Screening for Depression with Retrospectively Harvested Private versus Public Text

ML Tlachac
Data Science

May 7, 2020

About Me: ML Tlachac

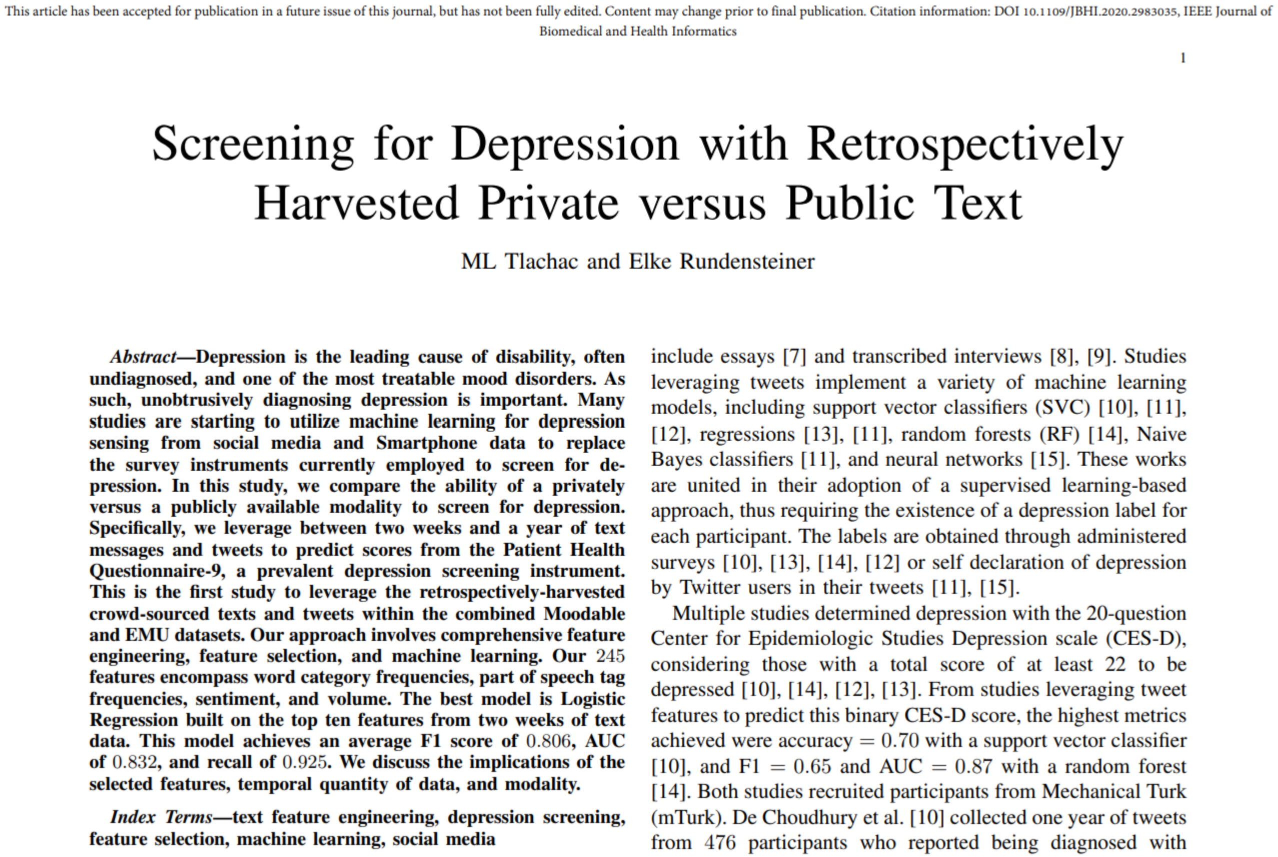
- BS in Applied Mathematics from U Wisconsin- Eau Claire
- 4th year in the combined MS and PhD Data Science program at WPI
- Member of the WPI Data Science Student Council
- First author of 7 accepted health informatics papers
- Favorite hobby is hiking with my border collie Bumper

About You

In the chat, tell me something about you

1. Major(s)
2. Career goals
3. Pets
4. Favorite hobbies
5. Etc

About This Talk



This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.2983035, IEEE Journal of Biomedical and Health Informatics

1

Screening for Depression with Retrospectively Harvested Private versus Public Text

ML Tlachac and Elke Rundensteiner

***Abstract*—Depression is the leading cause of disability, often undiagnosed, and one of the most treatable mood disorders. As such, unobtrusively diagnosing depression is important. Many studies are starting to utilize machine learning for depression sensing from social media and Smartphone data to replace the survey instruments currently employed to screen for depression. In this study, we compare the ability of a privately versus a publicly available modality to screen for depression. Specifically, we leverage between two weeks and a year of text messages and tweets to predict scores from the Patient Health Questionnaire-9, a prevalent depression screening instrument. This is the first study to leverage the retrospectively-harvested crowd-sourced texts and tweets within the combined Moodable and EMU datasets. Our approach involves comprehensive feature engineering, feature selection, and machine learning. Our 245 features encompass word category frequencies, part of speech tag frequencies, sentiment, and volume. The best model is Logistic Regression built on the top ten features from two weeks of text data. This model achieves an average F1 score of 0.806, AUC of 0.832, and recall of 0.925. We discuss the implications of the selected features, temporal quantity of data, and modality.**

***Index Terms*—text feature engineering, depression screening, feature selection, machine learning, social media**

include essays [7] and transcribed interviews [8], [9]. Studies leveraging tweets implement a variety of machine learning models, including support vector classifiers (SVC) [10], [11], [12], regressions [13], [11], random forests (RF) [14], Naive Bayes classifiers [11], and neural networks [15]. These works are united in their adoption of a supervised learning-based approach, thus requiring the existence of a depression label for each participant. The labels are obtained through administered surveys [10], [13], [14], [12] or self declaration of depression by Twitter users in their tweets [11], [15].

Multiple studies determined depression with the 20-question Center for Epidemiologic Studies Depression scale (CES-D), considering those with a total score of at least 22 to be depressed [10], [14], [12], [13]. From studies leveraging tweet features to predict this binary CES-D score, the highest metrics achieved were accuracy $= 0.70$ with a support vector classifier [10], and F1 $= 0.65$ and AUC $= 0.87$ with a random forest [14]. Both studies recruited participants from Mechanical Turk (mTurk). De Choudhury et al. [10] collected one year of tweets from 476 participants who reported being diagnosed with

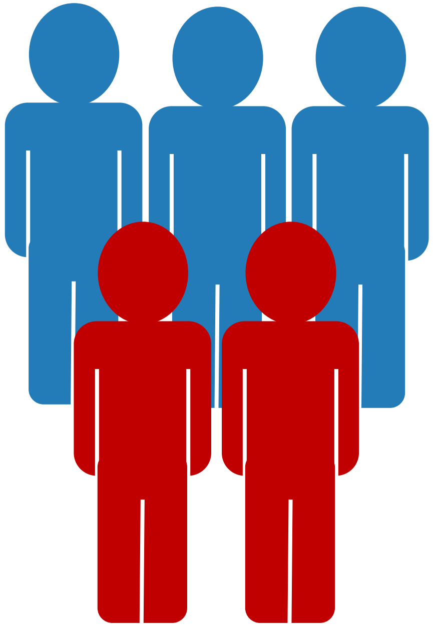
- On my paper accepted to IEEE journal of Biomedical and Health Informatics
- This is my first virtual presentation
- I want this talk to be interactive
  - Ask questions at any time
  - I will be asking for reactions

Depression is Prevalent and Costly

- Depression is prevalent, especially among students
- It takes 11 years on average to get treatment
- Depression is costly
  - $1 trillion/year globally
  - Leading cause of disability
  - 2nd leading cause of death for US adults under 30

2 in 5

graduate students suffer from depression



Evans, Bira, Gastelum, Weiss, Vanderford. "Evidence for a Mental Health Crisis in Graduate Education," Nature Biotechnology, 2018.
National Alliance on Mental Health. "Mental Health By Numbers," 2019.

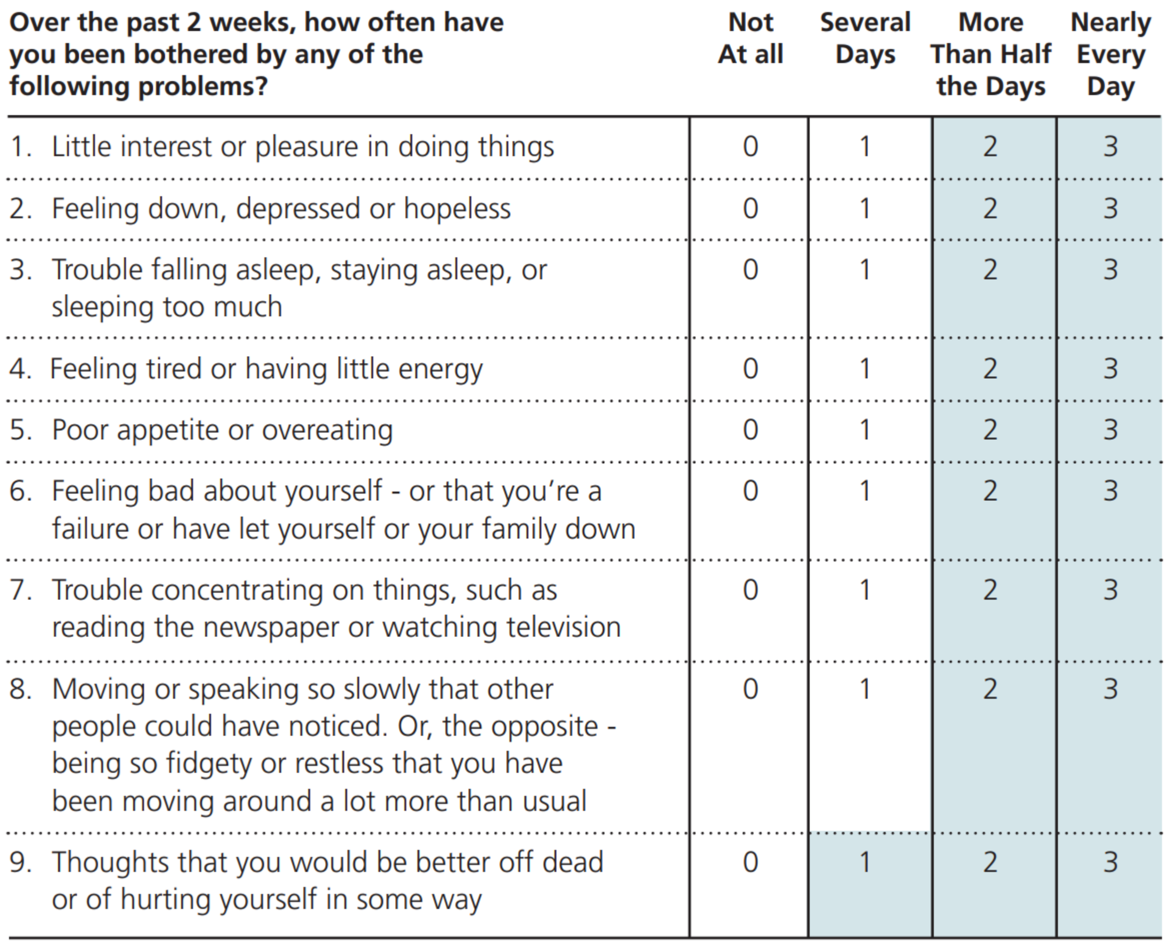
Evolution of Depression Detection

Interview

Survey

Smartphone

Patient-Health Questionaire-9 (PHQ-9)



