TweetWatch

We'll look into it.™

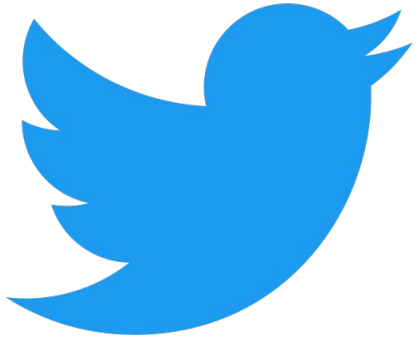
Content Warning

- Mental illness
- Self harm
- Suicidal ideation
- Suicide

The Problem

Over the last decade, many social media services, especially Twitter, have provided an unforeseen and potentially unhealthy outlet for those struggling with serious mental illness.

With young people today becoming increasing desensitized to the frustrated and nihilistic attitudes pervading the internet, it can be easy to completely miss warning signs that somebody may be in serious danger.



http://www.twitter.com/

Specific Examples

A quick search on Twitter quickly confirms this.

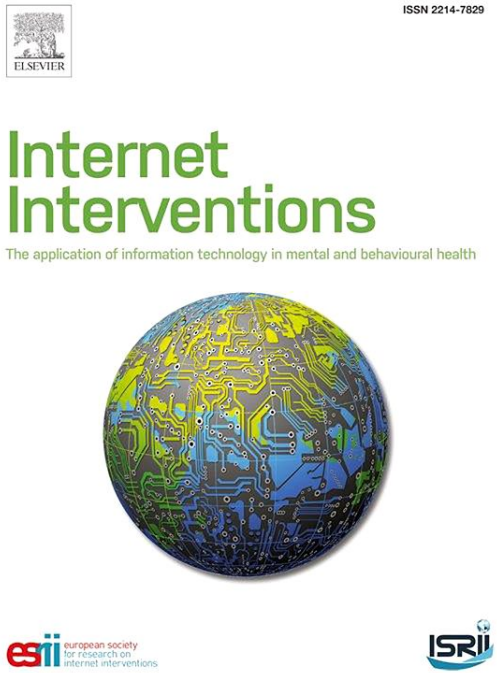
Concealed in plain sight among the archetypical babble of the site, it’s devastating to see how many people are really crying out for help.

http://www.twitter.com/

Our Solution

With modern advancements in machine learning, specifically in regard to text classification, we thought it a natural application of these tools to determine whether a seemingly distressed “tweeter” may be in serious danger.

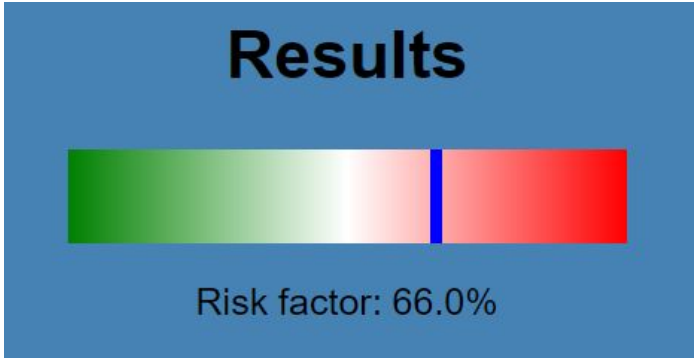
This belief, which has been echoed in recent psychology literature (O’Dea et. al. 2015), was the beginning of the conception of TweetWatch.



The Plan

Our next goal was to determine how best to use an ML model in conjunction with Twitter data to determine the potential risk that a given user exhibited toward themself or others.

We decided to examine the most recent tweets of the user in question, and based on how they were interpreted by the model , assign the user a “Risk Factor” between 0 and 100%.

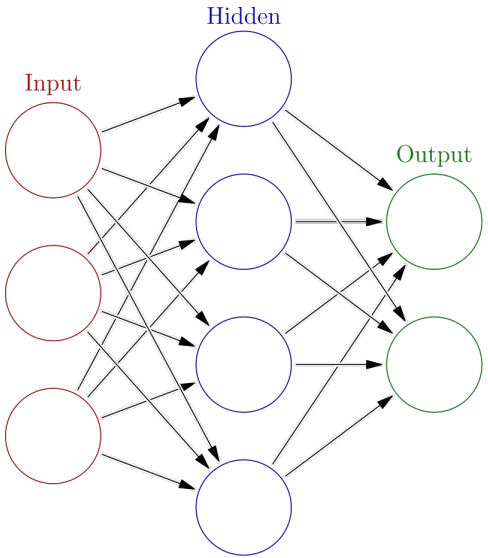


Two Approaches

Regarding the specific ML model that we wanted to implement, we were faced with two primary options.

We could search for a pre-existing, pretrained model that we could shape in a way that would approximately fit our needs.

Alternatively, we could build our own network from scratch, locating relevant datasets and training the model to meet exactly the specifications we would like.

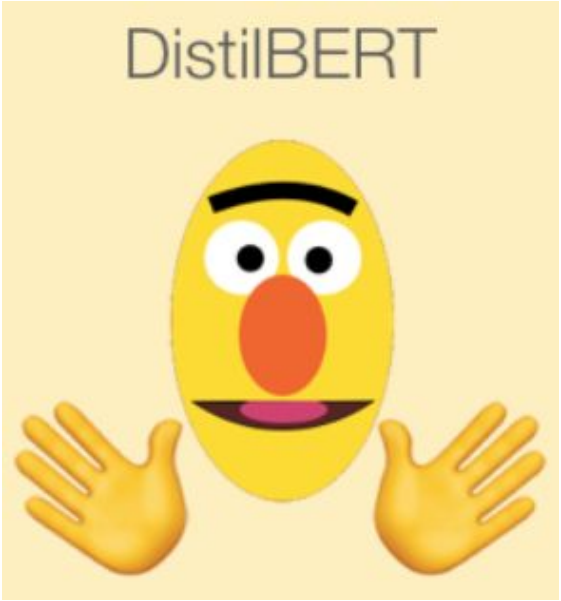


http://en.wikipedia.org/

First Approach: DistilBERT

After some initial research, we found a fine-tuned checkpoint of the DistilBERT, trained on SST-2. This model takes in a text string as input and returns two values: a “positivity” score and a “negativity” score.

Our use case of this model would require using these two scores to determine the aforementioned Risk Factor for a given Twitter user.



http://jalammar.github.io/

Second Approach: Tensorflow

We trained a machine learning model using tensorflow. Our dataset contained approximately 9000 sentences, each labeled with a 1 (corresponding to suicide risk) or a 0 (corresponding to no risk). By creating a one hot encoding of our labels, we trained a neural network using supervised learning. Our neural network had an embedding layer with 10000 neurons, a pooling layer, a dense layer with 16 neurons, and an output layer.

Demonstration

Final Thoughts

The models did a decent job of identifying potentially problematic tweets, but obviously there is plenty more optimization that needs to be done.

One success we found while testing was that if a given user added new tweets that were clearly problematic, their risk factor increased on the next run without fail.

Questions?

Literature Cited

O'Dea, B., Wan, S., Batterham, P. J., Calear, A. L., Paris, C., & Christensen, H. (2015). Detecting suicidality on Twitter. *Internet Intervention*, *2*(2), 183–188.