Nina Rao
CS 4395.001

ACL Paper Summary

The paper I chose to summarize was Improving the Generalizability of Depression Detection by Leveraging Clinical Questionnaires (Nguyen et al., ACL 2022). The authors include:

Thong Nguyen of University of Amsterdam, Amsterdam, Netherlands

Andrew Yates of Max Planck Institute for Informatics, Saarbrücken, Germany

Ayah Zirikly of Johns Hopkins University, Baltimore, Maryland

Bart Desmet of National Institutes of Health, Bethesda, Maryland

Arman Cohan of Allen Institute for AI, Seattle, WA

In their paper, the researchers identify the problem of identifying mental illnesses in a clinical setting using machine learning models. Their claim is that the current automated diagnosis tools face difficulties when applied to formal clinical settings because of poor generalization capabilities and the untrustworthiness of black box models. The first weakness they identified refers to the out-of-domain or OOD generalization problem which is when models lose performance when training and test data differ. The second weakness refers to black box models which are models that provide accuracy but are difficult to understand and interpret the facets of the prediction result. The researchers propose that their method of grounding the model's learning based off the PHQ9, a clinical questionnaire used in depression screening, will improve generalizability while still performing competitively with a standard BERT (Bidirectional Encoder Representations from Transformers) text-classification approach.

This research paper aims specifically to detect depression based off a user's social media posts. The introduction credits the domain of mental health as important but challenging in public health. In the introduction, the authors briefly mention previous work that has been developed to aid users and their corresponding clinicians in identifying symptoms and monitoring behavior through text. While these approaches have seen technical success, it is said that other studies have shown that practically, assessment of depression and suicidal peoples is a difficult job for clinicians, let alone algorithmic tools. Here, the authors present their two-pronged solution. First a Questionnaire model which is directly based off the PHQ9 depression screening questionnaire. The PHQ9 consists of questions that detect nine types of symptoms. These nine symptoms form the basis of nine symptom detection models. The second part is a Depression Detection model which classifies whether a user is depressed based off the patterns identified by the former model. The claim is that this approach succeeds in classification and generalization, unlike BERT.

The authors give credit to related work in the domain of mental health analysis and classification and posit their research as a novel and unique finding. These include numerous datasets including Twitter posts of depressed and PTSD affected users and ReachOut and Reddit posts of high risk and suicidal users. Other datasets which also utilize the PHQ9 questionnaire in some

ways are mentioned though authors claim that the existing usages have either different approaches or objectives. There are studies on language and linguistic patterns in depressed users and a similar tool LIWC (Using Linguistic Inquiry and Word Count) exists to identify depressed language. Finally, it is mentioned that utilization of contextualized embeddings, word vectors built with context in mind, has contributed to improvements in classifier performance. Despite this prior work, the authors emphasize that the Natural Language Processing models of the past still have poorer performance when it comes to generalization. They assert that when different datasets or even different dataset construction rules are used to apply to these models, performance declines. Furthermore, the black box problem remains with these prior works because a model must be explainable for it to have practicality in a clinical setting.

The authors identify three major contributions their research provides. First are thorough pattern sets for identifying the PHQ9 symptoms as well as heuristics for training symptom classifiers. Second are a set of methods with varying levels of constraint for running depression detection based on the nine symptoms. Third is a comprehensive evaluation of each of the depression detection methods that are tested and mentioned in the paper.

Subsequently, the paper delves into the methods used, firstly pattern-based methods that make classification decisions on the presence of a symptom pattern corresponding to the PHQ9. As mentioned earlier there is a Questionnaire model which acts as a pattern matcher. By crossing user posts and symptom patterns, a binary pattern matching matrix is produced. An index corresponds to a user's post and a symptom. A 1 in this index means that there is a match, or that this post exhibits the symptom. The Depression model receives the matrix produced by the Questionnaire model as input and has two variations. One is count based and produces a symptom score, the other is based on CNN (Convolutional Neural Network) and produces a symptom vector. Then there are classifier-constrained methods. The necessary constraint posed by the researchers is that the questionnaire model is trained on weakly-supervised data in order to produce a model that is based on the PHQ9 questionnaire. They claim that despite this, the model is capable of generalizing beyond the given symptom pattern sets. The Questionnaire model receives BERT token embeddings in order to predict answers for each of the 9 symptoms. Each classifier is a CNN trained on weakly labeled symptom data. Using the BERT token embeddings supplies the model with transfer learning and the weakly labeled data provides important context even apart from the symptom identifying phrase that can be used for further assessment.

Also crucial is a look into two models which the PHQ9 model will be tested against: PHQ9Plus and the unconstrained BERT. The PHQ9Plus is an extension of the PHQ9 with another symptom (neuron) that is trainable end-to-end and can learn other depressive signals apart from the nine PHQ9 symptoms. While this can be a more comprehensive approach, the authors claim there is greater risk for the model taking shortcuts thus limiting its generalizability. The BERT model faces a similar risk. To provide comparison to the PHQ9, this model replaces the questionnaire with just a BERT encoder. While access to raw embeddings produced by the BERT gives the

model more to learn from and thus more signals it can identify, there is again the risk of false signals being identified. For contrast, in the proposed Questionnaire model, these false signals would be ignored as they lie outside of the symptom pattern set.

The experiment consists of the researchers taking their models and the comparison/baseline models and testing each one's performance. The authors conducted their research based on three Reddit datasets, each with a different construction. These datasets are RSDD, eRisk2018, and TRT. For each dataset, the authors measure and report the positive (diagnosed) class and the area under the receiver operating characteristic curve. Here, the heuristics used to determine Positive class and Negative class are also described. Finally, the paper shows the results of the experiment. It is shown that PHQ9 and the pattern-based methods outperform BERT. Though BERT and PHQ9Plus performed the best when given train and test data from the same corpus, this further reinforces that these models were not built for generalizability.

I believe the work done by these resources is important as the capabilities of NLP models grow, an opportunity arises for their integration in other disciplines. In this case, a classification model such as the one described in this paper could be a vital diagnostic tool in the future. Applications lie not only in a clinical setting but also extend to social media algorithms. Identifying a high-risk or suicidal person based off their posts using this generalized model could help the person receive help quicker and potentially save lives.

The authors and the number of their citations on Google Scholar are given below:

Thong Nguyen, 31 citations

Andrew Yates, 2671 citations

Ayah Zirikly, 649 citations

Bart Desmet, 1591 citations

Arman Cohan, 6909 citations

Nina Rao
CS 4395.001

Citations

Thong Nguyen, Andrew Yates, Ayah Zirikly, Bart Desmet, and Arman Cohan. 2022. Improving the Generalizability of Depression Detection by Leveraging Clinical Questionnaires.
In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8446–8459, Dublin, Ireland. Association for Computational Linguistics.