Progress Report 2
ENSF 619-4

Willis CHEUNG
Justin WOODS
Julian MULIA

April 6, 2019

Instructor: Zahra Shakeri

1 Objectives

- Data Labeling
- Visualizations and comparison with last trial
- Unsupervised Learning
- Supervised Learning

2 Update

Since the last progress report, we were able to perform the following:

- Re-run our pre-processing script using the new keywords derived from various subreddits.
- Compare the distribution of data labels to our previous trial.
- Perform supervised and unsupervised learning.

3 Data Distribution

After re-running our pre-processing script, we were able to filter out 91,811 tweets. We believed that applying spam filters with commonly seen trigrams used in job postings and apartment listings (ex: ”click the link” and ”apartments for rent”) could improve the results of our NLP modeling. This time, we labeled 2045 data points using the same methodology as before:

- Indicative of psychological stress. (P)
- Indicative of sleep disorder. (Z)
- Both sleep and stress disorder. (B)
- Neither sleep nor stress disorder. (N)
- Unclear. (U)

During our previous labeling, we had seen the distribution of data points to be the following:

Table 1: Label Counts with Original Keyword Filtering

Label	Freq	Perc
N	1557	77.81
P	209	10.44
U	147	7.35
Z	55	2.75
B	33	1.65

Now, with our improved keyword filters, we were able to improve the distribution of relevant points, particularly for tweets containing indicators of stress.

Table 2: Label Counts with Refined Keyword Filtering

Label	Freq	Perc
N	1489	72.81
P	390	19.07
Z	95	4.65
B	45	2.20
U	26	1.27

4 Supervised Learning

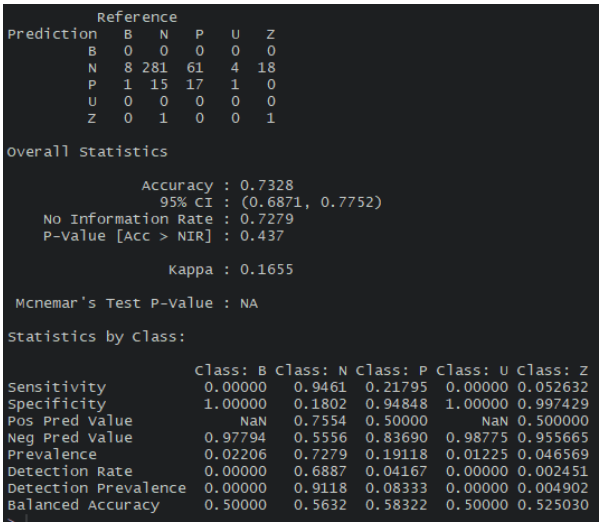
With our new labeled data set, we were able to perform the following supervised learning techniques:

- Naive Bayes
- Logistic Regression

4.1 Naive Bayes

For our Naive Bayes model, we used the variation known as the binary Naive Bayes which considers only the presence/absence of features, rather than the frequency. This is more effective for sentiment classification (particularly in shorter documents) as word occurrence is more important than word frequency. For our modeling, we considered a 80:20 training/test split.

We reduced the number of features by ignoring words appearing in less than five tweets. The resulting confusion matrices can be shown below:



```
          Reference
Prediction   B   N   P   U   Z
         B   0   0   0   0   0
         N   8 281  61   4  18
         P   1  15  17   1   0
         U   0   0   0   0   0
         Z   0   1   0   0   1

Overall Statistics

               Accuracy : 0.7328
                 95% CI : (0.6871, 0.7752)
    No Information Rate : 0.7279
    P-Value [Acc > NIR] : 0.437

                  Kappa : 0.1655

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: B Class: N Class: P Class: U Class: Z
Sensitivity           0.00000   0.9461  0.21795  0.00000 0.052632
Specificity           1.00000   0.1802  0.94848  1.00000 0.997429
Pos Pred Value            NaN   0.7554  0.50000      NaN 0.500000
Neg Pred Value        0.97794   0.5556  0.83690  0.98775 0.955665
Prevalence            0.02206   0.7279  0.19118  0.01225 0.046569
Detection Rate        0.00000   0.6887  0.04167  0.00000 0.002451
Detection Prevalence  0.00000   0.9118  0.08333  0.00000 0.004902
Balanced Accuracy     0.50000   0.5632  0.58322  0.50000 0.525030
```

Figure 1: Binary Naive Bayes Summary Results and Confusion Matrix

From the results shown above in Figure 1, we can see that the model is just slightly better than if one were to just predict a class of ”Neither”, which would result in being accurate 72.79% of the time. We believe this is due to the relative lack of tweets labeled as ”sleep” or ”stress” compared to ”neither”. Perhaps with more labeled points and improved spam filtering, this result could be improved.

