```
1  C:\Users\270708326\PycharmProjects\PythonProject\Sentiment\.venv\Scripts\python.exe C:\Users\
   270708326\PycharmProjects\PythonProject\Sentiment\sentiment_analysis.py
2  All libraries imported successfully!
3  Analysis Date: 2025-10-15 10:02:34
4
5  ==========================================================================
6  ONE NZ APP SENTIMENT ANALYSIS - BUSINESS ANALYTICS PROJECT
7  ==========================================================================
8
9  ==========================================================================
10 SECTION 2: DATA LOADING AND EXPLORATION
11 ==========================================================================
12
13 Dataset Shape: (9386, 9)
14 Total Reviews: 9,386
15
16 Column Names and Types:
17 reviewId            object
18 content             object
19 manually_labelled   object
20 score                int64
21 thumbsUpCount        int64
22 at                  object
23 replyContent        object
24 repliedAt           object
25 appVersion          object
26 dtype: object
27
28 First Few Records:
29                              reviewId  \
30 0  005447b5-7160-440d-8e35-9043944b61bf
31 1  44c57a8a-adb9-4ac3-b23c-5ec40017350b
32 2  2f5fe9e5-348b-42e9-a8f9-d6c2c7fa1f7e
33
34
   content  \
35 0                                                answered all my
   questions
36 1                  Charges you 13 months a year thinking each month is February, every 28
   days.
37 2  The app wont even open so i can pay my bill, this has been going on for months and months, and
    t...
38
39   manually_labelled  score  thumbsUpCount              at  \
40 0          positive      5              0   9/22/2025 4:43
41 1          negative      1              5  9/17/2025 21:39
42 2          negative      1              0   9/17/2025 3:14
43
44
   replyContent  \
45 0                          Awesome stuff Gilda, we are pleased that we were able to
   help!
46 1  Hi Ben, you'll need to login via our website to do that, I'll pass your feedback on to the dev
    t...
47 2

   NaN
48
49           repliedAt appVersion
50 0  9/23/2025 4:59      7.5.1
51 1  4/4/2016 21:22      7.5.1
52 2            NaN      7.5.1
53
54 Basic Statistics:
55                score  thumbsUpCount
56 count  9386.000000    9386.000000
57 mean      2.423610       1.903367
58 std       1.689345       9.593348
59 min       1.000000       0.000000
```

```
 60 25%         1.000000         0.000000
 61 50%         1.000000         0.000000
 62 75%         4.000000         1.000000
 63 max         5.000000       413.000000
 64
 65 Missing Values:
 66                    Missing Count  Percentage
 67 content                      11    0.117196
 68 manually_labelled            23    0.245046
 69 replyContent               6257   66.663115
 70 repliedAt                  6257   66.663115
 71 appVersion                  923    9.833795
 72
 73 ==============================================================================
 74 SECTION 3: DATA PREPROCESSING WITH ADVANCED NLP
 75 ==============================================================================
 76
 77 1. Processing date columns...
 78 2. Extracting temporal features...
 79 3. Creating binary features...
 80 4. Calculating response time...
 81 5. Preprocessing review text...
 82    a. Basic cleaning (for sentiment analysis)...
 83    b. Stopword removal (for word clouds and topic modeling)...
 84    c. Lemmatization (for semantic analysis)...
 85 6. Calculating text metrics...
 86 7. Grouping app versions...
 87 8. Creating sentiment labels...
