over Sampling, Synthetic Minority Over Sampling Technique (SMOTE) [24], or Adaptive Synthetic (ADASYN) [25]. In Random Over Sampling, the observations of the minority class are randomly selected and duplicated to create more instances of

that class. While this technique increases the number of samples of the minority class, and thus offset the bias of machine learning algorithms to predict the majority class, it does not add more information about the minority class to the dataset. This is because the new instances are exact duplicates of existing instances of the class. Algorithms like SMOTE and ADASYN add more information to the dataset by interpolating the minority class to create slightly different examples.

Like over sampling, under sampling can be achieved using stochastic methods like Random Under Sampling, and Cluster Centroid Under Sampling [26]. In Random Under Sampling, random instances of the majority class are chosen and removed from the dataset to make it more balanced. The centroid-based technique uses a more sophisticated technique for removing instances from the dataset. In this case, the instances of the majority class are clustered and the instances of the cluster are replaced by the cluster centroid, thus ensuring the distribution of instances of the majority class does not change and that minimal information is lost in balancing the datasets.

Table 5 shows result metrics for the experiments using various re-sampling techniques described in this section. The results are based on the bag-of-words features of the dataset for Random Forest classifier model.

Table 5. Comparison of the performance of various resampling techniques.

Type	F1 score (average)	Balanced Accuracy Score	Precision (Target class)	Recall (Target class)
Random Under Sampling	69.26	71.12	0.59	0.82
Cluster Centroids	90.08	89.93	0.94	0.85
Random Over Sampling	90.12	89.77	0.84	1.00
SMOTE	93.80	93.81	0.93	0.95
ADASYN	96.46	96.31	0.94	0.99

4.3 Experiment 3: Tf-idf feature vector representation

The aim of this experiment is to investigate the effect of weighting features in a bag-of-words representation, by incorporating an inverse document frequency factor to the vectors via Tf-idf. This weighting has the effect of reducing the impact of terms which appear in the majority of document in a corpus.

For the experiment, the feature vector consisted of only the 1000 most useful terms as ranked by Tf-idf weights. These terms were found from the text of user posts of the depressed users in the training set only. The features are created just using the depressed class users as these are present in minority in the dataset and contain all the significant terms for classification of the users. Terms that appear in just a single document or in more than 85% of the documents are removed. Once the features are identified, all users' text is converted into a vector representation, which is then passed on to the learning algorithm.

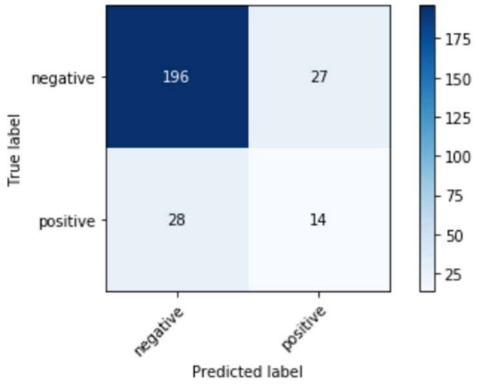

Fig. 2. Confusion matrix for Decision Tree model using Tf-idf features

The best result for this experiment using Tf-idf features is obtained using decision tree model and Figure 2 shows the confusion matrix for the same. On comparison with figure 1 of the best performing confusion matrix using bag-of-words features, we can see that significantly greater number of depressed users are identified by using Tf-idf scores. Hence, Tf-idf weighted features provide a better representation of the text than the bag-of-words features used in experiments 1 and 2.

4.4 Experiment 4: Linguistic metadata and N-grams feature representation

The structure of natural language is often captured by short phrases. As such, bigrams and tri grams can be useful features which better capture the contents of the text. In experiment 3, it was observed that certain terms like “good” and “feel” had a high Tf-idf score. These terms can change their context depending on the combination of words they are used with. The aim of this experiment is to use bigrams and trigrams to capture the context, which can then be used as features for the classification models.

The feature vector of bigrams and trigrams contains the 250 most useful n-grams by Tf-idf score. The value of feature length is determined by varying number of features between 100 and 1000 and evaluating the performance on the training set. As in the previous experiments, the features are found using just the posts of depressed users. As seen in Table 6, the features capture more meaningful phrases and contextual information, which could be helpful in distinguishing between users of the two classes.