4.2 Logistic Regression

For our logistic regression, we took the following features as potential independent variables:

- Total Sentiment Score (normalized using AFINN lexicon scores)
- account_life_days (normalized)
- latitude (normalized)

- longitude (normalized)
- province
- tweet_day (day of month)
- tweet_hour (0 = midnight)
- tweet_weekday (0 = monday)
- user_followers_count (normalized)
- user_friends_count (normalized)
- user_listed_count (normalized)
- user_statuses_count (normalized)
- user_verified (normalized)

From the fit of the linear model to our labeled data, we can see if there are any significant correlations to the sleep or stress labels. From Table 11, we notice that the following:

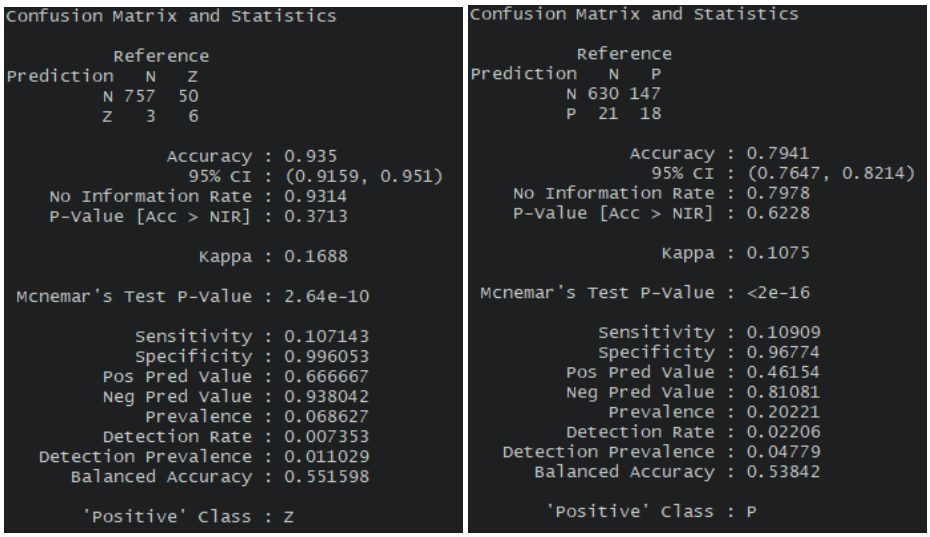
- Authors from Manitoba had a higher odds of writing a tweet being indicative of a sleep disorder ($p < 0.05$).
- Tweets posted between 1am-4am ($p < 0.01$) and 5am-7am ($p < 0.05$) had a higher odds of being indicative of a sleep disorder.
- Users with a higher friends count had a lower odds of publishing tweets being indicative of sleep disorder ($p < 0.05$)

If more tweets were able to be labeled, one might find additional significant findings from this particular model (for example, further variations for tweet hours, province, or other independent variables).

A similar model can be built with tweets that were labeled as showing signs of psychological stress. From Table 12, we can extract some more interesting findings:

- Having a higher sentiment score resulted in a lower odds of the tweet being indicative of stress ($p < 0.01$).
- Tweets published from Manitoba, New Brunswick, Nova Scotia, Ontario, and Quebec all had a higher odds of indicating Stress ($p < 0.05$).
- Tweets published between, 12pm-1pm had a lower odds of being indicative of stress ($p < 0.05$).
- Tweets published on Wednesdays ($p < 0.01$), Thursdays ($p < 0.05$), and Sundays ($p < 0.05$) all had higher odds of indicating stress.