Over the past 2 weeks, how often have you been bothered by any of the following problems?	Not At all	Several Days	More Than Half the Days	Nearly Every Day
1. Little interest or pleasure in doing things	0	1	2	3
2. Feeling down, depressed or hopeless	0	1	2	3
3. Trouble falling asleep, staying asleep, or sleeping too much	0	1	2	3
4. Feeling tired or having little energy	0	1	2	3
5. Poor appetite or overeating	0	1	2	3
6. Feeling bad about yourself - or that you're a failure or have let yourself or your family down	0	1	2	3
7. Trouble concentrating on things, such as reading the newspaper or watching television	0	1	2	3
8. Moving or speaking so slowly that other people could have noticed. Or, the opposite - being so fidgety or restless that you have been moving around a lot more than usual	0	1	2	3
9. Thoughts that you would be better off dead or of hurting yourself in some way	0	1	2	3

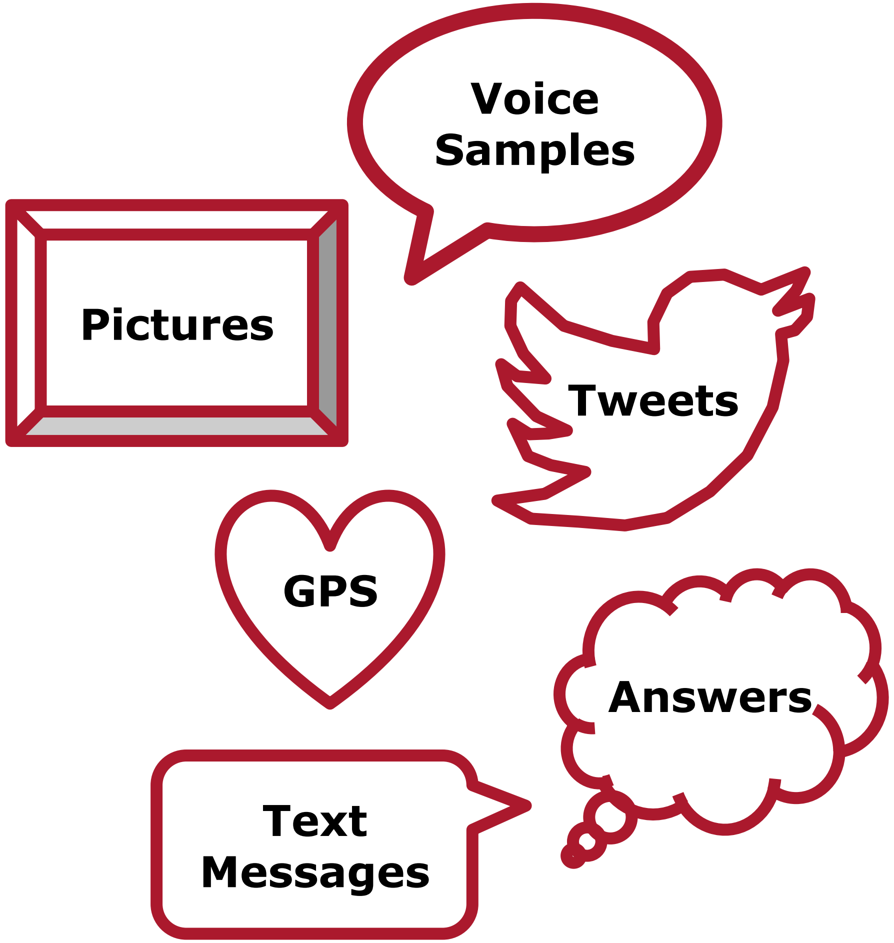
PHQ-9 Score	Interim Diagnosis	Treatment
0-4		
5-9	Symptomatic	Monitor
10-14	Mild Depression	Support or Treatment
15-19	Moderate Depression	Treatment
20-27	Severe Depression	Treatment

http://www.cqaimh.org/pdf/tool_phq9.pdf

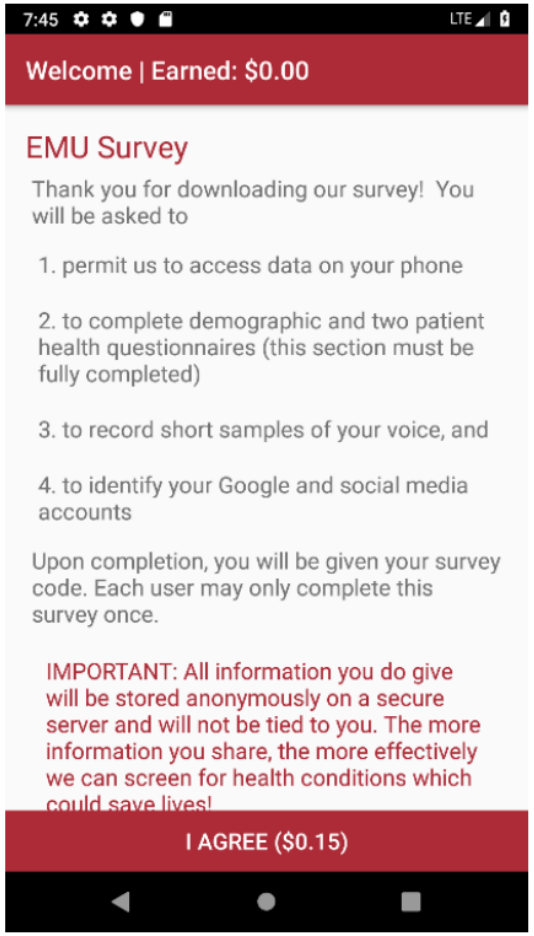
Detecting Depression with Smartphone Data

Machine learning can detect depression from Smartphone sensor data

- Collected Smartphone modalities under WPI IRB
- Used to predict surveys
- Inexpensive and unobtrusive
- Can be unbiased and passive



Data Collection with Mobile App



Welcome | Earned: $0.00

EMU Survey

Thank you for downloading our survey! You will be asked to

1. permit us to access data on your phone

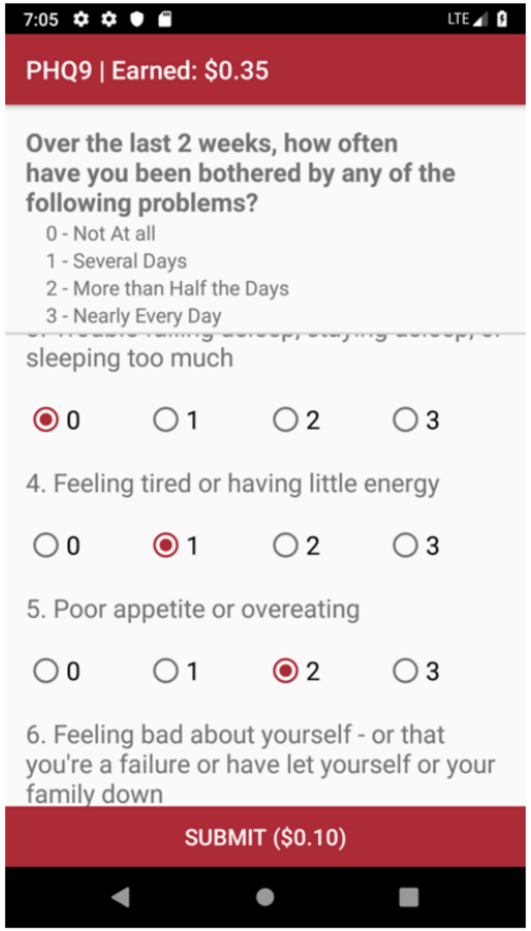
2. to complete demographic and two patient health questionnaires (this section must be fully completed)

3. to record short samples of your voice, and

4. to identify your Google and social media accounts

Upon completion, you will be given your survey code. Each user may only complete this survey once.

IMPORTANT: All information you do give will be stored anonymously on a secure server and will not be tied to you. The more information you share, the more effectively we can screen for health conditions which could save lives!



PHQ9 | Earned: $0.35

Over the last 2 weeks, how often have you been bothered by any of the following problems?