Table 6. Top 50 n-grams in the text for each class.

Type	Top N-grams found
Depression related	‘make uncomfort’,‘support group', 'major depress', 'pretty cool', 'read comment’, 'people friend', 'live parent', 'sever anxiety', ‘person life', 'talk friend', 'time talk', 'blah blah', 'take care', 'suicide thought', 'good reason', 'good point', 'give time', 'ma

	ke joke', 'week feel', 'bad idea', 'phone call', 'person disord', 'good advice', 'family friend', 'hurt feel', 'hope help', 'make depress', 'deal depress', 'sex drive', 'great idea', 'make fun', 'hope find']
Non-depression related	['make life', 'people read', 'birth control', 'move back', 'day life', 'game play', 'read comment', 'hey guy', 'take long', 'work people', 'work day', 'phone call', 'lot money', 'spend money', 'period time'

Trotzek [27], [28] explored the impact of linguistic meta data features for detecting depression on the CLEF dataset. This work used features like usage of personal pronouns, possessive pronouns, and past tense verbs, in conjunction with depression lexicon words for the classification of users. In this experiment, the features explored are word count, use of words like (I, me, mine) and (they, them, their) in conjunction with n-grams for classification.

Table 7. Experiment 4 results using linguistic metadata and n-gram features

Technique	**Model**	**Balanced Accuracy Score**	**F1 score**
N-grams + Linguistic feature	Naïve Bayes	78.66	83.97
N-grams + over sampling	Naïve Bayes	96.05	96.04
N-grams + under sampling	Random Forest	82.04	82.05

Table 7 shows the results obtained by using n-grams and linguistic metadata as inputs features to two different classifiers. The results show an increase in the balanced accuracy score by 26%, when compared to the baseline model used in experiment 1. Thus, we conclude that using these features aids the classifiers in distinguishing between the two groups of users.

4.5 Experiment 5: Feature Creation using LDA Topic Modelling

The aim of this experiment is to create features using an unsupervised topic modeling technique called Latent Dirichlet Allocation (LDA). As discussed previously in Section 2.1, LDA is used to discover latent topics in a set of documents. Once these topics are found using LDA, a feature vector is created for each user which contains the probability of each latent topic to be present in the user document. These feature vectors are then used as inputs to the classifier models.

Topic modelling is an unsupervised technique to learn the latent topics represented by a document or a group of documents. These topics are identified by clustering similar words that appear in the documents. A document may be made up of multiple topics and different documents may vary in the distribution of these topics. Latent Dirichlet Allocation is a topic modelling technique that has been used in the research to identify depression in text [19, 17].

In this experiment, we vary the number of topics discovered in the corpus between 2 and 25 topics to find the optimal number of topics. An optimal and mutually exclusive set of topics is obtained when we find 5 topics.

Table 8. Topics identified for the dataset by LDA modelling

Topic	Top 10 words identified by LDA
1	star war, feel free, long time, tv show, watch movie, good job, mass effect, hope help, good luck, day ago
2	depress anxiety, make feel, feel sad, anxiety depress, diagnose depress, clinic depress, social anxiety, panic attack, mental health, people depress
3	make sense, high school, good luck, pretty good, lot people, long time, play game, answer question, he will, good idea
4	play game, hard time, game play, make sense, feel bad, reason who, real life, make fun, give shit, time play

Table 8 illustrates the topics found by the model. It can be inferred from the table that topic 2 refers to discussion about depression and anxiety. For the classification task, the probability distribution of user documents for these topics is used as a feature vector. The results obtained for two different classifiers are shown in Table 9.

Table 9. Results for experiment using the probabilities of each topic fond using LDA as features to represent each user in the dataset

Technique	Model	Balanced Accuracy Score	F1 score
LDA Topic Modelling	Random Forests	72.41	87.96
LDA + Over Sampling	Random Forests	95.06	94.70
LDA Topic Modelling	Decision Trees	71.61	84.61
LDA + Over Sampling	Decision Trees	92.50	92.27

These results show that when the topics discovered using LDA are used as standalone features, they give good classification results. The balanced accuracy and f1 scores

obtained are reliable. We can thus conclude that the topics used, and their associated probabilities, provide useful features for the classification task.