 88
 89 Cleaned dataset shape: (9386, 25)
 90
 91 Text processing summary:
 92    - Average words (original): 19.0
 93    - Average words (no stopwords): 9.0
 94    - Stopwords removed: 52.6%
 95
 96 Preprocessing complete!
 97
 98 ==============================================================================
 99 SECTION 4: EXPLORATORY DATA ANALYSIS
100 ==============================================================================
101
102 1. Review Score Distribution:
103 score
104 1    4831
105 2     925
106 3     691
107 4     701
108 5    2238
109 Name: count, dtype: int64
110
111 Average Score: 2.42
112 Median Score: 1.00
113
114 2. Sentiment Distribution:
115 sentiment_label
116 Negative   5756
117 Positive   2939
118 Neutral     691
119 Name: count, dtype: int64
120
121 Negative Review Percentage: 61.3%
122
123 3. Reviews Over Time:
124 Peak Month: 2016-12 (391 reviews)
125 Lowest Month: 2013-04 (1 reviews)
126
127 4. Top App Versions:
128 appVersion
```

```
129 3.5        435
130 4.1        274
131 2.5.3      274
132 3.15       269
133 3.3        243
134 3.31       242
135 3.30.1     204
136 6.1.0      184
137 4.18.1     172
138 3.10.1     167
139 Name: count, dtype: int64
140
141 5. Company Response Rate:
142 Overall Response Rate: 33.3%
143 Average Response Time: -547.0 hours (-22.8 days)
144
145 6. Review Length Statistics:
146 Average Word Count: 19.0
147 Average Character Length: 99.1
148
149 ================================================================================
150 SECTION 5: ADVANCED SENTIMENT ANALYSIS
151 ================================================================================
152
153 1. VADER Sentiment Analysis...
154 2. TextBlob Sentiment Analysis...
155
156 3. RoBERTa/BERT Sentiment Analysis...
157    This may take several minutes for large datasets...
158    Loading RoBERTa model...
159 Some weights of the model checkpoint at cardiffnlp/twitter-roberta-base-sentiment-latest were
    not used when initializing RobertaForSequenceClassification: ['roberta.pooler.dense.bias', '
    roberta.pooler.dense.weight']
160 - This IS expected if you are initializing RobertaForSequenceClassification from the checkpoint
    of a model trained on another task or with another architecture (e.g. initializing a
    BertForSequenceClassification model from a BertForPreTraining model).
161 - This IS NOT expected if you are initializing RobertaForSequenceClassification from the
    checkpoint of a model that you expect to be exactly identical (initializing a
    BertForSequenceClassification model from a BertForSequenceClassification model).
162 Device set to use cpu
163    Processing 9386 reviews...
164    Processing batch 10/294...
165    Processing batch 20/294...
166    Processing batch 30/294...
167    Processing batch 40/294...
168    Processing batch 50/294...
169    Processing batch 60/294...
170    Processing batch 70/294...
171    Processing batch 80/294...
172    Processing batch 90/294...
173    Processing batch 100/294...
174    Processing batch 110/294...
175    Processing batch 120/294...
176    Processing batch 130/294...
177    Processing batch 140/294...
178    Processing batch 150/294...
179    Processing batch 160/294...
180    Processing batch 170/294...
181    Processing batch 180/294...
182    Processing batch 190/294...
183    Processing batch 200/294...
184    Processing batch 210/294...
185    Processing batch 220/294...
186    Processing batch 230/294...
187    Processing batch 240/294...
188    Processing batch 250/294...
189    Processing batch 260/294...
190    Processing batch 270/294...
191    Processing batch 280/294...
```

```
192    Processing batch 290/294...
193    RoBERTa analysis complete!
194
195 4. Sentiment Validation:
196 VADER Accuracy vs Actual Ratings: 54.62%
197 TextBlob Accuracy vs Actual Ratings: 44.89%
198 RoBERTa Accuracy vs Actual Ratings: 79.12%
199
200 5. Sentiment Analysis Results:
201
202 VADER Sentiment Distribution:
203 vader_sentiment
204 Positive    4095
205 Negative    3030
206 Neutral     2261
207 Name: count, dtype: int64
208
209 TextBlob Sentiment Distribution:
210 textblob_sentiment
211 Positive    3822
212 Neutral     3625
213 Negative    1939
214 Name: count, dtype: int64
215
216 RoBERTa Sentiment Distribution:
217 roberta_sentiment
218 Negative    5715
219 Positive    2408
220 Neutral     1263
221 Name: count, dtype: int64
222
223 ==============================================================================
224 SECTION 6: TOPIC MODELING
225 ==============================================================================
226
227 Processing 9386 reviews for topic modeling...
228
229 1. Creating TF-IDF matrix...
230    TF-IDF matrix shape: (9386, 1000)
231
232 2. Performing LDA with 5 topics...
