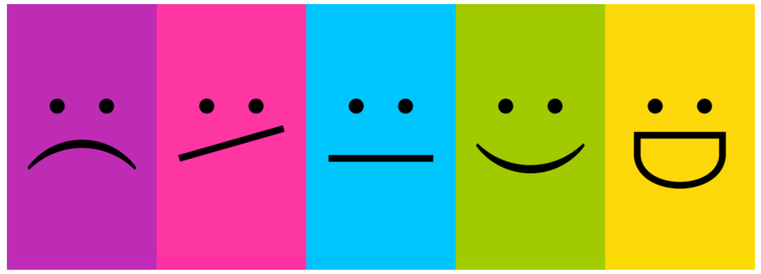


Data Science Collaborative @ BIDS

Berkeley Institute of Design

DeStress Mining Stress Meaning from Social Media

Main Objective

Identify causes, interventions and archetypes for stress and well being

Work in Progress

Succesful word2vec semantic querying for advanced featurization

Re-training classifiers with post-level randomization

Future Work

PCA reduction to 6 universal emotions (discrete emotion classification) → ICA reduction to 2 fundamental dimensions (dimensional emotion classification)

Mechanical Turk automated query result selection for LiveJournal word2vec

Causal inference between Life Events Questionnaire (LEQ) features and stress/mood labels

Data Wrangling

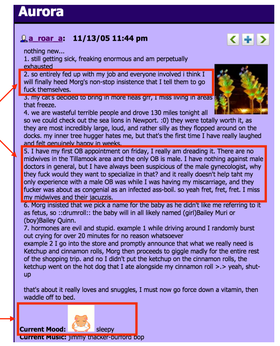
LiveJournal Dataset

Over 1.2 Million Users
More than 64 Million Posts

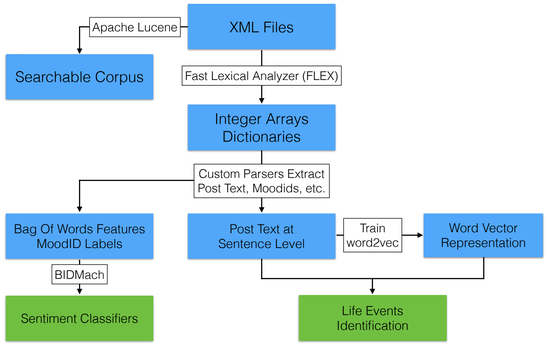
Frank discussions of major life events

Post data:
Colloquial, incomplete, cryptic, misspelled, personal/anonymous

Tagged with self-reported moods



Data Pipeline



Labeling

50% of posts tagged by users with a mood tag

Labels: 132 moodID number representing emotions such as: “happy, “cranky”, “frustrated”, etc.

Features: Bag of words of each post

Classifier: Logistic Regression with L1 regularization

BIDMach: CPU and GPU-accelerated Machine Learning Library

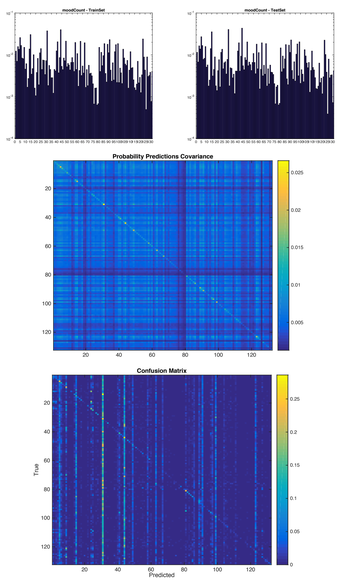
Results: 60% mean Area Under the Curve

Conclusion: Hard task:

i) Too many moodIDs (several are synonyms – not orthogonal);

ii) Post content is very noisy (html, misspellings, slang, etc.)

Potential Solution: Dimensionality reduction (PCA/ICA) & better features (word2vec)



Advanced Featurization

Theory: Life Events Questionnaire (LEQ): Clinical tool with 79 questions about chronic and acute (positive and negative) stress events.

Challenge: Discover LEQ content within post data (colloquial).

Learning Model: Word2Vec · Semantic word representation into vectors

vec(“king”) – vec(“man”) + vec(“woman”) = “Queen”

Model 1: Google News corpus · 300 dimensional

· Reordered vector to match our dictionary → word2vec * dictionary

· Two iterations: 1) Single Word/Phrase → 2) Handpicked Phrases

Model 2: Custom model trained on LiveJournal dataset + user IDs

· Gensim / 300-dimensional / typos, colloquialisms, etc.

· Add user IDs every 3 words → different semantics per user

1st Iteration – Single word/phrase

scala> query("illness", 5)
0.705 -- and attempting to sleep of illness .
0.683 -- did you suffer illness or injury ?
0.680 -- sickness !
0.670 -- they may fake terminal illness , pregnancy , or disease .
0.652 -- my mother who fell ill to a mental illness twice .

2nd Iteration – Handpicked Phrases

scala> multi_query(illness_queries, 7, true, 2)
0.632 -- my father was diagnosed with cancer in early june lung cancer that has spread to the bones .
0.628 -- my mom is in the hospital with a kidney infection .
0.628 -- my brother is sick with bronchitis .
0.625 -- i went to the doctors , and i ended up having double ear infections and really bad strep throat .
0.623 -- he passed away last thursday due to complications with bronchitis in connection to a rare lung disease .

Team: Pablo Paredes, Cory Schilaci, Ana Ferreira, Gene Yoo, Pierre Karaschuk, Dennis Xing,

In collaboration with: Daniel Seita, Prof. John Canny

BIDS Facilitators: Dav Clark, Anthony Wai