4.6 Experiment 6: Lexicon features based on depression lexicons

Experiment 6 aims to create a feature representation for user text based on lexicons relating to depression, and to analyze its impact on the identification of depressed users. Depression lexicons have been used to identify text related to depression, anxiety and other mental illnesses by measuring the usage of depression related words in the text. For this experiment, we use an existing depression lexicon [3] which was built on text extracted from Yahoo answers from the mental health category. Feature vectors are created to measure the usage of terms from the lexicon using Tf-idf weights for each user's document.

Extended lexicon using Word Embeddings

Next, we extend the lexicon above by using a word embedding model to find the words most similar to the ones present in the lexicon, and adding them to the list if they are not already present in it. The word embedding itself was created from 15 years' worth of Guardian News articles (2004-2018) downloaded from the Guardian Open Platform API[5]. A *word2vec* word embedding model [29] was trained on this data using the implementation provided by the Gensim package.

Table 10. Experiment 6 results using depression lexicon representation as features

Technique	Model	Balanced Accuracy Score	F1 score
Lexicon + N-grams	Random Forest	80.92	93.12
Lexicon +Over Sampling	Random Forest	96.23	96.24

Table 10 shows the classification result metrics obtained by using depression lexicon as features to represent the user text. There is a 2% increase in the balanced accuracy score and a 10% increase in the f1 score compared to the results of experiment 4 which shows that using a lexicon is an efficient representation of the text.

4.7 Experiment 7: Neural Network model using Word Embeddings

Deep learning techniques for detecting metal illnesses like depression and anxiety have been studied in the recent years [21, 19, 27, 28]. Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have been used for this task. Convolutional Neural networks have generally been used in the machine learning domain for tasks

[5] https://open-platform.theguardian.com/

like image classification and computer vision. Trotzek [27] used CNNs for the task of sentence classification for Reddit posts.

In this experiment, we use a word embedding to represent the text as word vectors and these vectors are then passed through the neural network. The embedding layer receives the text input and passes the vectorized output to a sequence of alternating convolutional and max pooling layers. This is followed by a fully connected dense layer and a softmax layer to predict the user classes. All the hidden layers have RelU active activations. The word embedding used for the experiment is a 100 dimensional GloVe embedding provided by Stanford[6] pre-trained on Wikipedia data. It contains 6 billion tokens and a 400k vocabulary.

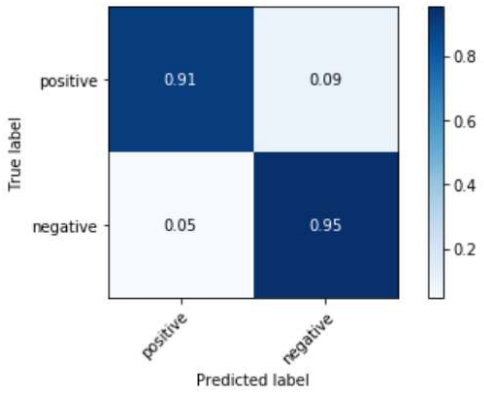

Fig. 3. Confusion matrix for experiment 7 using Deep-Learning model

The results of the experiment are evaluated on a validation set of 264 users after the model is trained on the remaining 70% of the dataset for 50 epochs. The confusion matrix (Figure 3) shows the results obtained on the validation set. The model achieves an overall f1 score of 92.26% and has a good balanced accuracy score of 92.89%. It classifies 95% of the depressed users accurately and classifies 91% of the control group accurately. A breakdown of the metrics is shown in the Table 11. The support mentioned in the table refers to the number of instances of each class in the validation data.

Table 11. Classification scores for experiment 7 using a deep learning model.

Class	Precision	Recall	f1 score	Support
Negative	1.00	0.91	0.95	243
Positive	0.47	0.95	0.62	21

The recall scores are impressive for both classes, as most of the users are correctly classified in each group without being affected by the unbalanced class distribution. It should be highlighted that these scores are obtained without resampling the data to interpolate the minority class. However, the precision score for depressed users' class is

[6] https://nlp.stanford.edu/projects/glove/

low, as only 47% of the users classified as belonging to this class were depressed. Since this tool is generally used as a screening or initial test to identify users who may be at risk, a higher recall and accuracy score would be more useful and can overlook the lower precision score. If this tool was to be used as a last step in diagnosis, the low precision score may not be ideal. However, we would not expect the method to be used in this way. Since the general context of usage is as an initial screening tool, the results from this experiment are rather useful and impressive in distinguishing between the classes of users.