233
234 LDA Topics:
235 Topic 1: good, easy, use, easy use, download, happy, thank, fine, uninstall, simple
236 Topic 2: update, time, working, data, like, deal, work, load, open, need
237 Topic 3: awesome, thanks, service, bad, slow, helpful, customer, best, apps, want
238 Topic 4: work, doesnt, useless, doesnt work, nice, open, bill, pay, never, cool
239 Topic 5: great, update, love, log, time, data, keep, much, say, let
240
241 3. Performing NMF with 5 topics...
242
243 NMF Topics:
244 Topic 1: good, good service, service, thanks, pretty good, deal, really good, really, pretty,
       thank
245 Topic 2: update, use, data, time, even, open, account, easy, phone, top
246 Topic 3: great, easy, thanks, great service, work great, service, great easy, use, easy use,
       love
247 Topic 4: awesome, thanks, easy, deal, service, love, thank, cool, guy, easy use
248 Topic 5: work, doesnt, doesnt work, never work, never, work well, well, time, work time, still
       doesnt
249
250 4. Assigning topics to reviews...
251
252 Topic Distribution:
253 dominant_topic
254 0.0    1518
255 1.0    2225
256 2.0    1245
257 3.0    1279
```

```
258  4.0     3119
259  Name: count, dtype: int64
260
261  ================================================================================
262  TOPIC IMPACT ANALYSIS - FEATURE IMPORTANCE & RATING CORRELATION
263  ================================================================================
264
265  ================================================================================
266  SECTION 7.1: ASSIGNING DOMINANT TOPICS TO REVIEWS
267  ================================================================================
268
269  Total reviews processed: 9,386
270
271  Reviews per Topic:
272    Topic 0:  1,518 reviews ( 16.2%)
273    Topic 1:  2,225 reviews ( 23.7%)
274    Topic 2:  1,245 reviews ( 13.3%)
275    Topic 3:  1,279 reviews ( 13.6%)
276    Topic 4:  3,119 reviews ( 33.2%)
277
278  ================================================================================
279  SECTION 7.2: AVERAGE RATING BY TOPIC
280  ================================================================================
281
282  Overall Average Rating (All Reviews): 2.42/5.0
283
284  Average Rating by Topic:
285
286                 Mean_Rating  Median_Rating  Std_Dev  Review_Count  Min_Rating  Max_Rating
287  dominant_topic
288  0                    3.728            5.0    1.622          1518           1           5
289  1                    2.076            1.0    1.510          2225           1           5
290  2                    2.947            3.0    1.802          1245           1           5
291  3                    2.135            1.0    1.564          1279           1           5
292  4                    1.946            1.0    1.449          3119           1           5
293
294
295  Best Topic: Topic 0 (Avg: 3.73)
296  Worst Topic: Topic 4 (Avg: 1.95)
297  Difference: 1.78 stars
298
299  ================================================================================
300  SECTION 7.3: EXTRACTING TOPIC KEYWORDS
301  ================================================================================
302
303  Top 15 Words per Topic:
304
305  Topic 0:
306    good, easy, use, easy use, download, happy, thank, fine, uninstall, simple
307
308  Topic 1:
309    update, time, working, data, like, deal, work, load, open, need
310
311  Topic 2:
312    awesome, thanks, service, bad, slow, helpful, customer, best, apps, want
313
314  Topic 3:
315    work, doesnt, useless, doesnt work, nice, open, bill, pay, never, cool
316
317  Topic 4:
318    great, update, love, log, time, data, keep, much, say, let
319
320
321  ================================================================================
322  SECTION 7.4: AVERAGE RATING BY KEYWORD PRESENCE
323  ================================================================================
324
325  Average Rating by Keyword Presence:
326
```

	topic_id	keywords	with_keywords_avg	without_keywords_avg
327				
		rating_difference		
328	0	good, easy, use, easy use, download	2.464334	2.396932
		0.067402		
329	1	update, time, working, data, like	1.843696	3.142448
		-1.298752		
330	2	awesome, thanks, service, bad, slow	2.199101	2.523641
		-0.324540		
331	3	work, doesnt, useless, doesnt work, nice	1.783194	2.886727
		-1.103533		
332	4	great, update, love, log, time	2.021549	2.964277
		-0.942727		

```
333
334 ================================================================================
335 SECTION 7.5: STATISTICAL SIGNIFICANCE TESTING
336 ================================================================================
337
338 Overall Average Rating: 2.42/5.0
339
340 Independent Samples T-Tests (Topic vs. Other Reviews):
341
342 --------------------------------------------------------------------------------
