**Summary of “Natural Language Processing of Social Media as Screening for Suicide Risk”**

by: Lukas Martinsson, 980203-9678

**What is the goal of the project described in the paper?**

The goal of the research paper “*Natural Language Processing of Social Media as Screening for Suicide risk”* is to research how efficient a machine learning model could be in predicting suicide attempts by analyzing social media post and how the deployment of such a model could improve current screening methods for suicide attempts. Since the model analyzes private data from individuals the paper further discusses the possible ethical dilemmas of this and illustrates the difficulty in determining a justified privacy-prevention trade-off.

**How is the learning task defined?**

The purpose of the deep learning model is to determine whether an individual will attempt suicide or not. The model analyzes the words and their relationship in a user’s social media to gives an aggregate score that determines whether it will predict that an individual will attempt suicide or not. Since the prediction is discrete with the prediction being either yes or no the model is a classification model. However, since the aggregated score is a continuous value, the paper seems to use the resulting percentages of false positives or true positives to determine what the suitable aggregate score is for the model to predict if an individual will attempt suicide.

**What kind of data is used? How was the data collected?**

The data utilized in this paper is a collection of post from various social media platforms (Facebook, Twitter, etc.). The dataset itself is a collection of data from two different domains. One is OurDataHelps.org which is a website where individuals sign up themselves and, in some cases, their loved ones who have passed away from suicide to allow their social media post to be analyzed to help mental health research. The other dataset is collected from public domains (or the paper implies this at least) from individuals that publicly discusses their previous suicide attempts, and the dates of these attempts. To find a control group in both datasets each individual data is matched with a similar individual concerning ages and gender. However, if the control group is also from OurDataHelps.org or not in the first dataset is unclear.

**Is there any information about how the data is represented?**

It is unclear how the data is represented. However, since the data itself are post on social media which means that it consists of sentences it is probably represented in some form of list.

**What type of machine learning algorithms are used? Is the choice justified by the authors?**

The paper uses a deep machine learning model for the prediction. Since deep learning, as explained by the authors is able to find “complex non-linear relationships between input data” which for the task of understanding the connection between words themselves and a human’s mental state would seem justified. However, whether other possible machine learning algorithms could be used instead was not subject for discussion in this paper. Whether this is due to domain knowledge, i.e., that deep learning is currently regarded as the optimal choice for NLP and hence need no discussion or not is difficult to answer.

**How is the quality of the system evaluated? Is the system compared to any baseline?**

The quality of the system is measured through a 10-fold cross-validation with a ROC curve demonstrating the true positive rate versus the false positives depending on how close to the suicide attempt the model receives its data (in this case it was from 1, 2…,6 months to all available data possible that existed). The ROC curve also illustrates a baseline curve that would depend on chance. However, since this does indicate whether the model itself is viable for real usage it is compared to the rate what clinical diagnosticians would get. The paper notices the difficulty in this comparison, however since the model is significantly better than a clinician the model has a viable use.

**Are the drawbacks of the system discussed? Any particular errors that are discussed in detail?**

The paper noted that the model, whilst highly effective on the test set might not be similarly viable for the whole population. Two main pointes were presented, first that the genders of the majority of the individuals that attempted suicide in the dataset were women. However according to the paper, the model would still get a similar, albeit a bit worse accuracy for men. The second point concerned the age of the dataset. Since the majority of individuals in the dataset were women in the ages of 18 to 24 it is difficult to know if the model generalizes across ages. Because firstly, since the behavior indicating suicide tendencies of various age groups could vary and secondly their interaction with social media might not be similar to that of young women.

**Do the authors have any general take-home message?**

The main issue, not concerning the algorithm precision itself is its implication on privacy. Since it is not a question whether a model like this will be implemented but when. Hence the paper argues that safety issues will have to be address in the authors opinion a combination of government, clinician mental health activists, and so forth should work in tandem for such a model and its deployment would be optimal.

**Do you have any general opinion? Do you think this is the right approach? Do you see any drawback to applying machine learning to this problem?**

Suicide is currently one of the leading causes of death in the developed world among youngsters and hence researching into preventive methods are crucial. As displayed in the research paper the advancement of machine learning gives a possible method of screening individuals that are at risk for suicide attempts. One could argue that a problem with machine learning models is that whilst they might be accurate even the developer cannot always understand why a model gives a certain prediction. I would argue that this however is irrelevant, if there is a proven effective suicide minimization strategy that can be deployed which utilized machine learning it should be used. The only question is the method of its deployment, which is the possible drawback of machine learning in general. Should it be deployed in an opt-in or opt-out fashion and what is a justifiable true positive - false positive tradeoff? For example, testing the majority of the population if it would result in a hundred percent true positive is not viable. Therefore, research into what a viable true positive - false positive tradeoff is would be interesting to read.

*Link to the paper:* [*https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6111391/*](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6111391/)