Finally, the models can be applied to our test data set to evaluate performance. We started by conducting a single fold prediction for a 60:40 train/test split.

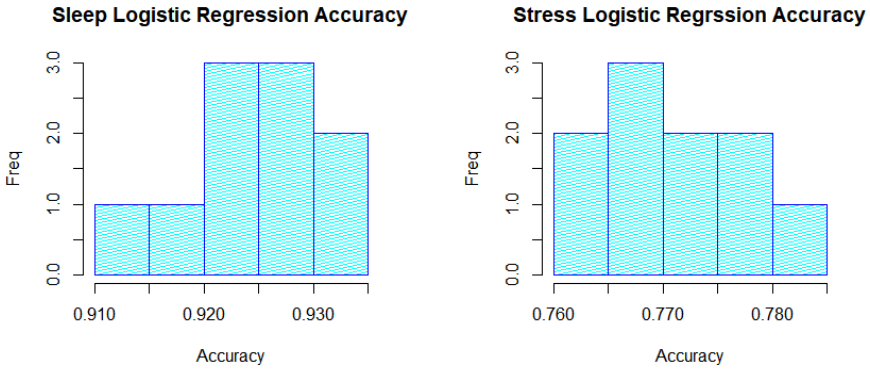


(a) Logistic Regression Model Sleep Confusion Matrix

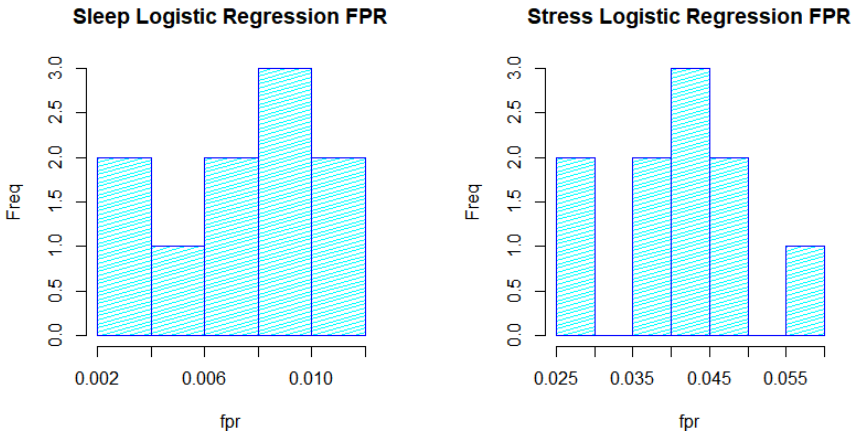
(b) Logistic Regression Model Stress Confusion Matrix

Figure 2: Logistic Regression Results

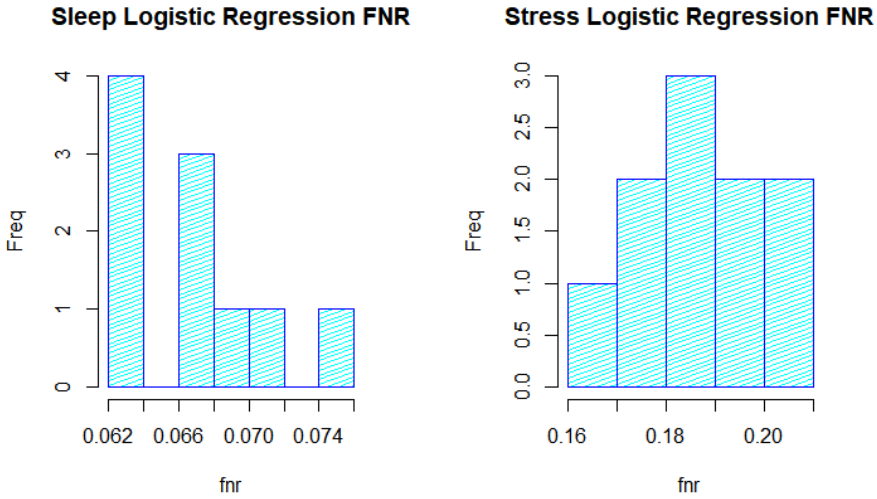
The models both perform quite poorly in predicting for sleep and stress, due to the heavy bias in the labeled data. As a result, the models are really only good for predicting whether the tweet is labeled as ”neither”. Again, better performance might be achieved through refined keyword filtering and additional data labeling. As an exercise, we performed a 10 fold cross validation to see if there would be any variation in accuracy or false positive and false negative rates. In each of the cases, we did not see that much variation in the accuracy, or false positive and negative rates.