343 Topic 0 (Keywords: good, easy, use, easy use, download...)
344   Reviews with keywords: 3715 (Avg: 2.46)
345   Reviews without keywords: 5671 (Avg: 2.40)
346   T-statistic: 1.8914
347   P-value: 0.058607
348   Cohen's d: 0.0399
349   Result: No significant difference
350 --------------------------------------------------------------------------------
351 Topic 1 (Keywords: update, time, working, data, like...)
352   Reviews with keywords: 5195 (Avg: 1.84)
353   Reviews without keywords: 4191 (Avg: 3.14)
354   T-statistic: -38.9230
355   P-value: 0.000000
356   Cohen's d: -0.8319
357   Result: Significantly lower ratings
358 --------------------------------------------------------------------------------
359 Topic 2 (Keywords: awesome, thanks, service, bad, slow...)
360   Reviews with keywords: 2893 (Avg: 2.20)
361   Reviews without keywords: 6493 (Avg: 2.52)
362   T-statistic: -8.8223
363   P-value: 0.000000
364   Cohen's d: -0.1929
365   Result: Significantly lower ratings
366 --------------------------------------------------------------------------------
367 Topic 3 (Keywords: work, doesnt, useless, doesnt work, nice...)
368   Reviews with keywords: 3939 (Avg: 1.78)
369   Reviews without keywords: 5447 (Avg: 2.89)
370   T-statistic: -34.5602
371   P-value: 0.000000
372   Cohen's d: -0.6900
373   Result: Significantly lower ratings
374 --------------------------------------------------------------------------------
375 Topic 4 (Keywords: great, update, love, log, time...)
376   Reviews with keywords: 5383 (Avg: 2.02)
377   Reviews without keywords: 4003 (Avg: 2.96)
378   T-statistic: -27.1408
379   P-value: 0.000000
380   Cohen's d: -0.5806
381   Result: Significantly lower ratings
382 --------------------------------------------------------------------------------
383
384 ================================================================================
385 SECTION 7.6: IMPACT SUMMARY
386 ================================================================================
387
388                         POSITIVE IMPACT TOPICS
389 ================================================================================
```

```
390  Topics associated with significantly HIGHER user ratings:
391
392     No topics with significant positive impact found.
393
394
395                         NEGATIVE IMPACT TOPICS
396  ================================================================================
