Dictionaries and Decision Trees for the 2019 CLPsych Shared Task

Micah Iserman, Taleen Nalabandian, and Molly E. Ireland
Department of Psychological Sciences, Texas Tech University
{micah.iserman, taleen.nalabandian, molly.ireland}@ttu.edu

Task

For Task A of the shared task, predict the expert-rated level of risk from language used in r/SuicideWatch Reddit posts.

Methods

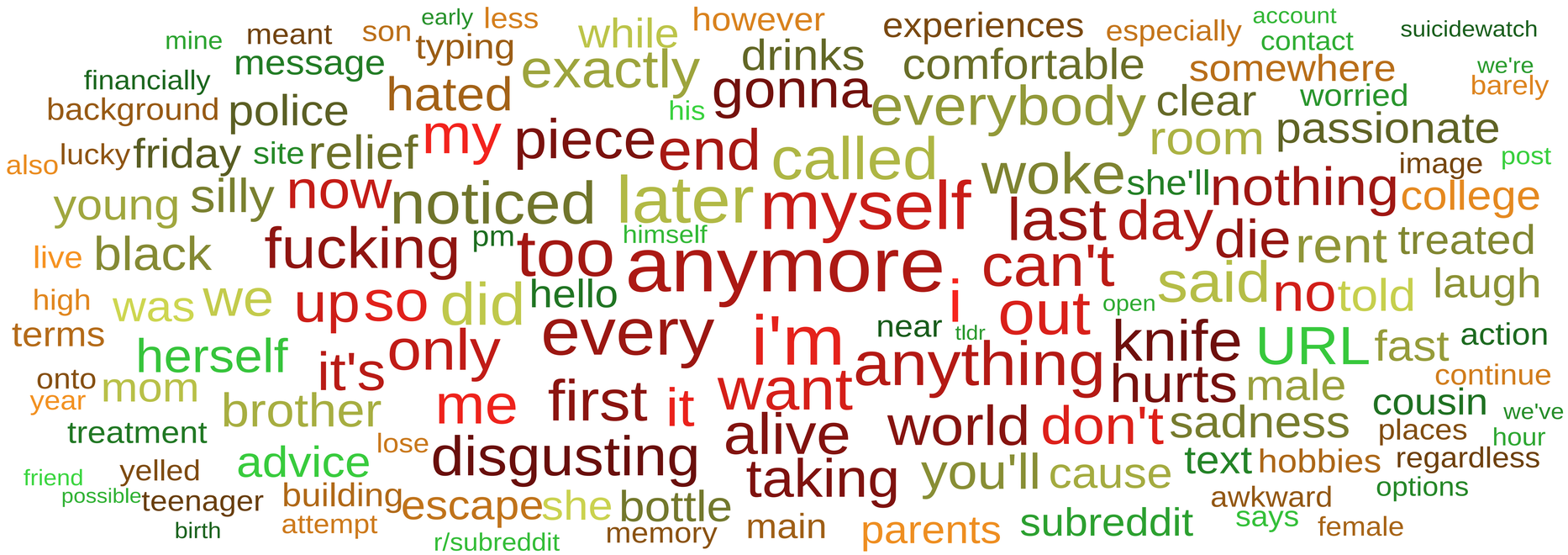
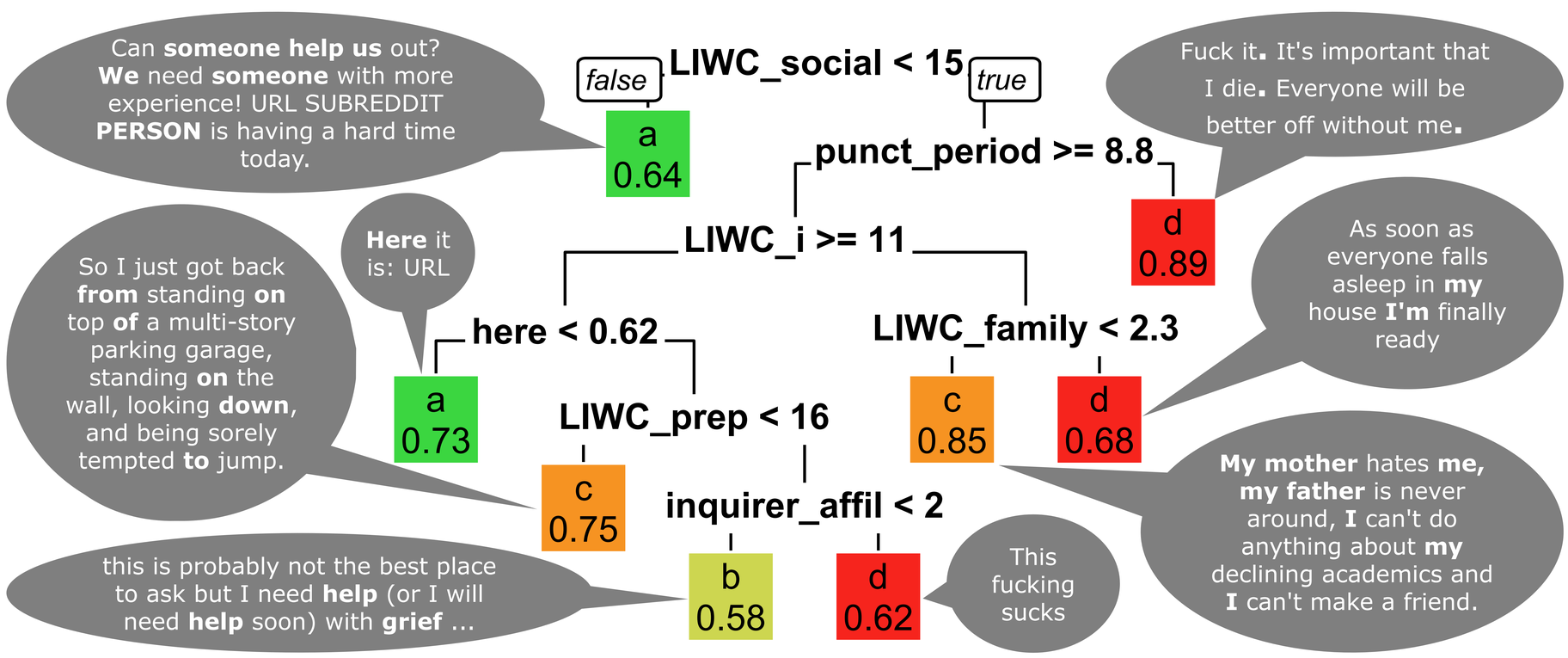
With unigrams and dictionary category scores as features:

1. Correlate each feature with each rating level.
2. Fit a decision tree (in-sample accuracy = .72, F1 = .52).

Results

No risk (a) posts often request advice about, or offer help to others, which was associated with social words (*friend*, *herself*, *advice*), including those relating to the platform (*subreddit*, *post*, *account*) and communication (*contact*, *text*, *message*).

High risk (d) posts were associated with terseness (many short sentences, or short posts with less structure), use of first-person singular pronouns (*I*, *me*, *my*), and indicators of intensity (*fucking*, *disgusting*, *knife*).



1. Color = rating, size = correlation, shade = frequency.
2. Nodes show the dominant rating and its probability within the subset.