# COMP47200 DISSERTATION: LITERATURE REVIEW AND PROJECT SPECIFICATION

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 is a platform to share one's thoughts and emotions presents a rich opportunity to understand the author's state of mind. As such, it can be used to diagnose mental illnesses in users of such platforms and can be a beneficial tool for mental health clinicians worldwide. In this thesis we aim to investigate linguistic latent features in Reddit posts of users which will be used to classify the users into two categories: Depressed and non-depressed users using 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 early detection of mental health problems a significant 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 users/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.

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

The work of Park et al. [2] showed that a healthy individual perceives Twitter as a tool for information sharing and consumption, however depressed individuals perceive it as 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 to raise awareness and decrease the stigma attached to 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. 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. 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.

### 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 [3] 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 can provide insight into their detection based on writings of the author. For instance, Stirman and Pennebaker [4] 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. [5]

examined the linguistic patters in essays written by college students who are currently depressed, were previously depressed or have never been depressed. According to his 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 (SNS) 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 a lot of research in using the UGC to detect onset of mental health diseases like depression, bipolar disorder and Post Traumatic Stress Disorder (PTSD). Reece et al. [6] showed that we can detect first stage depression with an accuracy of 87% accuracy months before its onset. Another work by Reece et al. [7] uses machine learning to analyze image features signaling depression from Instagram posts.

De Chaudhary et al. [8] investigate the potential of using social media platforms to diagnose major depressive disorder in individuals. The data used for the research was collected by crowdsourcing a set of Twitter users who were suffering from clinical depression. Experiments use a machine learning classifier based on linguistic and behavioral attributes like emotion, social engagement, linguistic style, ego network and references to antidepressant medications. They establish that the behavior of a depressed user is characterized by decrease in social activity, raised negative affect, high self-attention, increased medicinal concerns, and heightened expression of religious thought. The study by Coppersmith et al. [9] introduces a novel method for automatic twitter data collection to examine a range of mental disorders like depression, bipolar disorder, PTSD and seasonal affective disorder (SAD). They analyze the content by extracting features using language models, analyzing pattern of life etc. and building classifiers to separate the diagnosed users from the control group. [10] use the CLEF/eRisk 2017 dataset which consists of reddit user data for a certain time period. They focus on developing features using bag-of-words, word embedding and bigrams to classify users into two groups: Risk case of depressed users and non-risk case. They also consider additional stylometric and morphological features and evaluate their applicability.

The research by Hanwen and Frank [11] is primarily focused on detecting anxiety amongst users of the reddit platform, however many of the techniques can also be used to detect depression. They introduce the use of more sophisticated features using vector space embeddings (word2vec), LDA topic modelling and compare the performance with LIWC and n-gram features. For classification they use Linear regression model, Support Vector Machines with a linear kernel and compare its performance with a 2 layer MLP containing 256 hidden units per layer and sigmoidal activations. The research reports an accuracy of 91% using the word2vec feature alone and an accuracy of 98% when combined with lexicon-based features and the Multi-Layer Perceptron.

While most research in this area has been focused on developing features to be used by classifiers, [12] experiments with a deep learning approach where the features are learnt

within the architecture itself. They use two open datasets, CLPsych2015 [13] dataset and the Bell Let's Talk dataset for the experiments which show that CNN- based models with optimized word embeddings give the best results. The research by Trotzek et al. describes results obtained from five models for the CLEF 2017 task [14]. These models use latent linguistic features derived from individual user posts using Latent Semantic Analysis (LSA), bag-of-words, paragraph vectors and a deep learning Long Short Term Memory (LSTM) model.

### 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 use these features to distinguish depressed users from the control group.

The work of De Chaudhary et al. [8] employs social engagement measures like volume of posts, proportion of reply posts and retweets and an Insomnia Index which is defined as the normalized difference in the number of postings made between the day window and night window. 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 [15]. Other features like Linguistic style, Depression Lexicon and antidepressant usage were also engineered for each user. Feature vectors were constructed per user by representing the above mentioned features as a time series containing 4 numbers per measure: the mean, variance, mean-momentum and entropy. The feature vectors are standardized to have zero mean and unit variance.

Coppersmith [9] uses LIWC, language models and measures of pattern of life as features which are subsequently used in the machine learning models: LIWC provides clinicians with a tool for gathering quantitative data regarding the state of a patient from the patient's writing [16]. 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 method to create features for the classifiers is using Language Models. Language models are used to estimate how likely a given sequence of words is. 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 in the 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 terms, presence and valence

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

of sentiment words from a sentiment lexicon [17].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), [10] introduces the use of stylometric and morphological features. The research studies the merits of data representation 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 feature is based on word embeddings. A 100-dimensional word embedding trained on twitter messages (Pennington et al. [18]) 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 hypothesis for using these stylometric features is that persons 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 using Latent Dirichlet Allocation (LDA). LDA is a generative technique used for discovering abstract topics from bodies of texts. In this model, text is perceived as a blend of latent topics, each topic being characterized as a cluster of similar words (Blei et al. [19]). Hanwen [11] 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. [20] use LDA for detecting depression in Reddit posts and 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 [21]. 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 Approach

In the previous section we saw a few studies which employed the use of Multi-Layered perceptron with 1 or 2 hidden layers to model the classification of depressed users. In this section we explore some recent work done to leverage deep learning techniques to distinguish the depressed group from the control group.