397  Topics associated with significantly LOWER user ratings:
398
399  Topic 1: update, time, working, data, like
400     Average Rating: 1.84/5.0
401     vs. Other Topics: 3.14/5.0
402     Difference: -1.30 stars (NEGATIVE)
403     Effect Size (Cohen's d): -0.8319 (large)
404     Sample: 5195 reviews
405     P-value: 1.63e-302
406
407  Topic 3: work, doesnt, useless, doesnt work, nice
408     Average Rating: 1.78/5.0
409     vs. Other Topics: 2.89/5.0
410     Difference: -1.10 stars (NEGATIVE)
411     Effect Size (Cohen's d): -0.6900 (medium)
412     Sample: 3939 reviews
413     P-value: 1.84e-246
414
415  Topic 4: great, update, love, log, time
416     Average Rating: 2.02/5.0
417     vs. Other Topics: 2.96/5.0
418     Difference: -0.94 stars (NEGATIVE)
419     Effect Size (Cohen's d): -0.5806 (medium)
420     Sample: 5383 reviews
421     P-value: 4.82e-155
422
423  Topic 2: awesome, thanks, service, bad, slow
424     Average Rating: 2.20/5.0
425     vs. Other Topics: 2.52/5.0
426     Difference: -0.32 stars (NEGATIVE)
427     Effect Size (Cohen's d): -0.1929 (negligible)
428     Sample: 2893 reviews
429     P-value: 1.46e-18
430
431
432  Topics with NO significant impact (p ≥ 0.05): 1
433
434  Generating impact visualizations...
435  ✓ Saved: topic_impact_analysis.png
436
437  ================================================================================
438  ANALYSIS COMPLETE
439  ================================================================================
440
441  Generated Files:
442     1. topic_statistics.csv
443     2. keyword_presence_statistics.csv
444     3. topic_impact_test_results.csv
445     4. topic_impact_analysis.png
446       topic_id                              keywords  topic_avg  other_avg  \
447  1          1          update, time, working, data, like   1.843696   3.142448
448  3          3   work, doesnt, useless, doesnt work, nice   1.783194   2.886727
449  4          4                   great, update, love, log, time   2.021549   2.964277
450  2          2          awesome, thanks, service, bad, slow   2.199101   2.523641
451
452       topic_n  other_n  t_statistic        p_value  cohens_d  is_significant  \
453  1       5195     4191   -38.923003  1.633281e-302  -0.831912           True
454  3       3939     5447   -34.560162  1.836511e-246  -0.690040           True
455  4       5383     4003   -27.140756  4.819658e-155  -0.580565           True
456  2       2893     6493    -8.822262   1.458893e-18  -0.192860           True
457
458         impact
```

```
459 1    NEGATIVE
460 3    NEGATIVE
461 4    NEGATIVE
462 2    NEGATIVE
463    topic_id                              keywords  topic_avg  other_avg  \
464 0          0       good, easy, use, easy use, download   2.464334   2.396932
465 1          1             update, time, working, data, like   1.843696   3.142448
466 2          2         awesome, thanks, service, bad, slow   2.199101   2.523641
467 3          3  work, doesnt, useless, doesnt work, nice   1.783194   2.886727
468 4          4              great, update, love, log, time   2.021549   2.964277
469
470    topic_n  other_n  t_statistic        p_value  cohens_d  is_significant  \
471 0     3715     5671     1.891402   5.860694e-02  0.039904           False
472 1     5195     4191   -38.923003  1.633281e-302 -0.831912            True
473 2     2893     6493    -8.822262   1.458893e-18 -0.192860            True
474 3     3939     5447   -34.560162  1.836511e-246 -0.690040            True
475 4     5383     4003   -27.140756  4.819658e-155 -0.580565            True
476
477             impact
478 0  NOT SIGNIFICANT
479 1         NEGATIVE
480 2         NEGATIVE
481 3         NEGATIVE
482 4         NEGATIVE
483
484 ================================================================================
485 SECTION 7: MACHINE LEARNING MODELS
486 ================================================================================
487
488 1. Preparing training data...
489 Training set size: 7508
490 Test set size: 1878
491
492 2. Vectorizing text data...
493
494 3. Training and evaluating models...
495
496    Training Logistic Regression...
497    Cross-validating Logistic Regression...
498    Accuracy: 0.8360
499    Precision: 0.8136
500    Recall: 0.8360
501    F1-Score: 0.8061
502    CV F1-Score: 0.8097 (±0.0068)
503 Training completed in 0.26 seconds.
504
505    Training Random Forest...
506    Cross-validating Random Forest...
507    Accuracy: 0.8168
508    Precision: 0.7812
509    Recall: 0.8168
510    F1-Score: 0.7867
511    CV F1-Score: 0.7815 (±0.0025)
512 Training completed in 7.91 seconds.
513
514    Training SVM...
515    Cross-validating SVM...
516    Accuracy: 0.8397
517    Precision: 0.8203
518    Recall: 0.8397
519    F1-Score: 0.8108
520    CV F1-Score: 0.8090 (±0.0031)
521 Training completed in 20.65 seconds.
522
523    Training Naive Bayes...
524    Cross-validating Naive Bayes...
525    Accuracy: 0.8094
526    Precision: 0.8068
527    Recall: 0.8094
```

