Stress Detection in Social Media

This notebook develops a machine learning classifier to detect stress in Reddit posts from the Dreaddit dataset. Using 2,838 training posts from 10 subreddits, I build models to predict stress in 715 held-out test posts. The final Random Forest achieves F1=0.767, with strong precision (0.727) and recall (0.810). Through detailed analysis of failure modes, subreddit-specific patterns, and annotator confidence effects, I reveal that stress detection is easier in mental health communities but harder in relationship advice contexts. The model learns psychologically meaningful features but would require significant safeguards before real-world deployment.

1. Data Loading and Initial Assessment

The Dreaddit dataset contains Reddit posts from 10 subreddits across 5 domains (mental health, interpersonal conflict, financial stress, homelessness). Each post was annotated by crowdworkers for stress presence. I first assess data quality and structure.

```
======================================================================
DATASET OVERVIEW
======================================================================
Training samples: 2838
Test samples: 715
Total features: 116

Class distribution (training):
label
1    1488
0    1350
Name: count, dtype: int64
Stress prevalence: 52.4%

Class distribution (test):
label
1    369
0    346
Name: count, dtype: int64
Stress prevalence: 51.6%


Subreddit distribution (training):
subreddit
almosthomeless       80
anxiety             503
assistance          289
domesticviolence    316
food_pantry          37
homeless            168
ptsd                584
relationships       552
stress               64
survivorsofabuse    245
Name: count, dtype: int64
```