5 Discussion of Results

Table 12. Summary of performance for all models.

Technique	Model	Balanced Accuracy Score	F1 score
Lexicon + N-grams	Random Forest	80.92	93.12
Lexicon +Over Sampling	**Random Forest**	**96.23**	**96.24**
N-grams + Linguistic feature	Naïve Bayes	78.66	83.97
N-grams + Over sampling	**Naïve Bayes**	**96.05**	**96.04**
N-grams + under sampling	Random Forest	82.04	82.05
LDA Topic Modelling	Random Forests	72.41	87.96
LDA + Over Sampling	Random Forests	95.06	94.70
LDA + Over Sampling	Decision Trees	92.50	92.27
Word Embeddings	**ConvNet**	**92.89**	**92.26**

Table 12 summarizes the results obtained from the experiments described in Section 4. The three rows highlighted in the table show the best recorded performances in the experiments conducted. The convolutional network model using word embeddings for text representation gives the best results without having to re-sample the dataset to handle the minority class issue. Since the original dataset is not re-sampled, there is less probability of overfitting the model on the data. Thus, we believe that state-of-the-art techniques, like convolutional networks with word embeddings, can provide the most efficient and useful way of distinguishing between the two classes of users and predicting depression amongst users of the Reddit platform.

6 Conclusion

The utility of social media posts to detect mental illnesses, like depression, anxiety and post-traumatic stress disorder, has been studied in recent research to aid mental health clinicians and researchers. This research requires social media data of users for the purposes of training and building models to distinguish between users who are depressed and those who are not, which presents a question about the user privacy and ethical usage of sensitive data. For this research work, the data used is a publicly available

dataset designed for such research. It contains no personal information about users and neither does it include any of their private messages. Only data used are the actual post text, title and time when the post was made. All users are referenced using a system generated id for each user, not their usernames on the platform. Thus, every effort is made to ensure no personal data gets used for the work done.

The research done in this thesis highlights the fact that deep learning networks with good text representation can outperform classical models using hand crafted features. However, techniques like n-grams, lexicons and topic modelling are useful for analyzing the text, as these features can reveal characteristic behavior of people suffering from depression. This analysis may be useful to experts in the fields of psychology to augment their own expertise.

Our work has highlighted a few key issues present in this area of research. These challenges include distinguishing between characteristic behaviors of bad mood or general sadness from diagnosable mental illnesses like depression and anxiety. Another challenge encountered was distinguishing users who want to create awareness about depression and friends and relatives of people suffering from depression, from actually depressed users. Some users are very active on social media platforms and post very regularly on it. In such cases, the majority of posts may not indicate behavior of depression, but a few significant posts may get lost in the huge number of unrelated posts. To resolve this issue, it might be of use to consider each post as a separate document, rather than considering the entire user history.

The challenge of limited data for users diagnosed with depression was examined in this research, and techniques like resampling the data and using pretrained word embeddings did help resolve the problem to a reasonable extent. Further work could include research on the use of other word embeddings trained specifically on texts relating to mental illness in the form of natural language, like books from authors who overcame depression. Also, the use of sentiment and affect features from LIWC has proved beneficial in the recent research [6]. Integrating that data with the state-of-art deep learning approaches would be an interesting extension of this research.

References

[1] Cacheda F, Fernandez D, Novoa FJ and Carneiro V, "Early Detection of Depression: Social Network Analysis and Random Forest Techniques," *Journal of Medical Internet Research,* vol. 21, no. 6, 2019.

[2] Park Minsu, David W. McDonald and Meeyoung Cha, "Perception differences between the depressed and non-depressed users in twitter," in *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[3] Munmun De Choudhury, Michael Gamon, Scott Counts and Eric Horvitz, "Predicting depression via social media.," in *Seventh international AAAI conference on weblogs and social media*, 2013.

[4] Alessandro Rossi and Mannarini Stefania, "Assessing Mental Illness Stigma: A Complex Issue.," in *Frontiers in psychology vol. 9 2722.*, 11 Jan. 2019.