```
528     F1-Score: 0.7757
529     CV F1-Score: 0.7776 (±0.0109)
530 Training completed in 0.01 seconds.
531
532   Training Gradient Boosting...
533   Cross-validating Gradient Boosting...
534   Accuracy: 0.7902
535   Precision: 0.7697
536   Recall: 0.7902
537   F1-Score: 0.7586
538   CV F1-Score: 0.7581 (±0.0065)
539 Training completed in 18.54 seconds.
540
541 4. Best Model: SVM
542    F1-Score: 0.8108
543
544 5. Detailed Classification Report for SVM:
545             precision    recall   f1-score    support
546
547    Negative      0.84      0.96      0.89      1152
548     Neutral      0.56      0.04      0.07       138
549    Positive      0.84      0.80      0.82       588
550
551    accuracy                          0.84      1878
552   macro avg      0.75      0.60      0.59      1878
553 weighted avg     0.82      0.84      0.81      1878
554
555
556 ==============================================================================
557 SECTION 8: KEY BUSINESS INSIGHTS
558 ==============================================================================
559
560 1. TOP ISSUES IN NEGATIVE REVIEWS:
561
562 Detailed Issue Breakdown:
563            Issue Category   Mentions   Percentage
564   Installation/Compatibility     1340    23.280056
565             Update Problems       1219    21.177901
566           Data Usage Issues        738    12.821404
567        App Crashes/Freezing        732    12.717165
568      Billing/Charging Issues       728    12.647672
569      Customer Service Issues       680    11.813760
570              Account Issues        499     8.669215
571         Feature Not Working        475     8.252259
572       Balance/Credit Issues        438     7.609451
573     User Interface Problems        339     5.889507
574   Network/Connection Issues        316     5.489924
575  Login/Authentication Issues       311     5.403058
576            Slow Performance        254     4.412787
577            Missing Features        111     1.928423
578         Auto Top-up Problems         16     0.277971
579
580
581 Top 5 Issues with Examples:
582
583 Installation/Compatibility: 1340 mentions (23.3%)
584   Example: Have had lots of problems since 3 months ago. Had to reinstall several times before
   it would work. Now the app has hung again. Do I try yet again or w...
585
586 Update Problems: 1219 mentions (21.2%)
587   Example: The latest update is unable to be 'updated' or even installed and reinstalled on my
   sons phone, which is the same phone I have Samsung A14....
588
589 Data Usage Issues: 738 mentions (12.8%)
590   Example: Have had lots of problems since 3 months ago. Had to reinstall several times before
   it would work. Now the app has hung again. Do I try yet again or w...
591
592 App Crashes/Freezing: 732 mentions (12.7%)
593   Example: data speeds are terrible, even in the centre of auckland, and still unable to log
```

```
593  into the app, it just sits on loading forever, correct email, correc...
594
595  Billing/Charging Issues: 728 mentions (12.6%)
596     Example: Charges you 13 months a year thinking each month is February, every 28 days....
597
598
599  2. SENTIMENT BY APP VERSION:
600  sentiment_label  Negative  Neutral  Positive
601  major_version
602  Unknown            49.73    7.37    42.90
603  v1                 29.01   13.74    57.25
604  v2                 55.86   10.76    33.38
605  v3                 68.51    8.22    23.27
606  v4                 56.22    6.99    36.79
607  v5                 56.05    5.78    38.17
608  v6                 79.16    3.82    17.02
609  v7                 65.88    4.73    29.39
610
611  3. IMPACT OF COMPANY RESPONSES:
612                  mean   count
613  has_reply
614  0            2.511907   6257
615  1            2.247044   3129
616
617  4. SENTIMENT TRENDS OVER TIME:
618  sentiment_label  Negative   Neutral   Positive
619  review_date
620  2025-04            11         2         13
621  2025-05            14         1          9
622  2025-06            13         2         10
623  2025-07            17         0          8
624  2025-08            22         3          8
625  2025-09            10         1          6
626
627
628  Processed data saved to 'one_nz_processed_reviews.csv'
629
630  Saving ML results for visualization...
631  ML results saved to 'ml_results.pkl'
632
633  ================================================================================
634  ANALYSIS COMPLETE!
635  ================================================================================
636
637  Next step: Run the visualization notebook to generate all charts
638
639  Process finished with exit code 0
640
```