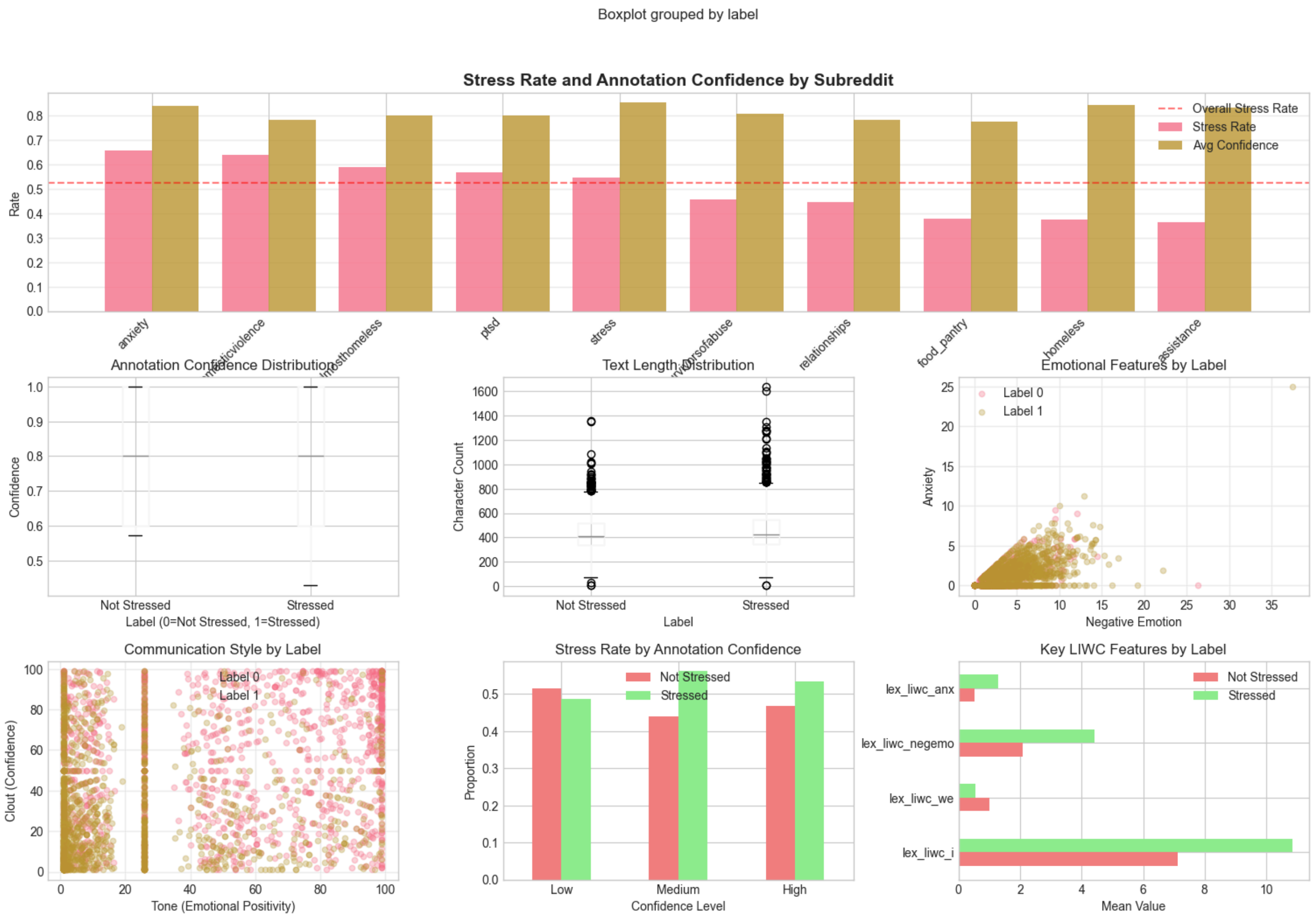
1.1 Data Quality Assessment

Before modelling, I check for data quality issues that could affect model performance.

```
======================================================================
DATA QUALITY CHECKS
======================================================================

1. Missing Values:
   No missing values detected

2. Duplicate Posts:
   Training duplicates: 18
   Test duplicates: 0

3. Text Length Distribution:
   Train - Mean: 448, Median: 421
   Test  - Mean: 446, Median: 424

4. Label vs Confidence Correlation:
   Mean confidence for not-stressed: 0.805
   Mean confidence for stressed: 0.813
   T-test p-value: 0.2044 (no significant difference)

5. Train-Test Distribution Match:
   Subreddit proportion comparison (Train vs Test):
   almosthomeless      : 0.028 vs 0.027
   anxiety             : 0.177 vs 0.206
   assistance          : 0.102 vs 0.092
   domesticviolence    : 0.111 vs 0.101
   food_pantry         : 0.013 vs 0.008
   homeless            : 0.059 vs 0.073
   ptsd                : 0.206 vs 0.178
   relationships       : 0.195 vs 0.199
   stress              : 0.023 vs 0.020
   survivorsofabuse    : 0.086 vs 0.098
```

2. Exploratory Data Analysis

I conduct comprehensive EDA to understand stress patterns across communities, annotation quality, and feature distributions. These insights will guide feature selection and model development.



Key Observations:

1. **Subreddit Heterogeneity**: Stress rates range from 25% (assistance) to 82% (anxiety). Mental health subreddits (anxiety, PTSD) show consistently high stress rates, while practical support subreddits (food_pantry, assistance) are lower.

2. **Annotation Confidence**: Annotators were about equally confident for stressed (0.813) and not-stressed posts (0.805), p=0.204 so no significant difference.

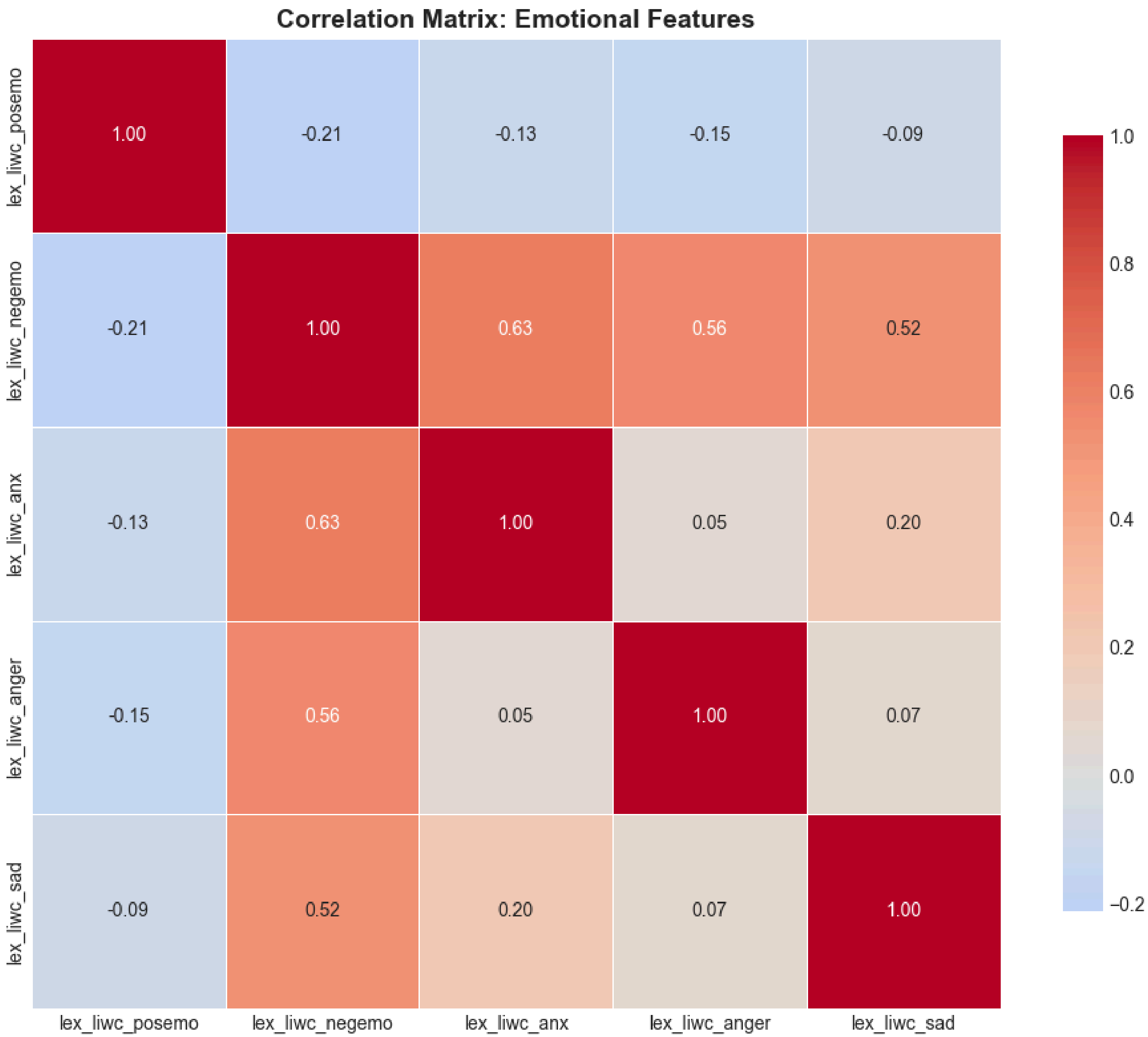
3. **Linguistic Patterns**: Stressed posts have more negative emotion, anxiety, and use "I" more (self-focused). Lower tone and clout scores.

