Progress Report: AI for Social Good (Professor Fei Fang)

Team 14

Prediction of Mental-Health Disorders From Social Media Posts

Why is AI Needed:

Ensuring healthy lives and promoting well-being is one of the United Nations sustainable development goals. By 2030, UN targets to reduce by one third premature mortality from non-communicable diseases through prevention and treatment and promote mental health. Contrary to other health conditions, mental health problems are overlooked and under-addressed. People often neglect to detect early signs of mental health issues. In today's society, social media has become an important means to communicate and spread information. In our paper, we aim to analyze social media data for early detection of mental illness, which could help the identification of individuals suffering from mental illness before they require help from healthcare professionals.

AI Methods:

A lot of work has been done around the mental health of social media users by analyzing their language usage on these platforms. These works have utilized data from platforms such as Twitter, Reddit and Facebook.

Many machine learning and Natural Language Processing techniques have been used previously. Many of these rely heavily on domain-dependent application-specific features and extensive feature engineering. These make the task less generalizable across platforms and does not scale well.

Some more recent works have used Deep Learning techniques for contextual representation and understanding of the texts. They do not require any feature selection and are highly generalizable. Yates et. al (https://arxiv.org/pdf/1709.01848.pdf) were able to outperform previous works using Convolutional Neural Nets as feature extractors on all posts by a user before classifying them as depressed or not.

Since, better techniques have emerged in Deep Learning and Natural Language Processing. Emergence of contextualized embeddings such as BERT, the rise of attention and more recently transformers, all seem to have surpassed older techniques at various tasks. We hope to experiment with these techniques and analyze if any of them can improve the accuracy of the prediction task. We hope that such models may get used for early detection of mental-health issues which affects about 350 million people today (WHO, 2010). [Ref: WHO. 2010. World Health Organization, World Health Statistics 2010. World Health Organization]

Progress towards milestones:

Acquiring Data

We plan to use data from a prior research study, *Depression and Self-Harm Risk Assessment in Online Forums*, written by researchers from Georgetown University. Georgetown University is willing to share data to other researchers who agree to follow usage requirements designed to protect the privacy of subjects in the datasets. The recipient researchers must have the authorized representatives from their institutions sign a specific Data Usage Agreement in order to gain access to the datasets.

We are going through a complicated process to have the right representatives from CMU to sign the DUA for our project. Multiple departments are involved and we need to satisfy various requirements along the process. Because the nature of the project concerns human subjects, we are also in close contact with CMU's Institutional Review Board to be in compliance. Each of our team members has finished the CITI Human Subject and Research training and gained certification. We are hoping to receive the datasets in another one or two weeks.

Architecture

Our work primarily will deals with two important steps when using neural networks for natural language processing:

1. Extract meaningful feature representations (vectors) for each post made by the user.
2. Merge feature vectors from each post for the user, use a pooling operation, and run through fully connected layers to learn classification.

To establish our baseline, we have picked up two research publications that complete the above steps.

1. Kim, Y. (2014). Convolutional neural networks for sentence classification. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. https://doi.org/10.3115/v1/d14-1181

   We have implemented this paper from scratch and achieved benchmarks results. To summarize this work, word embedding for each word is looked up in word2vec pretrained vectors. A 1-D convolution operation over word embeddings is performed followed by max-pooling. The max-pooling operation handles variable input sizes of the sentences. Finally, the output is progressed through a fully-connected layer to perform *N*-way classification.

   (Figure 1 taken from the original paper)

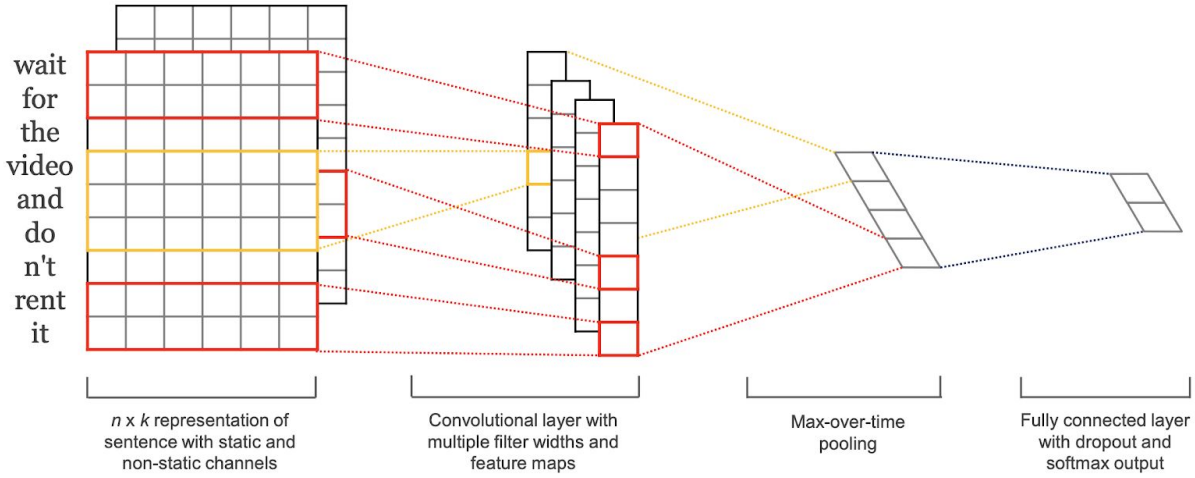

Figure 1: Model architecture with two channels for an example sentence.