0 - Not At all
1 - Several Days
2 - More than Half the Days
3 - Nearly Every Day

sleeping too much

● 0 ○ 1 ○ 2 ○ 3

4. Feeling tired or having little energy

○ 0 ● 1 ○ 2 ○ 3

5. Poor appetite or overeating

○ 0 ○ 1 ● 2 ○ 3

6. Feeling bad about yourself - or that you're a failure or have let yourself or your family down

My Focus is on the Text-based Modalities

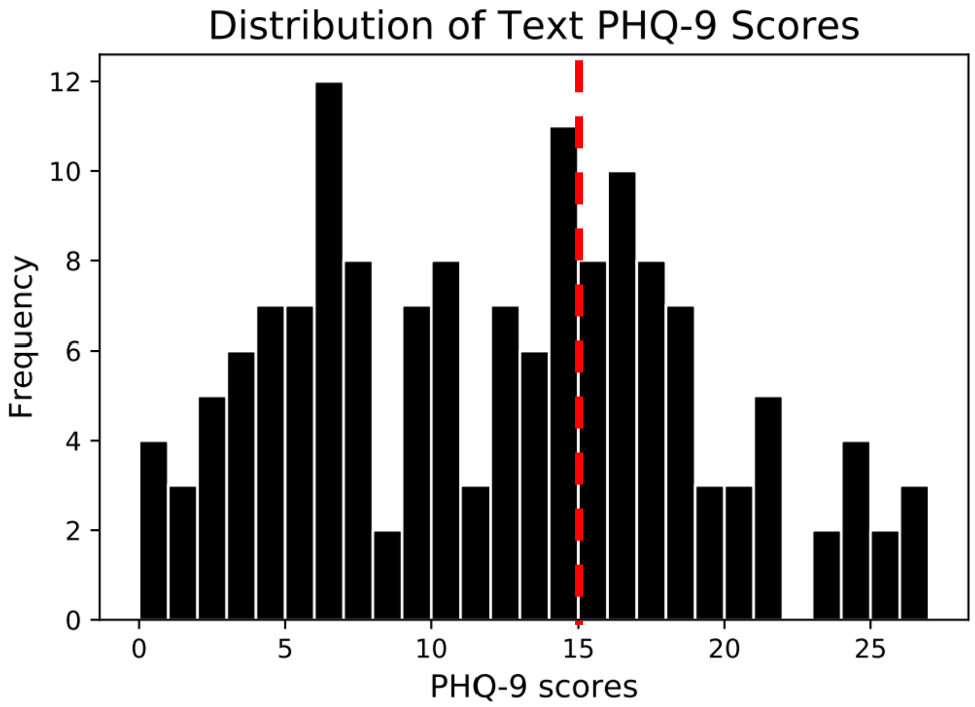
Text Messages

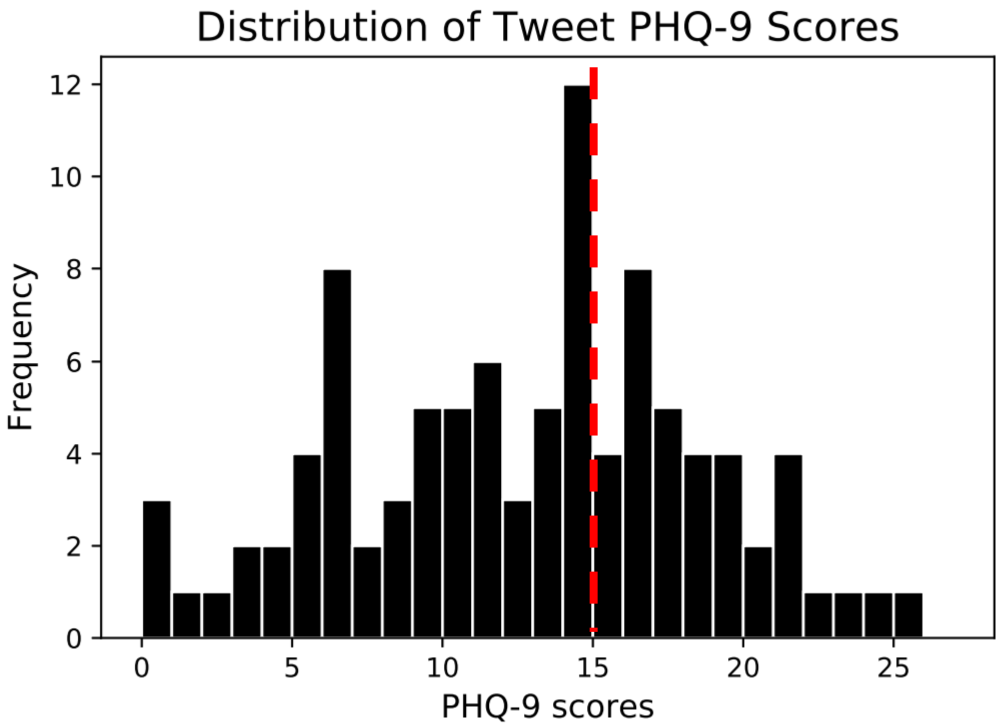
- Private
- More popular
- Less willing to share

Tweets

- Public
- Less popular
- More willing to share

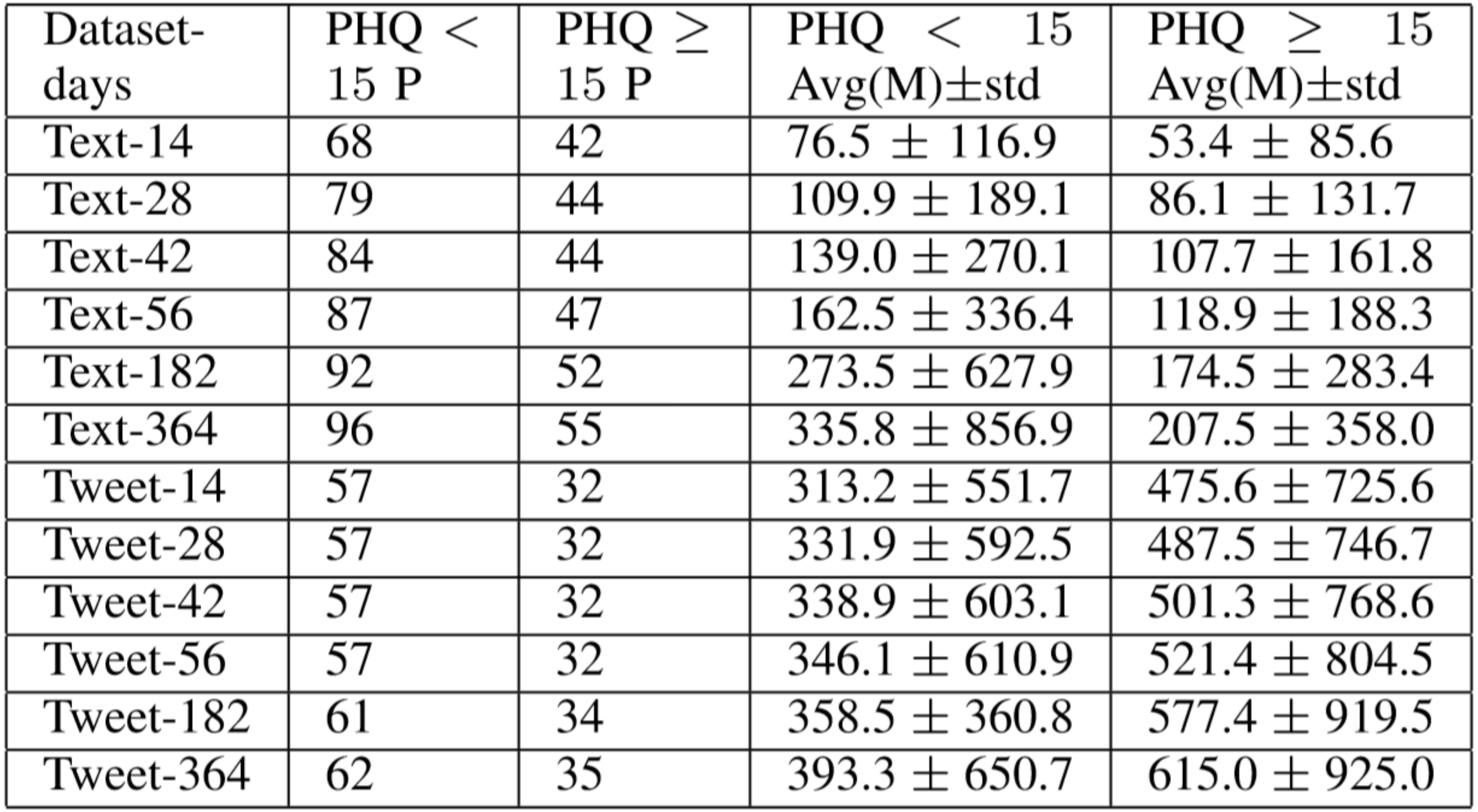
Participant Overview





Datasets

Participants P **Average Number of Messages M**

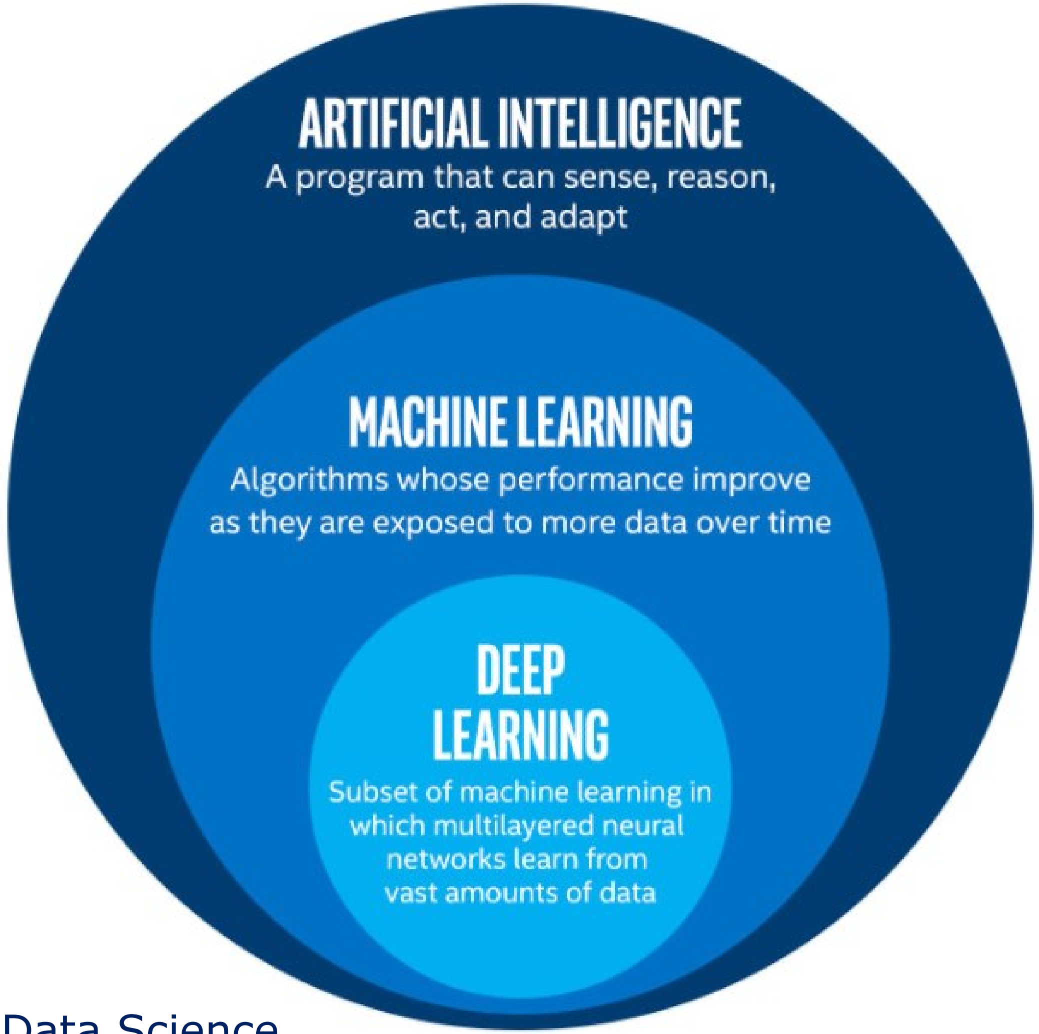


Dataset-days	PHQ $<$ 15 P	PHQ $\geq$ 15 P	PHQ $<$ 15 Avg(M)±std	PHQ $\geq$ 15 Avg(M)±std
Text-14	68	42	76.5 ± 116.9	53.4 ± 85.6
Text-28	79	44	109.9 ± 189.1	86.1 ± 131.7
Text-42	84	44	139.0 ± 270.1	107.7 ± 161.8
Text-56	87	47	162.5 ± 336.4	118.9 ± 188.3
Text-182	92	52	273.5 ± 627.9	174.5 ± 283.4
Text-364	96	55	335.8 ± 856.9	207.5 ± 358.0
Tweet-14	57	32	313.2 ± 551.7	475.6 ± 725.6
Tweet-28	57	32	331.9 ± 592.5	487.5 ± 746.7
Tweet-42	57	32	338.9 ± 603.1	501.3 ± 768.6
Tweet-56	57	32	346.1 ± 610.9	521.4 ± 804.5
Tweet-182	61	34	358.5 ± 360.8	577.4 ± 919.5
Tweet-364	62	35	393.3 ± 650.7	615.0 ± 925.0

Texts

Tweets

Psychotherapy Prefers Machine Learning



ARTIFICIAL INTELLIGENCE
A program that can sense, reason, act, and adapt

MACHINE LEARNING
Algorithms whose performance improve as they are exposed to more data over time

DEEP LEARNING
Subset of machine learning in which multilayered neural networks learn from vast amounts of data