(a) Sleep/Stress Cross Validated Accuracy



(b) Sleep and Stress Logistic Regression Models FPR



(c) Sleep and Stress Logistic Regression Models FNR

Figure 3: Logistic Regression Cross Validation Results

4.3 Unsupervised Learning

We employed the BDA and LDA unsupervised learning techniques to extract three topics from the full set of unlabeled tweets.

For the LDA method, we used a subset of 25000 of the tweets to extract topics due to memory constraints. For example, from Table 6, we can interpret the topics as follows:

- Topic 1: Mental Health
- Topic 2: Stress related to work
- Topic 3: Stress/Sleep related to weather, early morning
- Topic 4: General positive sentiment
- Topic 5: Tired/sick sentiment
- Topic 6: Angry sentiment (very stressed)

Table 3: LDA table 3 topics

	Topic 1	Topic 2	Topic 3
1	cold	life	shit
2	school	hate	fuck
3	job	heart	fucking
4	day	people	time
5	today	years	hours

Table 4: LDA table 4 topics

	Topic 1	Topic 2	Topic 3	Topic 4
1	life	school	shit	cold
2	time	heart	fuck	day
3	job	coffee	fucking	today
4	years	rest	hate	hours
5	mental	new	people	late

Table 5: LDA table 5 topics

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	life	heart	cold	shit	school
2	time	coffee	day	fuck	rest
3	job	mental	today	fucking	new
4	years	love	hours	hate	high
5	ago	health	late	people	year

Table 6: LDA table 6 topics

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
1	job	life	shit	time	cold	school
2	mental	heart	fuck	hours	day	rest
3	health	years	fucking	late	today	new
4	ever	love	hate	bed	coffee	high
5	someone	emojifacewithtearsofjoy	people	sick	good	year

Table 7: LDA table 7 topics

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
1	job	late	cold	life	time	shit	school
2	mental	years	day	heart	hours	fuck	new
3	health	ago	today	love	bed	fucking	high
4	ever	game	coffee	never	sick	hate	year
5	care	pressure	good	rest	emojifacewithtearsofjoy	people	kids

For the BDA method, we used a subset of 6511 of the tweets to extract topics due to memory constraints. The three to five topic models are shown below. Again we can try to interpret the topics by grouped tokens. For example, for Table 10:

- token: Complaining about weather/school/life
- token1: Complaining about weather, cold, snow
- token2: Difficult to interpret
- token3: Weather forecast
- token4: Public health announcement (likely related to BellLetsTalk Campaign)

Table 8: BDA table 3 topics

	token	token1	token2
1	life	mental	emoji_red_heart
2	cold	health	school
3	im	emoji_pizza	alarm
4	school	stigma	avenue
5	shit	emoji_woman	institution

Table 9: BDA table 4 topics

	token	token1	token2	token3
1	life	school	mental	cold
2	im	emoji_red_heart	health	wind
3	cold	avenue	emoji_pizza	emoji_heavy_minus_sign
4	school	dispatched	stigma	weather
5	shit	institution	fighting	emoji_snowflake

Table 10: BDA table 5 topics

	token	token1	token2	token3	token4
1	life	cold	emoji_red_heart	emoji_pizza	mental
2	im	coffee	school	wind	health
3	school	emoji_snowflake	avenue	pressure	stigma
4	cold	emoji_heavy_minus_sign	alarm	rain	fighting
5	shit	snow	institution	forecast	alcoholawareness

