Understanding the Posts and Comments on SuicideWatch Subreddit

Lerong Wang

Research Question

- What are some common suicidal risk factors implied by the original posts on SuicideWatch Subreddit?
- How do different linguistic characteristics affect the popularity of a given comment?

Past Literature

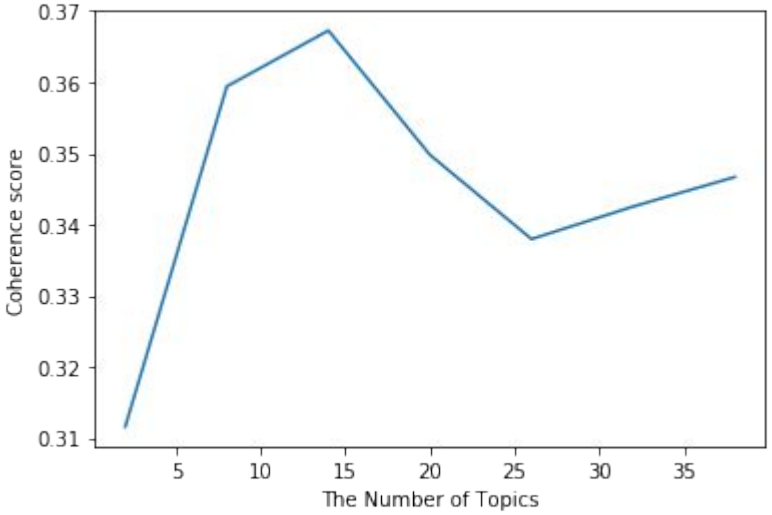
- An emphasis on issues related to depression, anxiety and detecting suicide ideation from social media posts
- A recent study (Grant et al., 2018) performed topic modeling on over 130000 original posts collected from SuicideWatch Subreddit: Word2Vec language models with k-means clustering

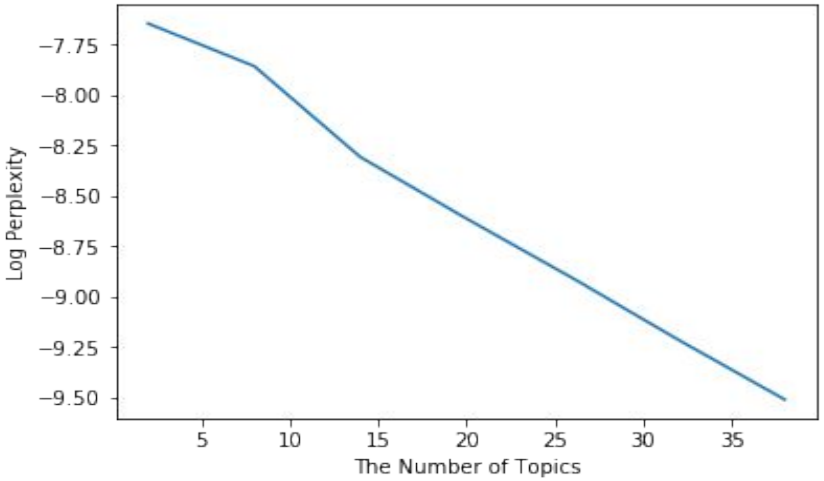
Data

- Google BigQuery
- Post data were collected monthly from April 2018 to October 2018
- 24207 posts in total
- Performed tokenization, lemmatization on the posts using SpaCy
- Comment data were collected from December 2018
- 26967 comments in total
- Used LIWC (linguistic inquiry and word count) to get the semantic categories of words

Methodology

- Topic modeling: Latent Dirichlet allocation (LDA) to discover topics from posts
- Choosing optimal number of topics
  - Coherence Score
  - Perplexity





LDA results

Topic	Terms
topic1	end, die, love, kill, hate, keep, care, ever, always, give
topic2	depression, suicidal, anxiety, deal, scared, afraid, reach, past, struggle, depressed
topic3	mother, father, idk, stab, bill, daughter, stomach, provide, reject, ass
topic4	final, somebody, difficult, steal, helpful, subreddit, peaceful, gut, garage, replace
topic5	attempt, man, entire, mental, survive, vent, physical, hotline, pull, site
topic6	work, job, amp, money, collge, move, pay, able, experience, situation
topic7	sleep, night, wake, eat, bed, eye, drug, drink, morning, doctor
topic8	friend, talk, school, guy, close, relationship, girl, high, stuff, fail
topic9	last, back, leave, still, parent, family, home, well, mom, old
topic10	blood, painful, overdose, wrist, planet, slit, beg, fire, painless, fighting