Machine Learning

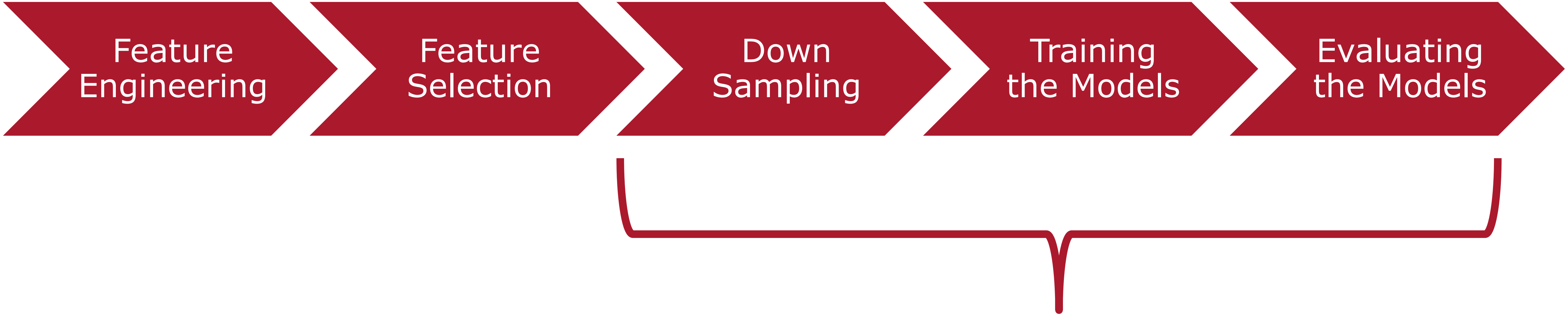
- Can work with limited data
- Requires features
- Interpretable

Deep Learning

- Need large quantities of data
- Can use features or raw data
- Not easily interpretable

Towards Data Science

Experimental Pipeline



Feature Engineering

Feature Selection

Down Sampling

Training the Models

Evaluating the Models

Repeat 100 times

1. Feature Engineering

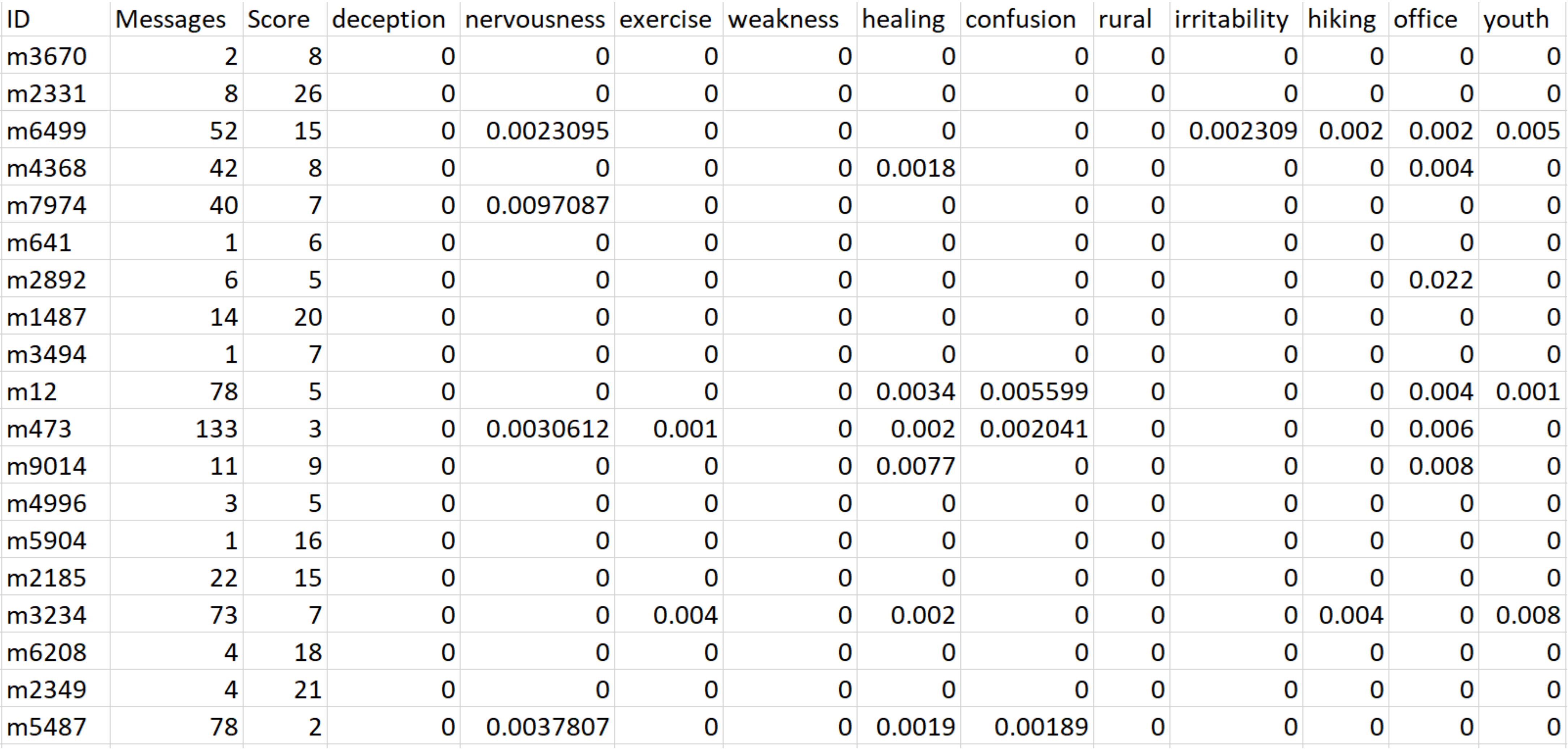
“I am angry. You make me so mad”

Feature	I	Am	Angry	You	Make	Me	So	Mad	%
Anger	0	0	**1**	0	0	0	0	**1**	0.25
Noun	**1**	0	0	**1**	0	**1**	0	0	0.375
Sentiment	0	0	**-1**	0	0	0	0	**-1**	0.25
Polarity	0	**0.5**	**1**	0	**0.5**	0	**1**	**1**	0.5

245 features involving

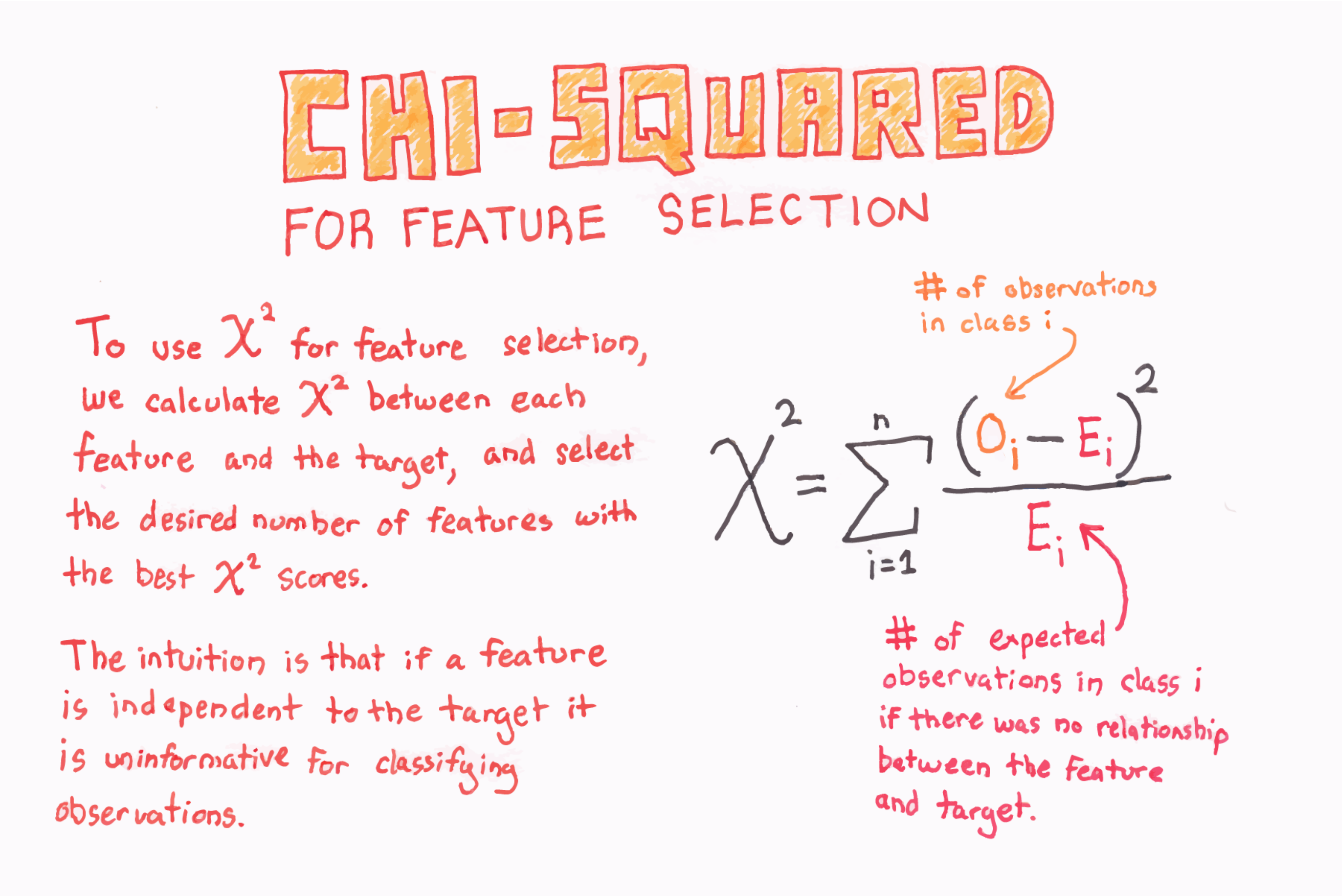
- Word category frequency
- Part of speech frequency
- Sentiment related
- Volume related

1. Feature Example



ID	Messages	Score	deception	nervousness	exercise	weakness	healing	confusion	rural	irritability	hiking	office	youth
m3670	2	8	0	0	0	0	0	0	0	0	0	0	0
m2331	8	26	0	0	0	0	0	0	0	0	0	0	0
m6499	52	15	0	0.0023095	0	0	0	0	0	0.002309	0.002	0.002	0.005
m4368	42	8	0	0	0	0	0.0018	0	0	0	0	0.004	0
m7974	40	7	0	0.0097087	0	0	0	0	0	0	0	0	0
m641	1	6	0	0	0	0	0	0	0	0	0	0	0
m2892	6	5	0	0	0	0	0	0	0	0	0	0.022	0
m1487	14	20	0	0	0	0	0	0	0	0	0	0	0
m3494	1	7	0	0	0	0	0	0	0	0	0	0	0
m12	78	5	0	0	0	0	0.0034	0.005599	0	0	0	0.004	0.001
m473	133	3	0	0.0030612	0.001	0	0.002	0.002041	0	0	0	0.006	0
m9014	11	9	0	0	0	0	0.0077	0	0	0	0	0.008	0
m4996	3	5	0	0	0	0	0	0	0	0	0	0	0
m5904	1	16	0	0	0	0	0	0	0	0	0	0	0
m2185	22	15	0	0	0	0	0	0	0	0	0	0	0
m3234	73	7	0	0	0.004	0	0.002	0	0	0	0.004	0	0.008
m6208	4	18	0	0	0	0	0	0	0	0	0	0	0
m2349	4	21	0	0	0	0	0	0	0	0	0	0	0
m5487	78	2	0	0.0037807	0	0	0.0019	0.00189	0	0	0	0	0

2. Feature Selection



CHI-SQUARED

FOR FEATURE SELECTION

To use χ^2 for feature selection, we calculate χ^2 between each feature and the target, and select the desired number of features with the best χ^2 scores.