5 Appendix A

Table 11: Sleep logistic regression results

	Dependent variable:
	Dependent variable:
	sleep_label
total_score	−0.171*
	(0.093)
account_life_days	−0.126
	(0.099)
‘provinceBritish Columbia‘	−0.360
	(0.717)
provinceManitoba	2.369**
	(1.146)
‘provinceNew Brunswick‘	5.181*
	(3.121)
‘provinceNewfoundland and Labrador‘	4.057
	(3.633)
‘provinceNorthwest Territories‘	−17.438
	(6,522.639)
‘provinceNova Scotia‘	3.802
	(3.241)
provinceOntario	3.488
	(2.268)
‘provincePrince Edward Island‘	−11.300
	(2,209.739)
provinceQuebec	4.309*
	(2.566)
provinceSaskatchewan	1.284*
	(0.724)
provinceYukon	−17.334

	(6,522.639)
latitude	0.336
	(0.320)
longitude	−1.412
	(1.180)
tweet_day2	−0.207
	(2,192.898)
tweet_day3	−32.027
	(3,055.275)
tweet_day4	0.792
	(0.954)
tweet_day5	14.985
	(2,023.118)
tweet_day6	−15.776
	(2,346.932)
tweet_day7	−33.393
	(3,055.275)
tweet_day8	−0.152
	(0.635)
tweet_day9	−0.241
	(2,192.899)
tweet_day10	−32.452
	(3,055.275)
tweet_day11	−15.768
	(1,640.505)
tweet_day12	15.890
	(2,023.118)
tweet_day13	−14.989
	(2,346.932)
tweet_day14	−33.384
	(3,055.275)

tweet_day15	−1.447*
	(0.871)
tweet_day16	−0.143
	(2,192.899)
tweet_day17	−32.460
	(3,055.275)
tweet_day18	−1.098
	(1.156)
tweet_day19	13.956
	(2,023.118)
tweet_day20	−16.195
	(2,346.932)
tweet_day21	−33.754
	(3,055.275)
tweet_day22	−0.341
	(0.892)
tweet_day23	−15.349
	(2,346.932)
tweet_day24	−31.988
	(3,055.275)
tweet_day25	0.158
	(0.649)
tweet_day26	0.142
	(2,192.899)
tweet_day27	−31.913
	(3,055.275)
tweet_day28	−0.178
	(0.915)
tweet_day29	15.390
	(2,023.118)

tweet_day30	−16.138
	(2,346.931)
tweet_day31	−32.362
	(3,055.275)
tweet_hour1	1.574**
	(0.619)
tweet_hour2	1.480**
	(0.723)
tweet_hour3	1.490**
	(0.693)
tweet_hour4	0.652
	(0.806)
tweet_hour5	1.569**
	(0.659)
tweet_hour6	1.270**
	(0.636)
tweet_hour7	0.998
	(0.650)
tweet_hour8	1.098*
	(0.640)
tweet_hour9	−0.077
	(0.720)
tweet_hour10	0.179
	(0.651)
tweet_hour11	0.474
	(0.634)
tweet_hour12	−0.318
	(0.716)
tweet_hour13	0.039
	(0.718)
tweet_hour14	−0.453

	(0.772)
tweet_hour15	−0.260
	(0.717)
tweet_hour16	−0.476
	(0.774)
tweet_hour17	−0.240
	(0.716)
tweet_hour18	−1.772
	(1.121)
tweet_hour19	−0.106
	(0.675)
tweet_hour20	0.208
	(0.633)
tweet_hour21	0.878
	(0.567)
tweet_hour22	0.385
	(0.618)
tweet_hour23	0.497
	(0.654)
tweet_weekday1	−32.297
	(3,055.275)
tweet_weekday2	−32.633
	(2,127.417)
tweet_weekday3	−0.445
	(0.577)
tweet_weekday4	−33.237
	(3,055.275)
tweet_weekday5	−48.405
	(2,289.476)
tweet_weekday6	−17.550
	(1,956.175)

user_followers_count	−0.014
	(0.291)
user_friends_count	−0.813**
	(0.389)
user_listed_count	−0.003
	(0.202)
user_statuses_count	−0.111
	(0.155)
user_verified	−0.867
	(1.120)
Constant	27.403
	(3,055.276)
Observations	2,045
Log Likelihood	−437.432
Akaike Inf. Crit.	1,034.864
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 12: Stress logistic regression results

