MSc Project - Reflective Essay

Project Title:	**Deep Learning for Detecting Mental Health Disorders using Social Media Generated Content**
Student Name:	**Luke Abela**
Student Number:	**200588919**
Supervisor Name:	**Arkaitz Zubiaga**
Programme of Study:	**Artificial Intelligence**

1. Analysis of Strengths/Weaknesses

1.1. Strengths

- Combined the pre-trained transformer based word embeddings with rule-based data augmentation methods to maximise the performance for a set of binary text classifiers.
- Conducted a comparison of performance between data augmented and non-data augmented methods, and transformer based models and non-transformer based models.
- Explored a common set of deep learning architectures including convolutional neural networks, long-term short-term neural networks and deep averaging neural networks. The pre-packaged classifiers from the transformers were used as a comparable control test.
- Made use of rule-based data augmentation methods to mimic some of the common grammatical errors and variation that exists in social media type text – missing words, misplaced words, use of synonyms, and thus improve performance.
- Tackled a potential multi-class text classification task by implementing a set of binary classifiers for each individual disorder. This overcame the limit in the dataset which provided a single label for each text observation. The implementation of the binary classifiers together could feasibly identify a user potentially displaying symptoms relevant to more than one disorder.

1.2. Weakness

- The manner in which the dataset was used was explicitly assumptive. The text obtained from a specific sub-Reddit was assumed to be representative of that sub-Reddit's disorder but this may not necessarily be the case. Nothing would stop a random user from simply posting a random comment on such a sub-Reddit.
- Each text observation was considered an independent data point. The concatenation of text observations made by the same user could provide a more accurate observation pertaining to that user's specific potential symptoms.
- The dataset was derived exclusively from Reddit. Thus the binary classifiers would be appropriate for Reddit and other Reddit style sites, but may not translate as effectively to Social Media platforms which are not forum based such as Facebook. Whilst Reddit is exclusively a forum posting website, sites such as Facebook, Instagram, and Twitter make use of postings, private messaging, and other services.

2. Presentation of Possibilities for further work

Further work should be considered to improve the work conducted in this study, and to improve the quality of research in the current field:

- Consider a concatenation of any prior comments made by a single user to an observation of text as context. Context would provide a more well-rounded observation.
- Alternative methods of data augmentation could be included to further craft a training dataset which would present more variations of positive cases.
- The black box nature of current mainstream deep learning methods persists in making in-depth questions regarding the actual function of large networks difficult to answer precisely. Combining the resources and methods used in the work presented here with a Hierarchical Attention Network, the benefit of which is that it provides insight to the trained neural network, could represent a path to addressing this issue. (Sekulic & Strube, 2019)
- Maintaining the method and approach but applying it to a hand crafted dataset based on the social media of subjects who have suffered from mental disorders coming forward would represent an intrinsically more reliable dataset than the assumptive dataset derived from Reddit.
- Beyond just textual observations, many users across various social media platforms typically include photos, videos, emojis, and gifs (graphic information format). Taking these additional facets into consideration could allow for a more well-rounded classification. On platforms such as Instagram and Snapchat, each post is required to contain an image, as the image is given more significance and importance than the text, thus to forgo image analysis whilst attempting to apply text classification for such platforms would be suboptimal.
- The realm of application for mental health classification could be expanded through neural net translation. By share of internet users, English users make up over a quarter of the internet, followed by Chinese at just under 20%, and Spanish at a distant 7.9%. Mental Health issues are not limited solely to users of any specific language, thus for users communicating in languages for which there may not be significant data to develop such classifiers, e.g. an English based classifier could be converted to a foreign language such as Maltese[1]. (Johnson, 2020)

3. Critical Analysis of the Relationship between Theory and Practical Work Produced

The salient subject of this dissertation was the classification of potential symptoms in mental health disorders through social media text. To achieve that, the method of deep learning would have to develop a combination of technical and deterministic features to capture the tell-tale signs of mental health through text. This is a very interesting notion as mental health is considered to be a deeply personal and highly subjective thing. Therefore, how could a machine, or a person go about quantifying such ideas?