The intuition is that if a feature is independent to the target it is uninformative for classifying observations.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

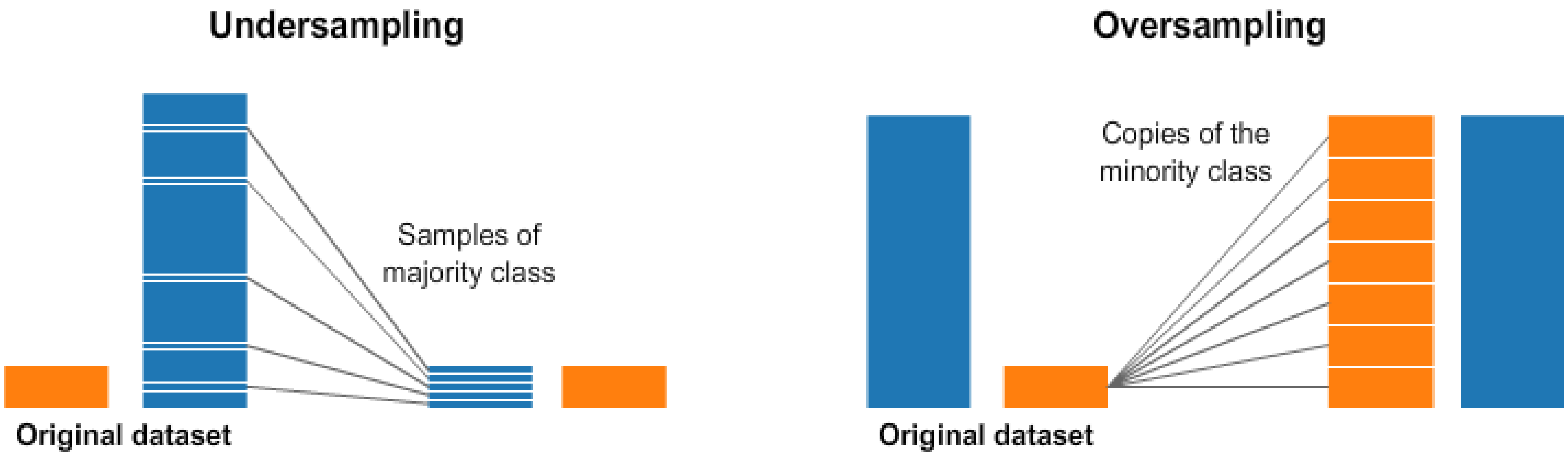
# of observations in class i

# of expected observations in class i if there was no relationship between the feature and target.

chrisalbon.com

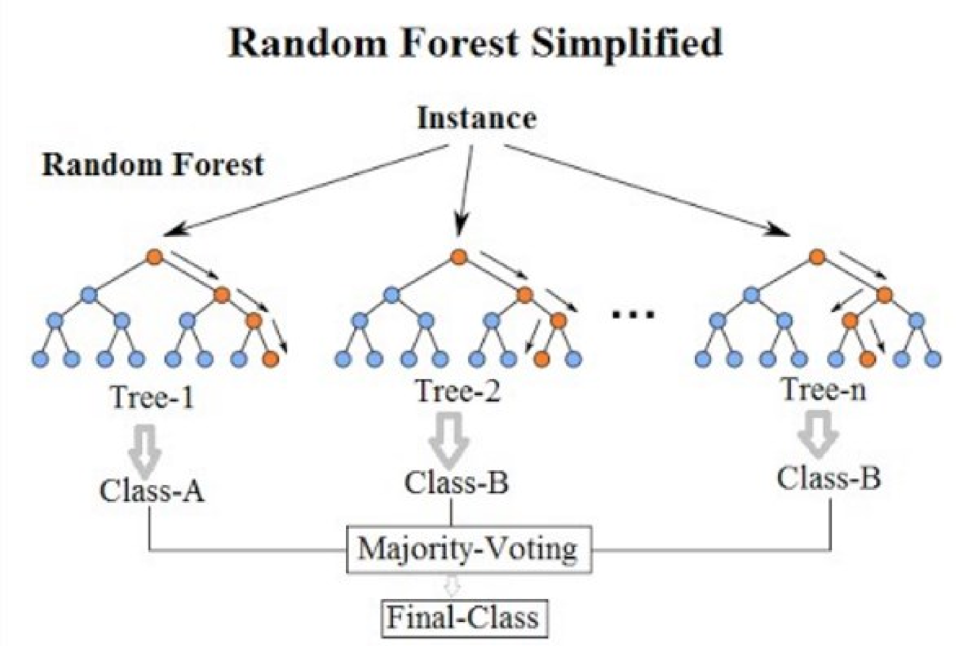
We don't know the ideal number of features, so we look at the f features with the highest chi-squared values for f in 1 to 245

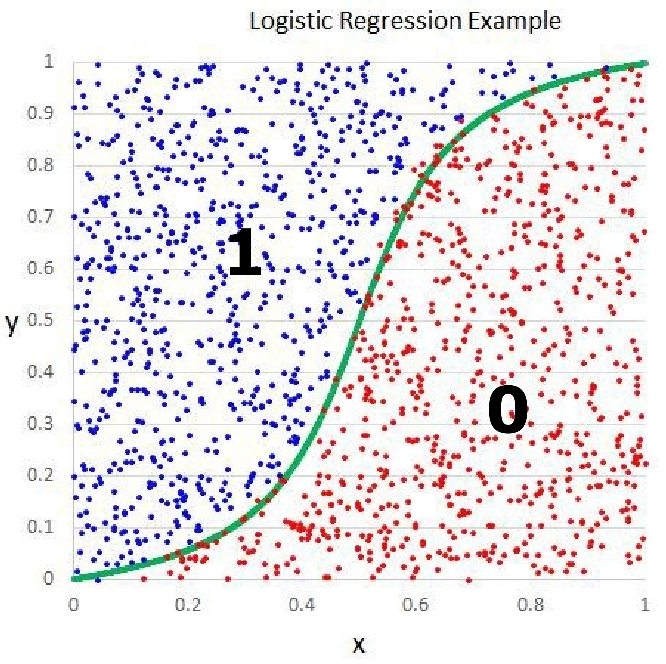
3. Down Sampling

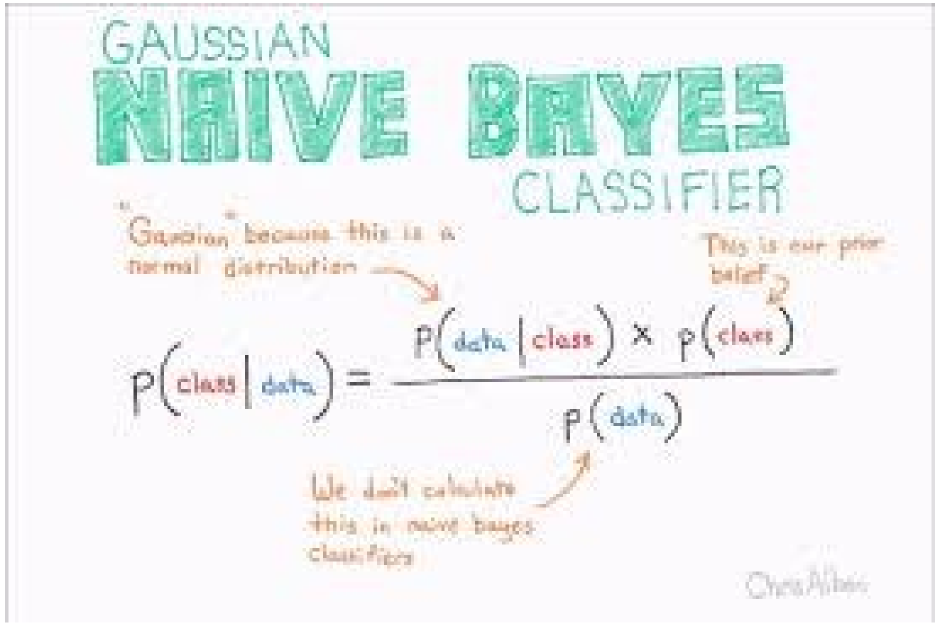


Kaggle.com

4. Training the Models

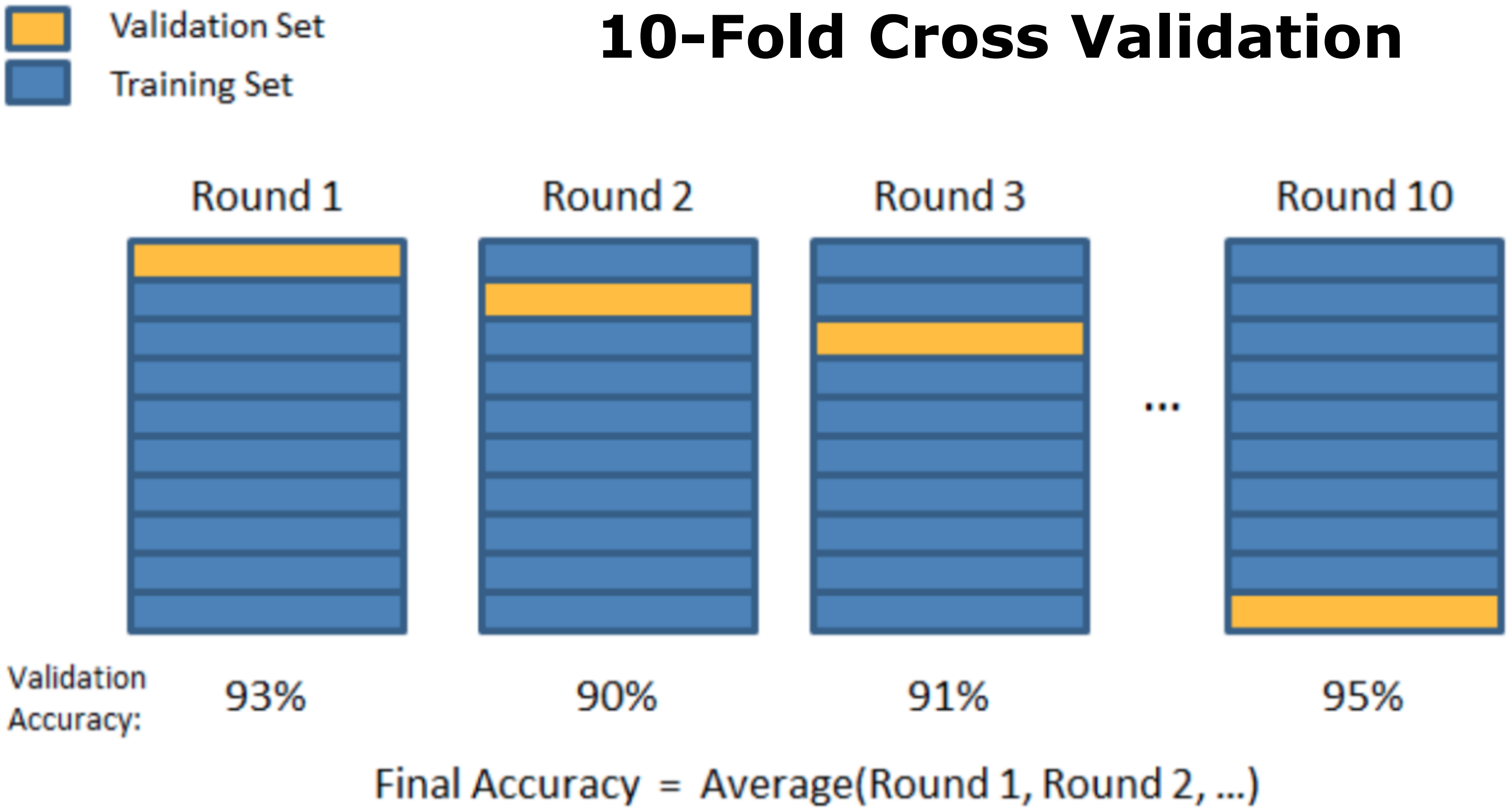





GAUSSIAN
NAIVE BAYES
CLASSIFIER
"Gaussian" because this is a normal distribution
This is our prior belief
P(class | data) = P(data | class) x p(class) / P(data)
We don't calculate this in naive bayes classifiers
Chris Albon

