Exploratory Analysis of Social Media Prior to a Suicide Attempt
Technical Appendix

Glen Coppersmith
Qntfy
glen@qntfy.com

Kim Ngo
Qntfy, University of Notre Dame
kim@qntfy.com

Ryan Leary
Qntfy
ryan@qntfy.com

Anthony Wood
Qntfy
tony@qntfy.com

1 Effects of Semi-supervised Training

The construction of the emotion classifier and the inclusion of a *no emotion* label called for a slightly more complicated training procedure. While the rest of the emotions have explicit labeling via the author employing an emotion word as a hashtag to explicitly label the tweet, the *no emotion* tweets were sampled from those that lack such an explicit emotion label. Since many tweets do convey emotion, yet lack an explicit hashtag indicating the emotion, we expect that this *no emotion* labeled dataset is highly contaminated. Thus, we employed semi-supervised training to remove some of the contamination.

Interestingly, prior to semi-supervised training, F1 was 56, but heavily biased towards misclassifying emotional tweets as *no emotion*. After a round of semi-supervised training, the F1 performance drops to 53, but the mistakes are far more balanced and evenly distributed amongst the labels. We valued the wider distribution of error over the slight dip in performance, and thus used the semi-supervised model for the analysis in the paper. Pragmatically, the trends called out in the Results section hold regardless of which model we used, though the absolute estimated percentage levels of each emotion does change.

Figure 1 shows the effects of semi-supervised training on the confusions made by the model. Unsurprisingly, the majority of the effect of semi-supervised learning can be seen in the improvement of classification of the *no emotion* category. Clearly errors still remain, but they are much more evenly spread across the emotion classes, instead of heavily represented in *no emotion*.

2 Error Analysis of Emotion Classifier

Emotion classification is a decidedly nontrivial task, complicated by two main factors. First, we are examining short statements (i.e., tweets) in isolation of any context. No information about a user's typical usage patterns, nor the tweets that surround it in conversation, nor the possible events the user is reacting to are included in the analysis. Pointedly, the same exact tweet could convey two very different emotions depending upon the context (e.g., see the top of Figure 2). Second, while the labels we apply are categorical and exclusive, many expressions of emotion are not. We see many tweets that are expressing more than one emotion (e.g., see the bottom of Figure 2). Figure 3 shows randomly selected misclassifications (paraphrased to maintain the privacy of the users) to shine light on the sorts of errors the classifier is prone to.

Also worth noting is the bidirectional high confusability of *anxiety* and *fear*. While there are subtle differences in the textbook definition of these two emotions, the Twitter population tends to use them relatively interchangeably. This highlights a limitation of this approach to generating emotion labels – it depends on the psychologically meaningful differences between emotions to be known and properly used by the general population. Depending upon the importance of this confusability, we would suggest either collapsing the two into a single emotional category or to employ a skilled human annotator to provide more direct guidance to the model training.

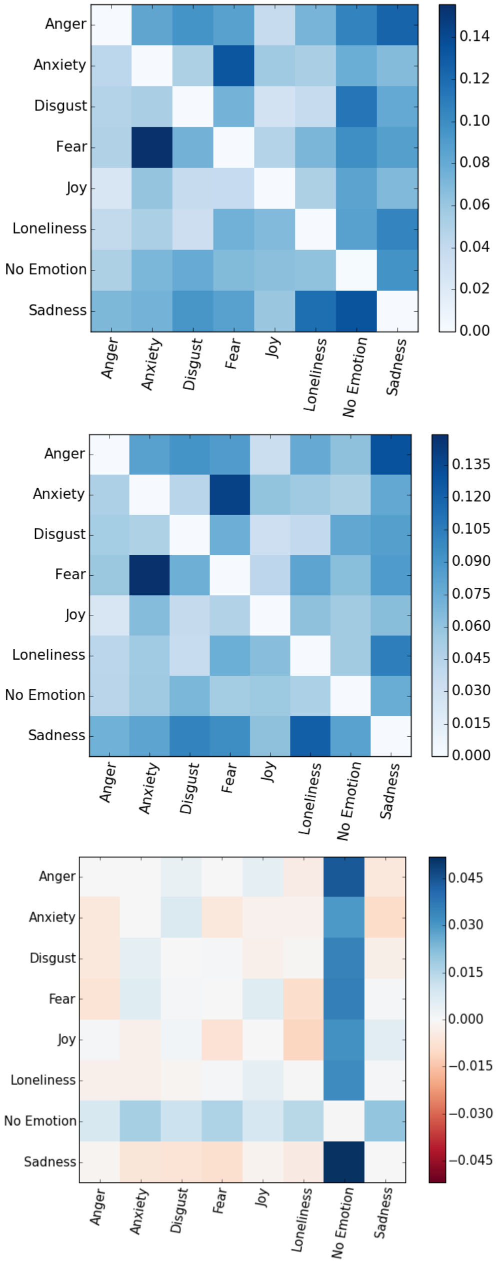

Figure 1: Confusion matrix for emotion classifier without (top) and with (middle) the semi-supervised training applied to *no emotion* tweets. The change between the two methods can be seen in the bottom, with minimal differences between the various emotion categories, and significant differences in the confusion of *no emotion* tweets across the board. The biggest change can be seen where *sadness* tweets are mislabeled as *no emotion* tweets, 5% of the total of *sadness* tweets were no longer mislabeled as *no emotion* after semi-supervised training.

That's been a very long time coming...
No sleep

I'm just pathetic, why can't I deal with those jerks?
No more snapchat for you!

Figure 2: Difficult tweets to classify, even for a human. Top tweets are ambiguous and could convey many emotions depending on the context. Bottom tweets are expressing more than one emotion simultaneously.

anger as *no emotion*
we practice on crappy days and cancel on gorgeous ones.
I need this shift to be done.

anger as *sadness*
could've gone without a certain something told to me.
can I sue a store in bankrupcy with lifetime warranty?

anxiety as *fear*
today's dentist is but a resident, I hope he knows enough
it's been a week since I crashed, driving will be weird

fear as *anxiety*
I hope I can walk after this workout
four days left.

disgust as *no emtotion*
that quarterback is fugly!
@ not always lol!

fear as *no emotion*
spiders? really?
It's raining hard over here in Texas... kinda!

fear as *sadness*
I just freaked out everyone with my phone in my drink!
I thought I ordered a vanilla shake, not a yellow one?

loneliness as *no emotion*
I'm going to cry, I think.
I'm ready for the spouse to come home from work.

loneliness as *sadness*
I messed up, but still feel. can I have 1 more chance?
I got blocked! That's not friendship!

sadness as *loneliness*
I miss my dad, some stranger is standing in his spot.
I wish you were who you are in my dreams.

sadness as *no emotion*
It seems the pontoon boat is gone for good.
Well, that's New York for you.

Figure 3: Fictitious edits of randomly selected misclassifications. For each section, the first emotion is the explicitly tagged label and the second is the emotion classified by the semi-supervised model.