2. Yates, A., Cohan, A., & Goharian, N. (2017). Depression and self-harm risk assessment in online forums. *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*. https://doi.org/10.18653/v1/d17-1322

   In this work, vector representations are first computed similar to the work discussed above by Kim et al. Then, all such representations from each post by the user are merged to create a *user representation vector*. Next, these vectors are again passed through N-Dense layers with dropout before classification. This code is publicly available at: https://github.com/Georgetown-IR-Lab/emnlp17-depression

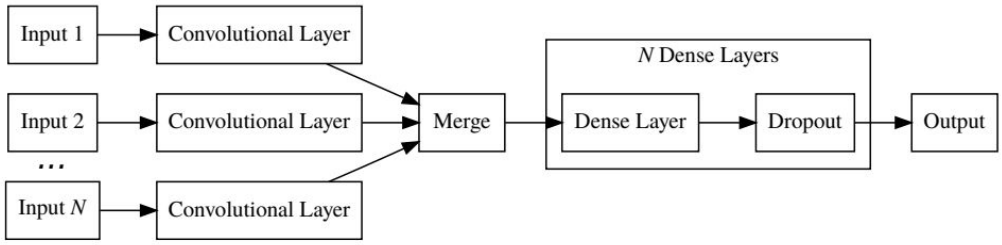

Figure 2: Architecture representing user-representation and classification.

3. *Future Improvements:*

Points above mention how we have established our baseline. In the work that follows, we are planning to use contextualized word embeddings like BERT or ELMo.

Further, we want to experiment with attention-based methods to see if it can improve the classification results.

Alternate data sources

Working on the feedback received on our initial proposal, we anticipated delays in data acquisition and the possibility of not reaching an agreement for acquisition. We looked at multiple other data sources as fallback options. However, given the medical and confidential nature of the issue we are tackling, publicly available large-scale labelled datasets (depressed/not depressed) were hard to come by.

We looked into a Reddit dataset of 3500 Reddit comments, that have been run through a Naive Bayes classifier and labelled Depressed/Not Depressed that we could use as training data for our model. [1] A possible issue with this dataset may be that the size is small for deep-learning models to converge, and we may need to do data augmentation to generate a large train set.

Another approach we decided to take is to run our baseline classifier against similarly labelled data from a different domain which would be Airline tweets labelled as positive, negative and neutral sentiment. This datasource is publicly available and contains 55,000 labelled records. [2] This can allow us to quickly prototype a learning model and pretrain it on the classification task. This pretrain model can then be used on a smaller mental health related labelled dataset and would help with the convergence problem discussed above.

We found another publicly available data source that would predict sentiment through 40,000 training data of text messages. [3] This may be used for a similar pre training or validation objective.

The fourth and final back up data source is a repository of reddit comments (in TB) that we would parse and clean, label Depressed subreddits accordingly and then we would be able to use it. [4] Because of the ancillary tasks that need to be done, we will use this as the last fallback option.

Other updates

In addition to the baseline implementation, we have been in communication with the authors of the original paper asking them for feedback and areas of improvement and they

confirmed our thinking of using BERT for contextualized word embeddings would be a good place to begin improvement. We plan to include this in our future work.

Outline of the Final Report

Our final report includes the following sections:

1. Abstract
2. Introduction & Background
3. Related Work
4. Model Architecture
5. Results & Comparisons
6. Conclusion

Tentative Plan of Next Steps

1. Gather data, clean and tidy, preprocess. [Week of 3/9 and 3/16]
2. Implement baseline model from scratch and establish results. [Week of 3/16 and 3/23]
3. Update the model and experiment with new methods for NLP. [Week of 3/23 and 3/30]
4. Make comparisons between results. [Week of 4/6]
5. Document results and generate visualizations, if applicable. [Week of 4/13]
6. Work on poster creation and presentation. [Week of 4/20]
7. Complete final report. [Week of 4/27 and 5/4]

Distribution Of Work:

Yue: Leading data gathering and preprocessing, fine-tuning model

Alina: Leading baseline implementation

Ashutosh: Leading new techniques implementation

All: Research new techniques, prepare poster, progress report, etc.

References

[1]https://github.com/AshwanthRamji/Depression-Sentiment-Analysis-with-Twitter-Data
[2]https://data.world/crowdflower/airline-twitter-sentiment
[3] https://data.world/crowdflower/sentiment-analysis-in-text
[4]https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/; Paper on the same: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5915669/#ref75