medium.com

5. Evaluating the Models



10-Fold Cross Validation

🟨 Validation Set

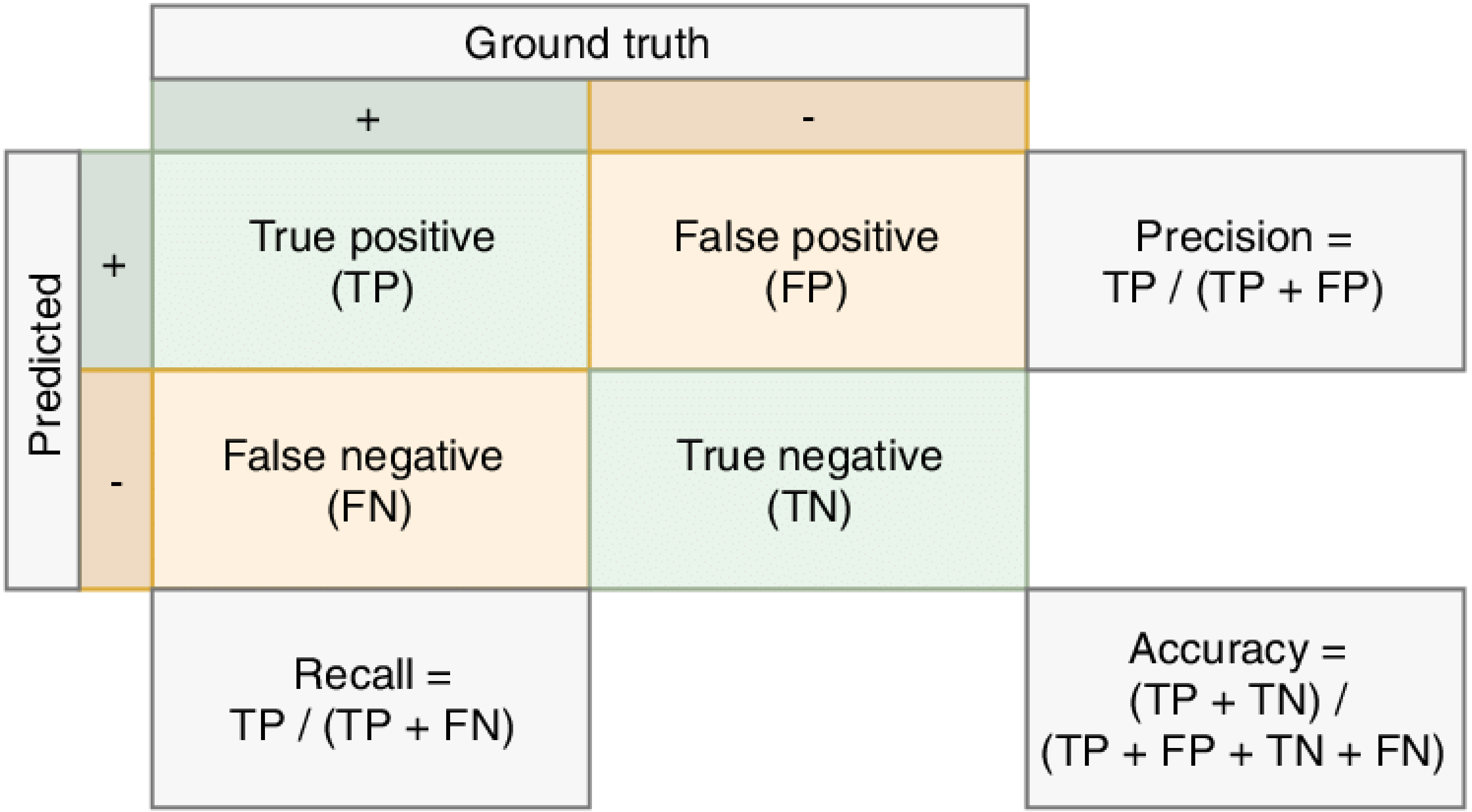
🟦 Training Set

Round 1 | Round 2 | Round 3 | ... | Round 10

Validation Accuracy: 93% | 90% | 91% | 95%

Final Accuracy = Average(Round 1, Round 2, ...)

5. Evaluation Metrics



		Ground truth		
		+	-	
Predicted	+	True positive (TP)	False positive (FP)	Precision = TP / (TP + FP)
	-	False negative (FN)	True negative (TN)	
		Recall = TP / (TP + FN)		Accuracy = (TP + TN) / (TP + FP + TN + FN)

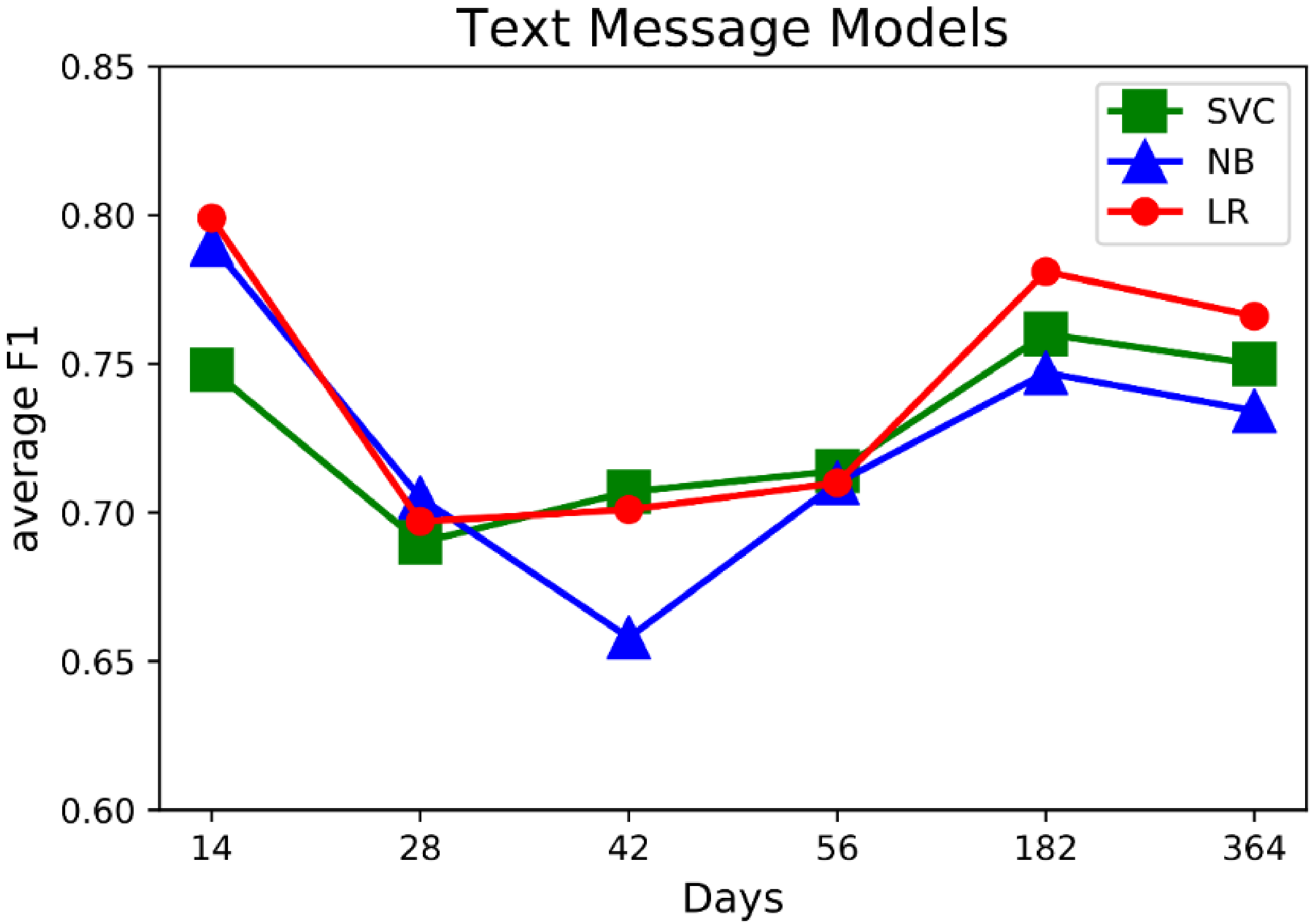
Want to make sure
healthy people
are not diagnosed

Want to make sure
people with depression
are diagnosed

$$\mathbf{F1} = 2\,\frac{\mathbf{Precision(Recall)}}{\mathbf{Precision + Recall}}$$