Majority of the work involved in the classical machine learning approaches discussed above is in creating cogent features that can distinguish a depressed user based on their behavioral, linguistic or morphological styles. 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 use deep learning models which can interpret these latent features within the network itself and give better classification results.

The work being done in this area is relatively recent. Orabi et al. [12] evaluates the effectiveness of four deep learning models, in which the first three use Convolutional Neural Networks (CNN) and the fourth uses a bi-directional Long Short Term Memory (LSTM) RNN model. These models are built on top of the 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 data we have about depressed users is generally small in size.

### 2.3 Machine Learning Models

**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%<br>Acc-90.7% | [10] |
| | Word2vec, n-grams | Reddit (Anxiety) | Acc – 91% | [11] |
| | LIWC + LDA + bigrams | Reddit [22] | F1 – 91%<br>Acc – 90% | [20] |
| Random Forest | TfIdf + morphology features | Reddit (CLEF eRisk 2017) | F1 – 62.79%<br>Acc-92.01% | [10] |
| | Lexicon, tweet timing, frequency, polarity | Twitter API | Acc – 81.04% | |
| MLP | N-Grams + LIWC | Reddit (Anxiety) | Acc – 98% | [11] |
| | LIWC + LDA + bigram | Reddit [22] | F1 – 93%<br>Acc – 91% | [20] |
| CNN with Max Pooling | Optimized Word Embedding (combining skip-gram and CBOW models) | Twitter (CLPsych 2015) | F1 – 86.96%<br>Acc – 95.1% | [12] |
| BiLSTM | | | F1 – 80.04%<br>Acc – 80.52% | |

## 3 Project Specification

### 3.1 Problem Statement

The research project will seek to explore and investigate the following:

- Automatic early detection of depression in individuals based on user-generated content from online communities.
- Identification of key characteristic features from the data extracted for individuals identified as belonging to a risk case of depression.
- 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

The CLEF early risk prediction dataset (2017) consists of 887 Reddit examples of which 135 are depressed users [23]. Users are 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 in every example, and so do the number of posts per example. Users contain as few as 10 messages or as many as 2000 messages. Overall, 15% of the examples are positive (i.e. they belong to the depression risk case) and 85% are the non-risk control group examples.

**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 |

We aim to use a secondary dataset for the validation of our models. This dataset was built by Inna Pirnia [22] and consists a list of depressed and non-depressed users on Reddit. The corpus contains 1293 depression-indicative posts and 548 standard posts.

### 3.3 Proposed Workflow

For this research thesis, we start with the Reddit datasets mentioned in the section above and preprocess the data by using standard Natural Language Processing (NLP) techniques like tokenizing, stemming and stop word removal. Fig.1 shows the workflow that uses this preprocessed data to classify the users as risk or non-risk.

This processed data can then be used to generate features for each user to understand their social media posts. As a pilot study, we plan to use a simpler set of features that are extracted from the text and evaluate its performance on a baseline model like Decision Trees. Based on the success of the pilot study, we will then generate more complex features to capture latent linguistic features from the data.

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 like random forests, SVM, Ensemble, and MLP neural network. 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.

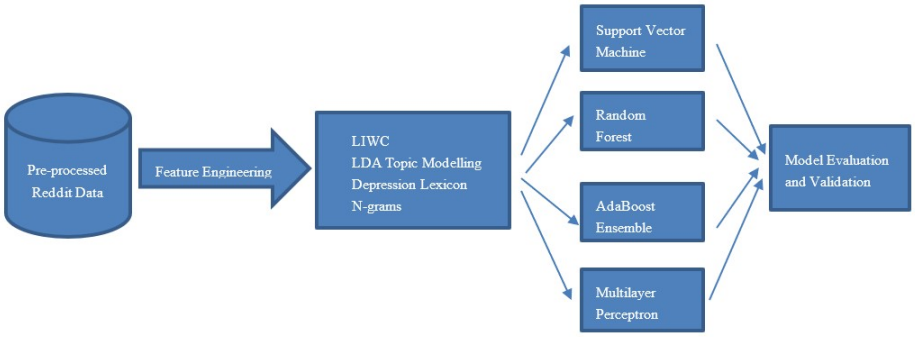


**Figure 1**: Proposed workflow model

### 3.4 Project Milestones

1. Application for ethics exemption to the Taught Masters Research Ethics Committee of the School of Computer Science (TMREC-SCS) – (25.06.2019)
2. Conduct a pilot study on the CLEF dataset, using a simple set of features extracted from the text, and evaluated using a baseline classification model – (02.06.2016)
3. Design of a richer extended feature set extracted from user data – (09.07.2019)
4. Implementation of code to extract the extended feature set – (23.07.2019)
5. Comprehensive evaluation of models built on the extended feature set using a range of classifiers, including experiments relating to feature importance – (30.07.2019)
6. Evaluation of the final model on a second external dataset – (07.08.2019)
7. Completion of final dissertation summarizing all methods and results – (18.08.2019)

## 4 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. In this thesis, we aim to detect depression amongst users of the Reddit platform by extracting useful linguistic and behavioral features from their texts and classifying the users using various machine learning models. We will then design experiments to evaluate the performance of these models and the importance of various features.

## 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 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

[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.