[1] Maltese is the national language of Malta, a southern European Island in the central Mediterranean sea, and the only Semitic official language of the European Union.

The prevalent authority on psychology is *The Diagnostic and Statistical Manual of Mental Disorders* (DSM–5), the product of more than 10 years of effort by hundreds of international experts in all aspects of mental health. This document describes various mental afflictions and describes the conditions and symptoms necessary for a person to be diagnosed with any singular disorder, effectively defining the disorder.

Let us consider a case study of Depression, one of the mental health disorders considered in this project. The DSM-5 stipulates two main criteria for the diagnosis of depression to be made: a depressed mood (the expression of sad and negative emotions), and anhedonia (the discontinuity of pleasure and interest in things previously enjoyed). Various secondary factors also exist including: sleep difficulties, changes in appetite and weight, poor concentration, fatigue, feelings of worthlessness, and thoughts of suicide or death amongst other things.

The DSM-5 stipulates that to make a diagnosis of depression, a mental health profession with years of training is required, as well as the demonstration of at least 5 symptoms listed. When classifying whether a user is potentially depressed or not based off text posted on social media, it would be naïve to assume that any person who does not immediately mention at least 5 symptoms in a single post is not depressed. Similarly, a person who does not show at least 5 symptoms cannot be labelled as depressed. This raises an issue of precision vs recall. A classifier could feasibly be trained to only label posts mentioning at least 5 recognised symptoms as potentially depressed, thus ensuring high precision, however that would mean labelling others who perhaps tend to be briefer in their words as not depressed, which would compromise recall.

The significance of recall and precision varies from application to application, however, in such a scenario it may not be immediately obvious as to what is the correct approach:

- A high precision, low recall approach: alert those who are most likely to suffer from a mental condition, and do not alert any others
- A medium precision medium recall approach: alert those are most likely to suffer from a mental condition, and alert those who are potentially at medium risk also, with the knowledge some people who probably aren't actually suffering from a mental health condition will also be reached out to in the process.

It should also be noted that being depressed and not being depressed is not strictly a binary matter. The recovery from depression is typically described as a journey, whereby the recognised symptoms of depression may begin to gradually cease in a person over a period of time, and their only content would feasibly feature less markers of potential depression. Similarly, a person sliding into depression may not immediately display sufficient markers to be diagnosed as clinically depressed, but to immediately classify them as not would ultimately be an injustice towards any struggles that are causing to become depressed. Finally, a person's transition into or out of a mental health disorder may not necessarily be linear either, significant improvement could be followed by moderate setbacks or vice versa.

Any task of machine learning/deep learning which attempts to truly capture whether someone has depression or not would have to somehow represent these criteria in a technical manner. The current black-box nature of deep learning means that it would be very difficult to precisely determine what features a classifier would be using to make its judgements. In traditional machine learning, the user typically selects the features, however, deep learning, in which the user does not explicitly select the features, typically significantly exceeds Machine Learning in performance.

4. Awareness of Legal, Social Ethical Issues, and Sustainability

For a topic as sensitive as Mental Health, achieving high performance whilst being answerable and transparent in approach is of great importance, and represents a challenge for the Social Data Science community moving forward. Furthermore, capturing the insight of a mental health professional which was not touched upon in this work, solidifying the idea that this work's purpose is ultimately to serve as an alert or early warning system, and not to ultimately perform a final diagnosis of a user.

The implementation of project such as this one on a social media platform would immediately invite questions regarding ethical conduct. Social media platforms differ in the data in they collect regarding their users. A mental health awareness system would mean that user posts are being assessed by an artificial mechanism, a notion which would might make users immediately uneasy. Platforms such as Facebook have already implemented anti-bullying filters, however, to bring in programs which relate to a person's mental health would represent a further, possibly intrusive step, into the current user experience.