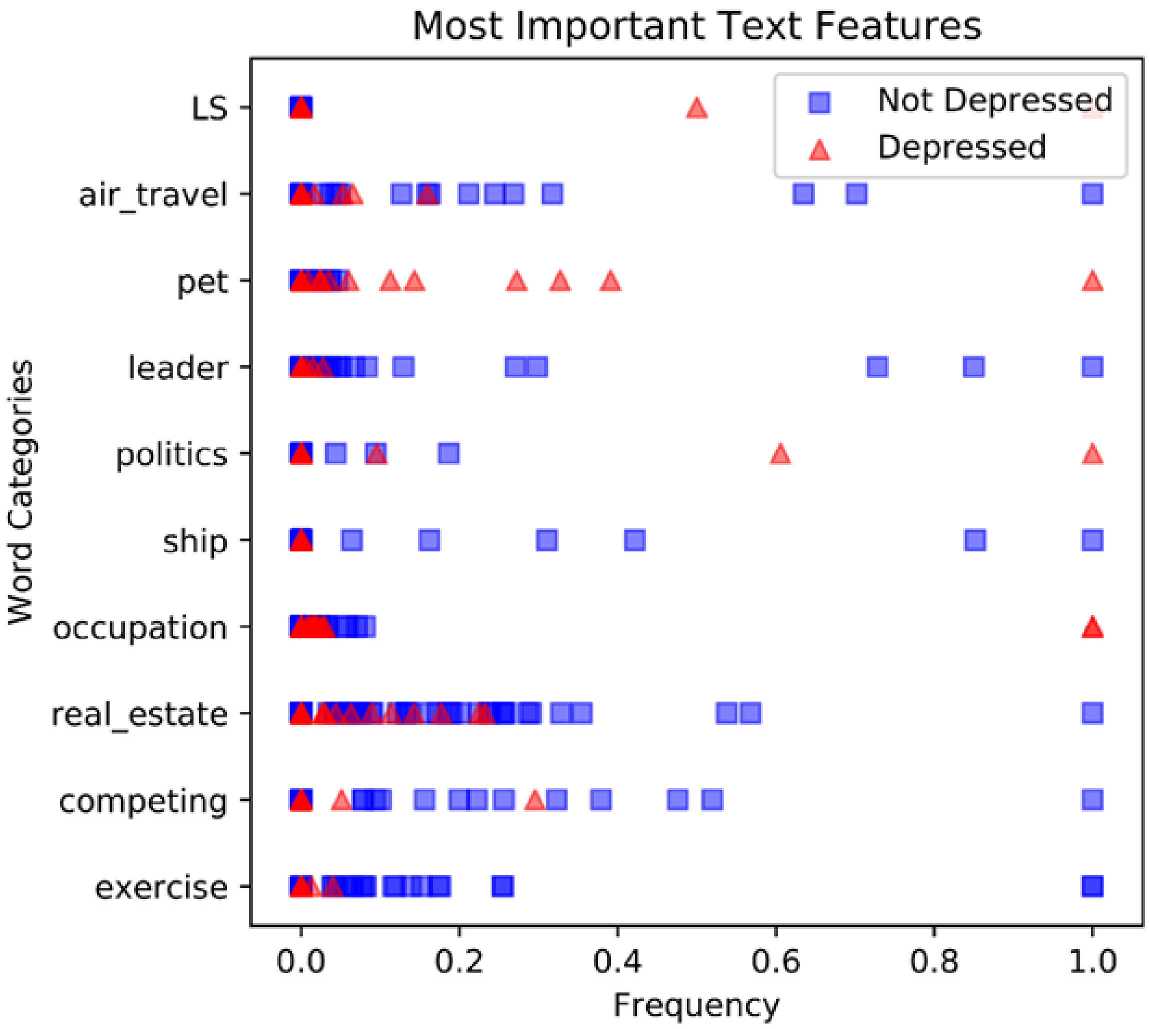
researchgate.net

Texts are More Predictive Than Tweets



Logistic regression models built with the last 14 days of text messages are the best at screening for depression

The Top 10 Chi-Squared Features

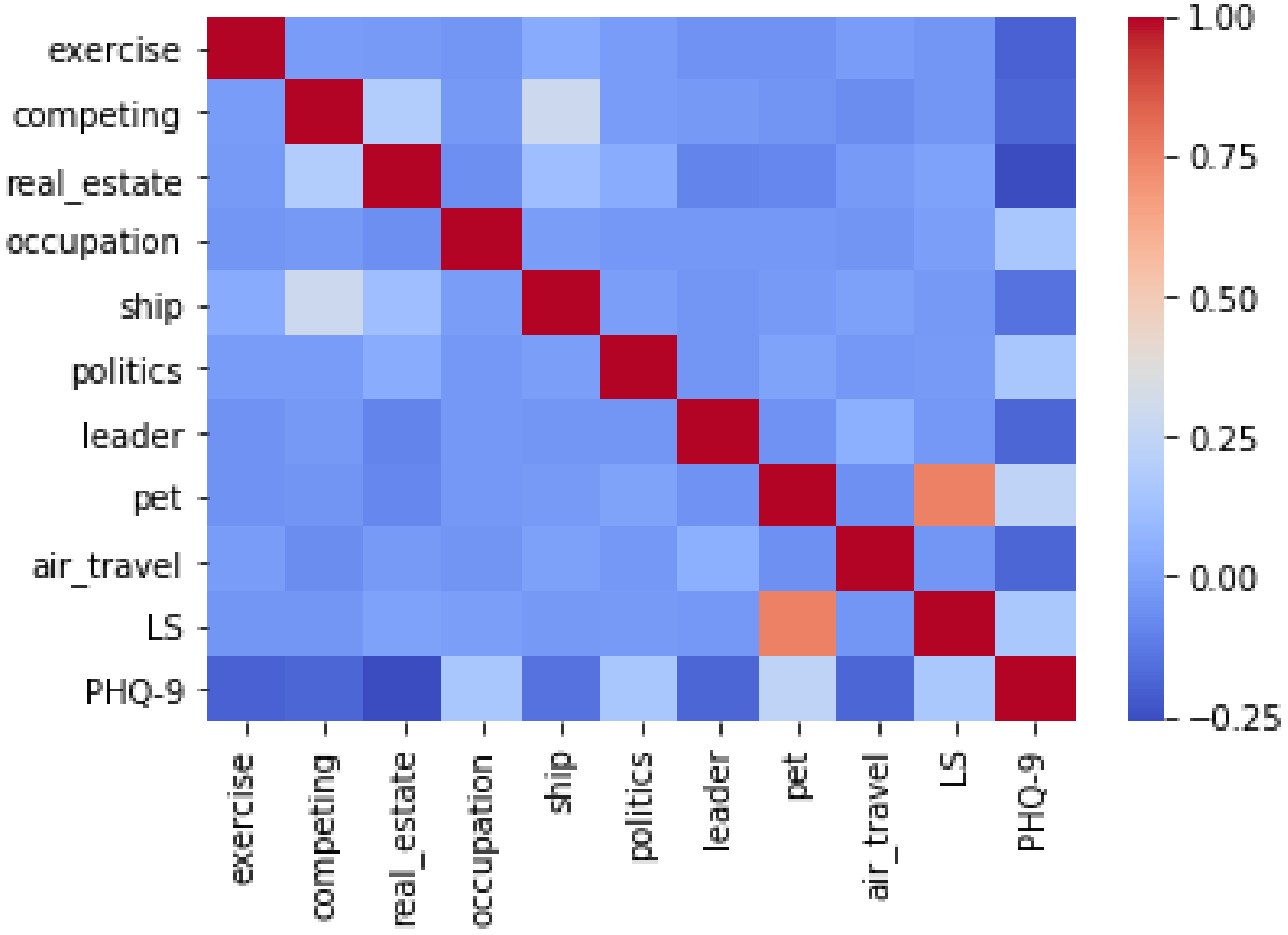


High usage of words in the categories

- Air travel
- Leader
- Real estate
- Competing
- Exercise

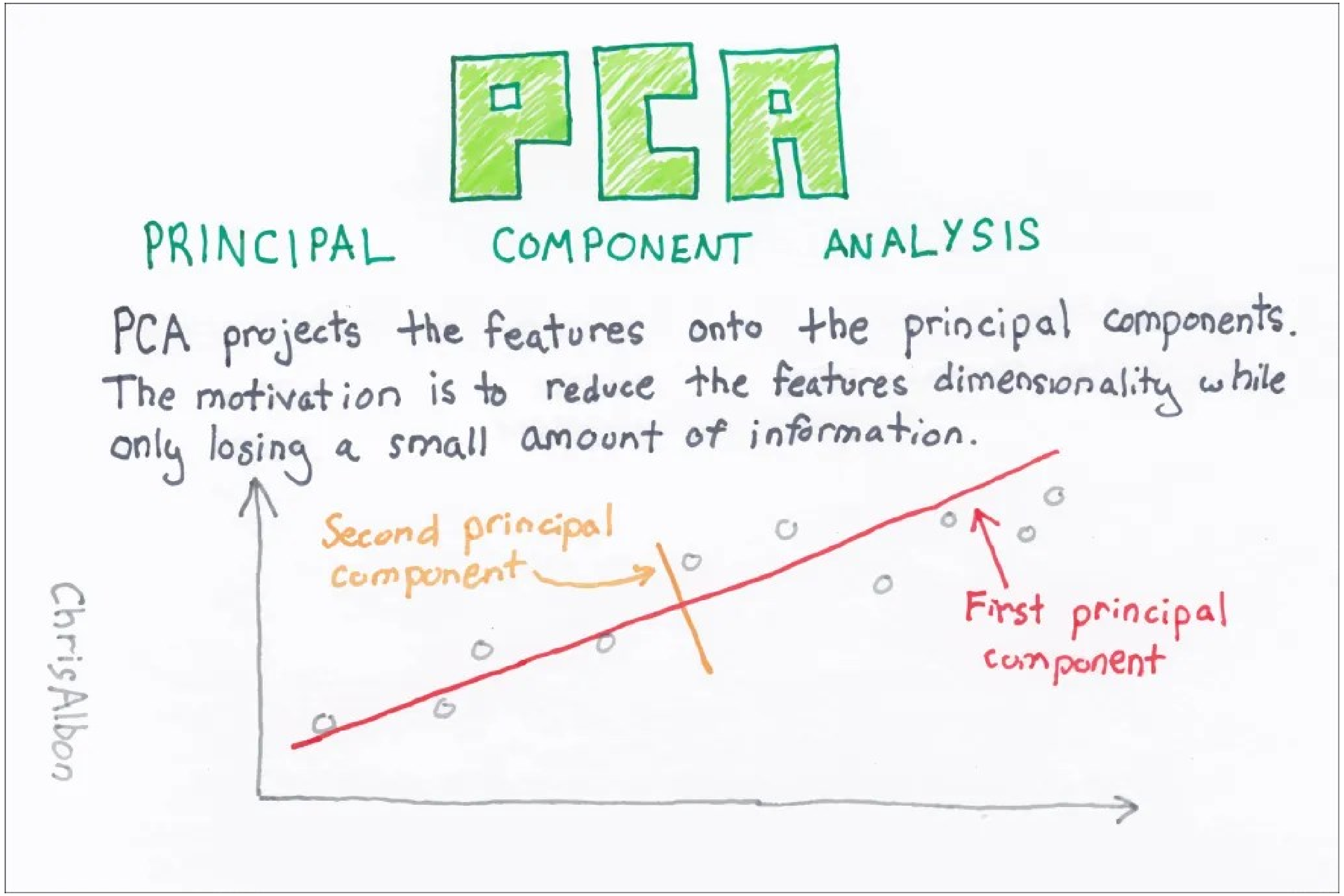
are indicative of not being depressed

Feature Correlation



There is a high Pearson's correlation between categories list item marker (LS) and pet

Principal Component Analysis (PCA)