	Dependent variable:
	stress_label
total_score	-0.653^{***}
	(0.066)
account_life_days	-0.093
	(0.060)
‘provinceBritish Columbia‘	-0.756^{*}
	(0.436)
provinceManitoba	1.892^{**}
	(0.748)
‘provinceNew Brunswick‘	4.943^{**}
	(2.035)
‘provinceNewfoundland and Labrador‘	4.287^{*}
	(2.349)
‘provinceNorthwest Territories‘	-12.593
	(535.412)
‘provinceNova Scotia‘	4.893^{**}
	(2.102)
provinceOntario	3.473^{**}
	(1.498)
‘provincePrince Edward Island‘	4.422^{*}
	(2.332)
provinceQuebec	3.719^{**}
	(1.687)
provinceSaskatchewan	0.591
	(0.479)
provinceYukon	-14.352
	(535.412)
latitude	0.202
	(0.206)

longitude	-1.600^{**}
	(0.752)
tweet_day2	-0.239
	(1.288)
tweet_day3	0.796
	(1.591)
tweet_day4	-0.062
	(0.739)
tweet_day5	-0.575
	(1.297)
tweet_day6	0.096
	(1.475)
tweet_day7	2.080
	(1.743)
tweet_day8	-0.182
	(0.453)
tweet_day9	-0.601
	(1.364)
tweet_day10	1.155
	(1.668)
tweet_day11	0.073
	(0.755)
tweet_day12	-0.381
	(1.295)
tweet_day13	-0.236
	(1.476)
tweet_day14	2.234
	(1.740)
tweet_day15	-0.958^{*}
	(0.539)

tweet_day16	−0.821
	(1.353)
tweet_day17	1.390
	(1.656)
tweet_day18	0.785*
	(0.460)
tweet_day19	−0.236
	(1.285)
tweet_day20	0.194
	(1.452)
tweet_day21	2.322
	(1.741)
tweet_day22	−0.657
	(0.609)
tweet_day23	0.303
	(1.453)
tweet_day24	1.774
	(1.747)
tweet_day25	−0.600
	(0.519)
tweet_day26	−1.244
	(1.384)
tweet_day27	1.251
	(1.662)
tweet_day28	−0.267
	(0.567)
tweet_day29	0.081
	(1.258)
tweet_day30	−0.626
	(1.357)
tweet_day31	1.084

	(1.655)
tweet_hour1	0.087
	(0.420)
tweet_hour2	−0.743
	(0.555)
tweet_hour3	−0.277
	(0.516)
tweet_hour4	−0.153
	(0.550)
tweet_hour5	−1.173*
	(0.610)
tweet_hour6	−0.298
	(0.432)
tweet_hour7	−0.050
	(0.415)
tweet_hour8	−0.515
	(0.436)
tweet_hour9	−0.610
	(0.399)
tweet_hour10	−0.151
	(0.363)
tweet_hour11	−0.661*
	(0.396)
tweet_hour12	−0.792**
	(0.399)
tweet_hour13	−0.776*
	(0.398)
tweet_hour14	−0.704*
	(0.404)
tweet_hour15	0.001
	(0.358)

tweet_hour16	−0.053
	(0.370)
tweet_hour17	−0.524
	(0.375)
tweet_hour18	−0.648*
	(0.368)
tweet_hour19	−0.633*
	(0.363)
tweet_hour20	−0.212
	(0.350)
tweet_hour21	−0.531
	(0.352)
tweet_hour22	−0.511
	(0.364)
tweet_hour23	−0.429
	(0.380)
tweet_weekday1	2.469
	(1.677)
tweet_weekday2	3.008***
	(1.099)
tweet_weekday3	1.213**
	(0.521)
tweet_weekday4	1.830
	(1.743)
tweet_weekday5	2.086*
	(1.194)
tweet_weekday6	2.026**
	(0.961)
user_followers_count	0.052
	(0.053)

user_friends_count	−0.047
	(0.090)
user_listed_count	−0.039
	(0.085)
user_statuses_count	−0.084
	(0.077)
user_verified	−0.732
	(0.608)
Constant	−5.433***
	(1.995)
Observations	2,045
Log Likelihood	−954.590
Akaike Inf. Crit.	2,069.180
Note:	*p<0.1; **p<0.05; ***p<0.01