Previously, scraping text from discussion forums was considered to be ethically low risk, due to the lack of interaction or direct contact with social media users – it was commonly used for student projects and inexperienced researchers who were not yet well versed in ethical research issues. There have been incidents in which data, which was thought to be anonymised, was easily de-anonymised and its users identified. (Zimmer, 2010)

Such an incident would naturally lead to concerns amongst users due to potential effects on their employment, social comfort, or the potential use of their data for financial gain by a third party. It is worth noting that data published on internet discussion platforms and social media has effectively been made public, and is of a different nature to clinical data originating from working with a mental health professional. Ford, Shephard, Jones and Hassan discussed the need for separate frameworks for the ethical consideration and handling of clinical data versus data originating from an online source, posted by a content creator. (Ford, et al., 2021)

There are other obstacles to consider in this task. Beyond any ethical issues, there also exist legal barriers to large companies simply harvesting large amounts of data. User privacy issues quickly arise in user data if such data is shared or collected without explicit user permission. The risk may be further exacerbated if data should be leaked to a third party or other internet source when this was not the explicit intention. The General Data Protection Regulation (GDPR) act provides a strict legal framework on such matters.

From a developmental perspective, there is a balance to be struck between the gathering of data for ease of development, experiment and ultimately the maximisation of performance, and respecting the privacy of users. Within the stipulation that a user should be made aware that their data is being used, they should also be given the option as to whether they would like their generated content to be assessed for potential symptoms of mental health disorders or not. Providing transparency on the functionality of a system without explicitly violating intellectual property rights would go a long way towards implementing and maintaining ethical standards in practice.

Finally, there is also the matter of short term and long term shifts in culture and communication. Humans are dynamic, social beings, who more so than ever before are being exposed to new people, changing ideas, and rapid development of society. Languages themselves are evolving, as new words make their way into recognised and reputable dictionaries, and notions such as political correctness move for the use of key

phrases to be reduced and removed, and replaced with more suitable alternatives. To train a classifier of this nature once, implement it, and to stop all further development would be to ignore the fact that humanity and language are constantly changing.

Terminology which is commonplace used today may have not even been prevalent or known to the majority a decade ago. IELTS, a recognised English language examination board, published a list of words and phrases, mostly recently updated in 2020, containing words and phrases such as “broigus”, “bukateria”, and “cancel culture”, which respectively refer to being angry and irritated, a roadside restaurant or street stall with a seating area, selling cooked food at low prices, and a call for withdrawal of support for a public figure, due to an accusation of a socially unacceptable action or comment. (IDP IELTS, 2020)

A classifier of sensitive nature should be retrained at regular intervals, appropriate for its field of application. The relationships between markers recognised by a deep learning / machine learning model for a mental health disorder in the year 2021 may no longer be as strong or relevant in the year 2031, or even 2026 for that matter.

The desire to undertake a project related to mental health stems from a desire to see people in need recognised, and given the help they require and deserve. Having said that, in attempting to reach out to these people through means such as artificial intelligence based text classifiers, care should be taken to effectively not misdiagnose and/or effectively cause more harm they would have had to endure had such a project never been undertaken in the first place.

References

Ford, E., Shepherd, S., Jones, K. & Hassan, L., 2021. Toward an Ethical Framework for the Text Mining of Social Media for Health Research: A Systematic Review. *Frontiers in Digital Health*, 26 January.

IDP IELTS, 2020. *100 New English Words And Phrases in 2020.* [Online] Available at: https://ielts.com.au/articles/100-new-english-words-and-phrases-updated-2020/
[Accessed 25 July 2021].

Johnson, J., 2020. *Most common languages used on the internet 2020.* [Online] Available at: https://www.statista.com/statistics/262946/share-of-the-most-common-languages-on-the-internet/

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).

Zimmer, M., 2010. "But the data is already public": On the ethics of research in Facebook. *Ethics and Information Technology,* 12(4), pp. 313-325.