PCA

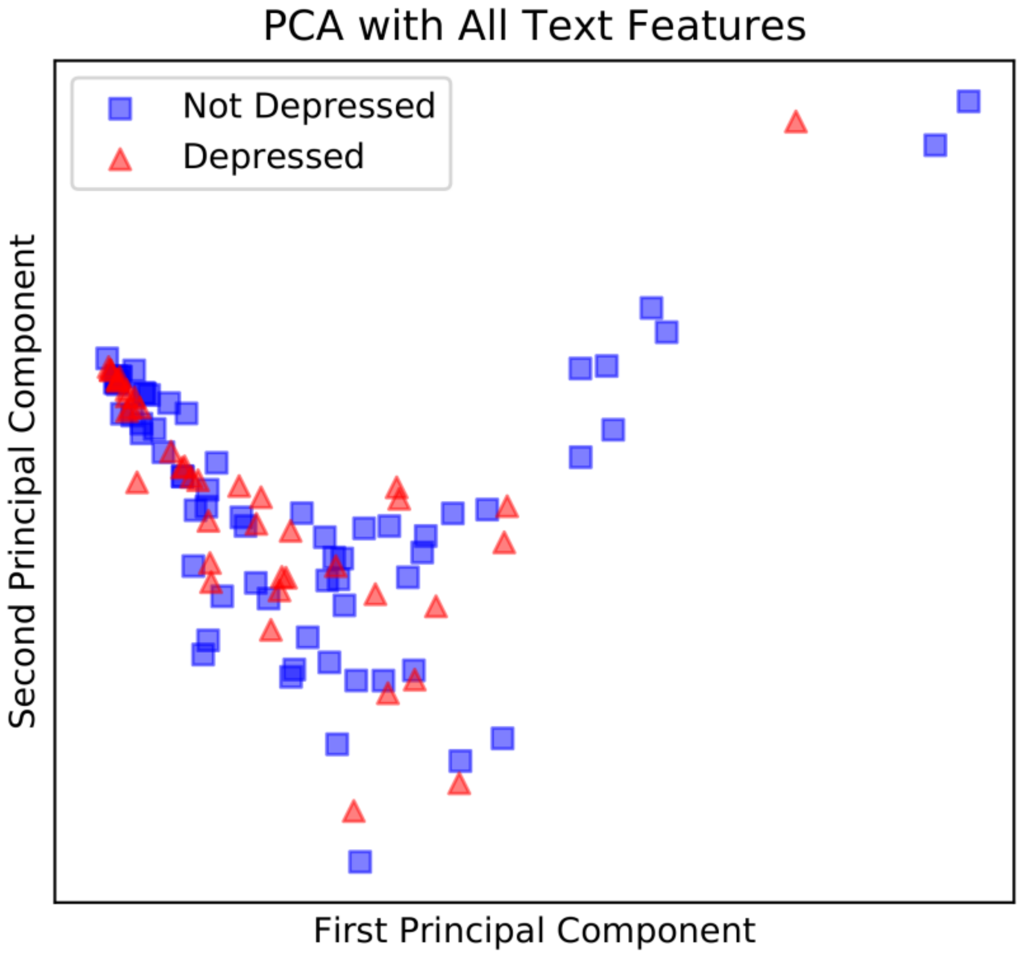
PRINCIPAL COMPONENT ANALYSIS

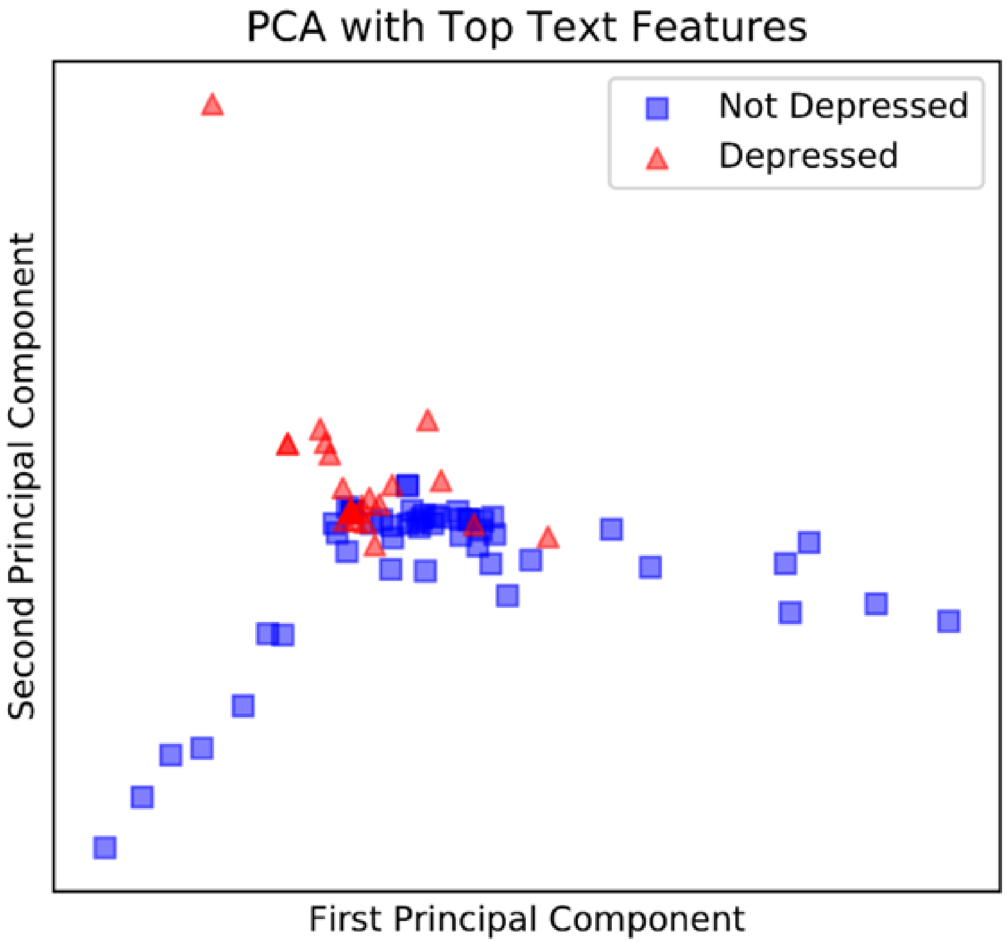
PCA projects the features onto the principal components. The motivation is to reduce the features dimensionality while only losing a small amount of information.

Each successive principal component covers less variance

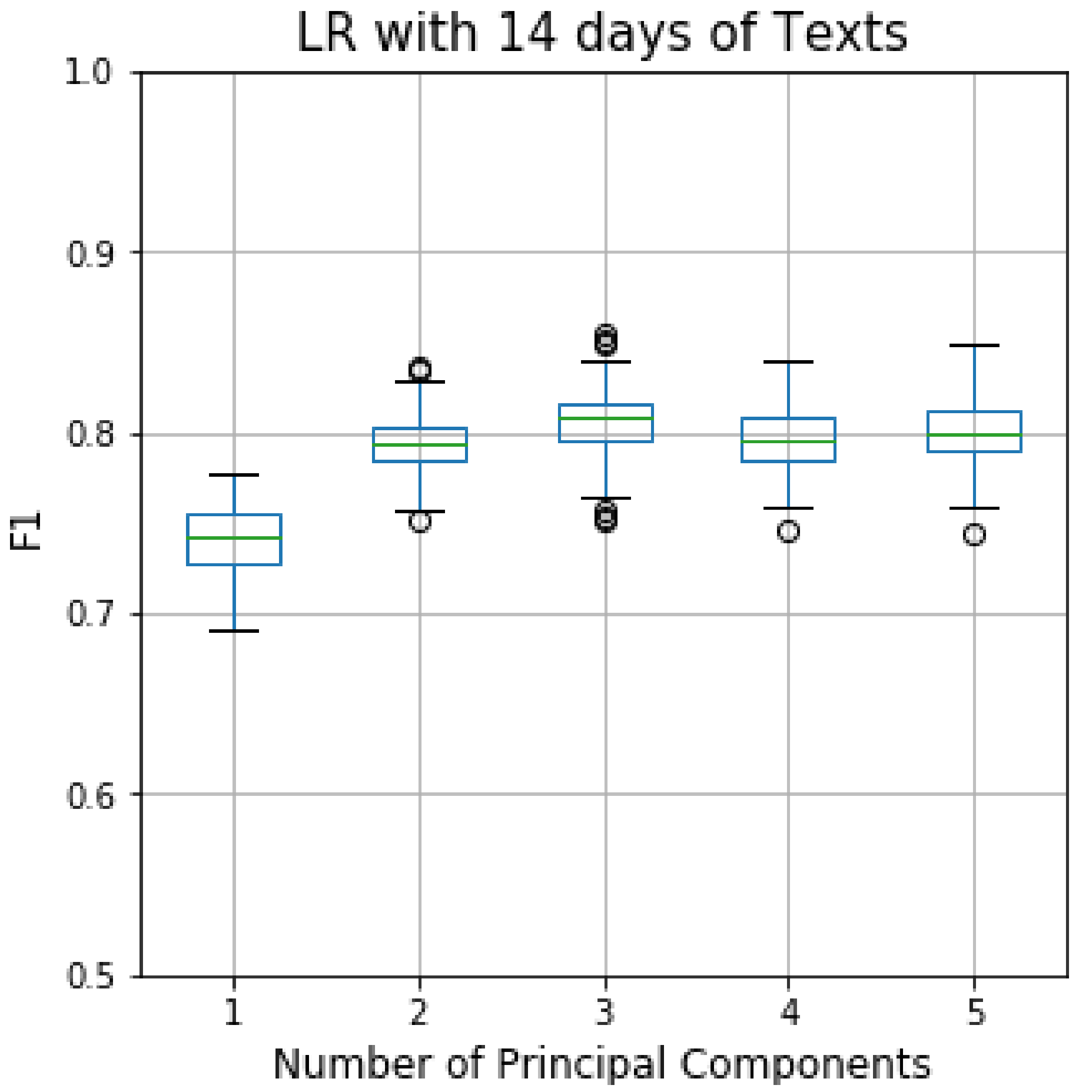
Can also be used to combat feature collinearity

Principal Component Analysis Results





Final Results



Best model is built with just three principal components

F1	0.806 ± 0.019
Precision	0.721 ± 0.027
Recall	0.925 ± 0.011
Specificity	0.620 ± 0.046
AUC	0.832 ± 0.022
Accuracy	0.773 ± 0.024

Conclusion

Takeaways

- Private messages are more predictive than public messages
- Two weeks of data was more predictive than greater temporal quantities of messages
- Machine learning is better for smaller datasets in domains where interpretability is important

Limitations

- Data quantity
  - Few participants submitted both texts and tweets, limiting any multi-modal analysis
  - Some participants shared very few text messages, making them challenging to classify
- Screening Tool as Ground Truth
  - Requires honest self-reflection
  - Limited in accuracy, especially around similar scores

Prior and Future Research

Other Modalities

- “You're Making Me Depressed: Leveraging Texts from Contact Subsets to Predict Depression” in Proceedings of IEEE biomedical and Health Informatics, 2019
- “Depression Screening from Text Message Reply Latency” in Proceedings of the IEEE Engineering in Medicine and Biology Society, 2020

Future Directions

- Collect more data
  - From college students
  - With clinician diagnosis
  - Multiple text modalities
  - More messaging apps
- Feature engineering targeted towards each text modality
- Only use conversational data
- Predict score or category rather than a binary depression cutoff

Any Questions?

- mltlachac@wpi.edu
- www.linkedin.com/in/mltlachac
- www.github/mltlachac

Acknowledgement: US Department of Education P200A150306: GAANN grant;
Toto, Bennett, Prof. Agu, Dogrucu, Peruic, Isaro, Ball, Gao,
Flannery, Resom, Assan, Wu, and DSRG community at WPI