[5] Glen Coppersmith, Mark Dredze and Craig Harman, "Quantifying mental health signals in Twitter.," in *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality.*, 2014.

[6] Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead and Margaret Mitchell, "CLPsych 2015 shared task: Depression and PTSD on Twitter," in *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2015.

[7] Munmun De Choudhury and Sushovan De, "Mental health discourse on reddit: Self-disclosure, social support, and anonymity.," in *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[8] Stirman Shannon Wiltsey and James W. Pennebaker, "Word use in the poetry of suicidal and nonsuicidal poets," *Psychosomatic medicine,* vol. 63, no. 4, pp. 517-522, 2001.

[9] Rude Stephanie, Eva-Maria Gortner and James Pennebaker, "Language use of depressed and depression-vulnerable college students," *Cognition & Emotion,* vol. 18, no. 8, pp. 1121-1133, 2004.

[10] Reece Andrew G, Andrew J. Reagan, Katharina LM Lix, Peter Sheridan Dodds, Christopher M. Danforth and Ellen J. Langer, "Forecasting the onset and course of mental illness with Twitter data," *Scientific reports,* vol. 7, no. 1, 2017.

[11] Reece, Andrew G and Christopher M. Danforth, "Instagram photos reveal predictive markers of depression," *EPJ Data Science,* 2017.

[12] Bradley Margaret M and Peter J Lang, "Affective norms for English words (ANEW): Instruction manual and affective ratings," in *Technical Report C-1, The Center for Research in Psychophysiology, University of Florida*, 1999.

[13] J. W. Pennebaker and Y. R. Tausczik, "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods.," *Journal of Language and Social Psychology ,* vol. 29, no. 1, pp. 24-54, 2010.

[14] Mitchell Margaret, Jacqui Aguilar, Theresa Wilson and Benjamin Van Durme, "Open domain targeted sentiment," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013.

[15] Maxim Stankevich, Vadim Isakov, Dmitry Devyatkin and Ivan Smirnov, "Feature Engineering for Depression Detection in Social Media," in *Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods(ICPRAM)*, 2018.

[16] Pennington, Jeffrey, Richard Socher and Christopher Manning, "Glove: Global vectors for word representation.," in *Proceedings of the 2014*

conference on empirical methods in natural language processing (EMNLP), 2014.

[17] David M Blei, Andrew Ng and Michael I Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research,* vol. 3, no. Jan(2003), pp. 993-1022, 2003.

[18] Judy Hanwen Shen and Frank Rudzicz, "Detecting anxiety on Reddit," in *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality*, 2017.

[19] Michael M. Tadesse, Hongfei Lin, Bo Xu and Liang Yang, "Detection of Depression-Related Posts in Reddit Social Media Forum," *IEEE Access,* vol. 7, pp. 44883-44893, 2019.

[20] M. McCallum, "A Machine Learning for Language Toolkit," MALLET 15(2), 2002.

[21] Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi and Diana Inkpen, "Deep learning for depression detection of twitter users," in *In Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, 2018.

[22] Inna Pirina and Çağrı Çöltekin, "Identifying depression on Reddit: the effect of training data.," in *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, 2018.

[23] Losada David E. and Fabio Crestani, "A Test Collection for Research on Depression and Language Use," in *In International Conference of the Cross-Language Evaluation Forum for European Languages*, 2016.

[24] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research 16,* p. 321–357, 2002.

[25] H. He, Y. Bai, E. A. Garcia and S. Li, "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning," in *International Joint Conference on Neural Networks (IJCNN 2008)*, 2008.

[26] S.-J. Yen and Y.-S. Lee, "Cluster-based Under-sampling Approaches for Imbalanced Data Distributions," *Expert Systems with Applications,* 2006.

[27] Marcel Trotzek, Sven Koitka and Christoph M. Friedrich, "Word Embeddings and Linguistic Metadata at the CLEF 2018 Tasks for Early Detection of Depression and Anorexia," 2018.

[28] Trotzek Marcel, Sven Koitka and Christoph M. Friedrich, "Linguistic Metadata Augmented Classifiers at the CLEF 2017 Task for Early Detection of Depression," in *CLEF (Working Notes).*, 2017.

[29] T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," in *NIPS'13 Proceedings of the 26th International Conference on Neural Information Processing Systems*, 2013.