Terms	Potential risk factors
depression, anxiety, scared, afraid, struggle, depressed	mental health conditions
mother, father, idk, stab, bill, daughter, stomach, provide, reject	relationship problems
attempt, man, entire, mental, survive, vent, physical, hotline	previous suicide attempts
work, job, money, college, pay, able, situation	financial difficulties
sleep, night, wake, eat, bed, eye, drug, drink, doctor	sleeping difficulties
friend, school, guy, close, relationship, girl, high, fail	school difficulties
last, back, leave, parent, family, home, mom, old	family violence/discord
blood, painful, overdose, wrist, slit, beg, fire, painless, fighting	drug abuse disorder

How do different linguistic characteristics affect the popularity of a given comment?

- Outcome variable: comment score = ups - downs
- Choosing independent variables: 93 variables about semantic categories from LIWC
- Choudhury & De (2014). Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity.

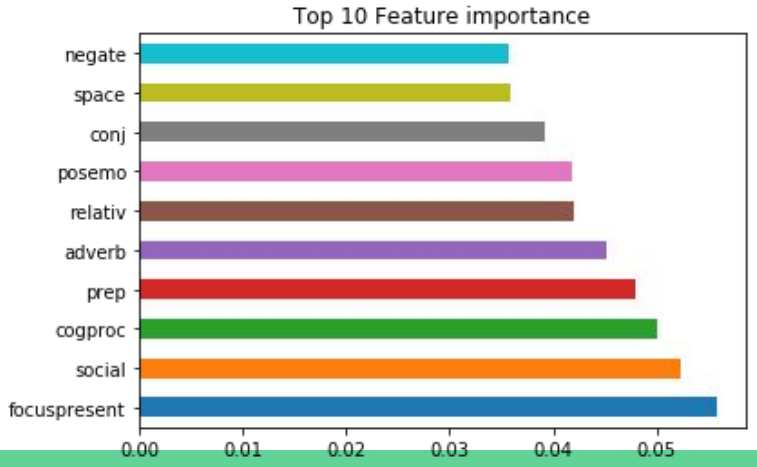
Independent variables			
1st person pronoun	Conjunction	Motion	Sexual
2nd p. pronoun	Death	Negation	Social
3rd p. pronoun	Discrepancy	Neg. emotion	Space
Achievement	Exclusion	Numbers	Swear
Adverbs	Health	Perception	Tense
Assent	Home	Pos. emotion	Tentative
Bio	Inclusion	Preposition	Time
Body	Ingestion	Quantitative	Work
Cause	Inhibition	Relationships	
Certainty	Leisure	Relativity	
Cognitive	Money	Religion	

Linear regression vs. Random Forest vs. Decision Tree

- Linear regression

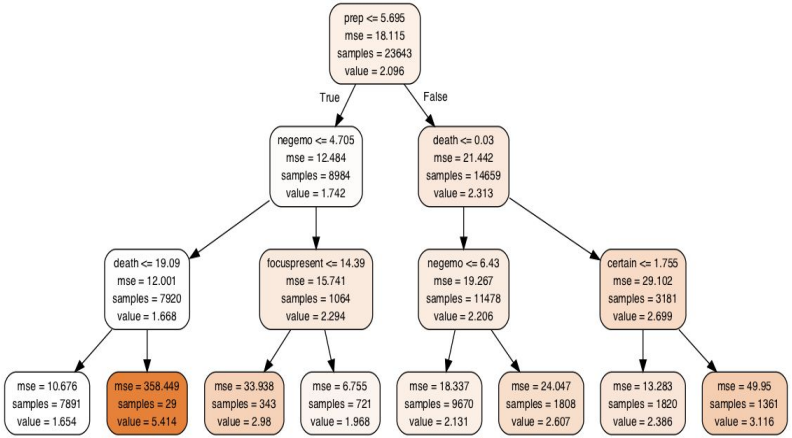
  Significant features: discrepancy, cause, death, negative emotion, preposition, focus on present, focus on past are significant at the level of 0.05

- Random forest



Linear regression vs. Random Forest vs. Decision Tree

- Decision Tree



Model	MSE
Linear Regression	16.128
Random Forest	15.191
Decision Tree	15.226

Future Work

- Conduct PCA on the semantic categories from LIWC
- Cross-validation
- Speed up LDA
- More insights on the interpretations and correlations between topics