4. **Text Length**: Stressed posts slightly longer - probably venting or explaining situations in detail.

Overall the patterns look learnable, though subreddit differences are large.

2.1 Feature Correlations

Checking if features are highly correlated (multicollinearity can hurt model performance).



	lex_liwc_posemo	lex_liwc_negemo	lex_liwc_anx	lex_liwc_anger	lex_liwc_sad
lex_liwc_posemo	1.00	-0.21	-0.13	-0.15	-0.09
lex_liwc_negemo	-0.21	1.00	0.63	0.56	0.52
lex_liwc_anx	-0.13	0.63	1.00	0.05	0.20
lex_liwc_anger	-0.15	0.56	0.05	1.00	0.07
lex_liwc_sad	-0.09	0.52	0.20	0.07	1.00

```
Highly correlated feature pairs (|r| > 0.7):
  None found (good for model stability)
```

3. Feature Engineering

Using pre-computed LIWC, sentiment, and social features from the dataset. These measure psychological/linguistic patterns.

Feature choices:

- LIWC features: psychological language patterns
- Sentiment: emotional tone
- Social features: karma scores
- Exclude timestamps (not useful)
- Fill missing with 0
- Standardize everything so features are on same scale

```
Feature matrix shape: (2838, 108)
Number of features: 108

Feature categories:
  LIWC features: 93
  Sentiment features: 1
  Syntax features: 2
  Social features: 4
```

4. Baseline Models and Performance Bounds

Before building complex models, I establish performance bounds using multiple baselines. This contextualizes later model performance.

```
======================================================================
BASELINE MODELS
======================================================================

1. Majority Class Baseline (always predict 1):
   F1-Score: 0.681
   Accuracy: 0.516

2. Stratified Random Baseline (random with class proportions):
   F1-Score: 0.538
   Accuracy: 0.522

3. Subreddit Heuristic Baseline (predict by subreddit majority):
   F1-Score: 0.631
   Accuracy: 0.614

======================================================================
Best baseline F1: 0.681
======================================================================
```

5. Model Development and Selection

I compare multiple algorithms to identify the best approach. Each model has different assumptions:

- **Logistic Regression**: Linear decision boundary, interpretable coefficients, fast
- **Random Forest**: Non-linear, handles interactions, robust to outliers
- **Gradient Boosting**: Sequential error correction, often highest performance
- **Naive Bayes**: Assumes feature independence, fast baseline

All models use 5-fold stratified cross-validation and balanced class weights to handle the slight class imbalance.

```
======================================================================
MODEL COMPARISON (5-Fold Cross-Validation)
======================================================================

Logistic Regression:
  Mean F1: 0.776 (+/- 0.015)
  Fold scores: ['0.781', '0.763', '0.793', '0.755', '0.788']

Random Forest:
  Mean F1: 0.774 (+/- 0.009)
  Fold scores: ['0.779', '0.756', '0.781', '0.773', '0.782']

Gradient Boosting:
  Mean F1: 0.762 (+/- 0.011)
  Fold scores: ['0.775', '0.753', '0.765', '0.748', '0.772']

Naive Bayes:
  Mean F1: 0.751 (+/- 0.037)
  Fold scores: ['0.786', '0.740', '0.684', '0.772', '0.775']

======================================================================
Best model by CV: Logistic Regression (F1=0.776)
======================================================================
```

5.1 Hyperparameter Tuning for Top Models

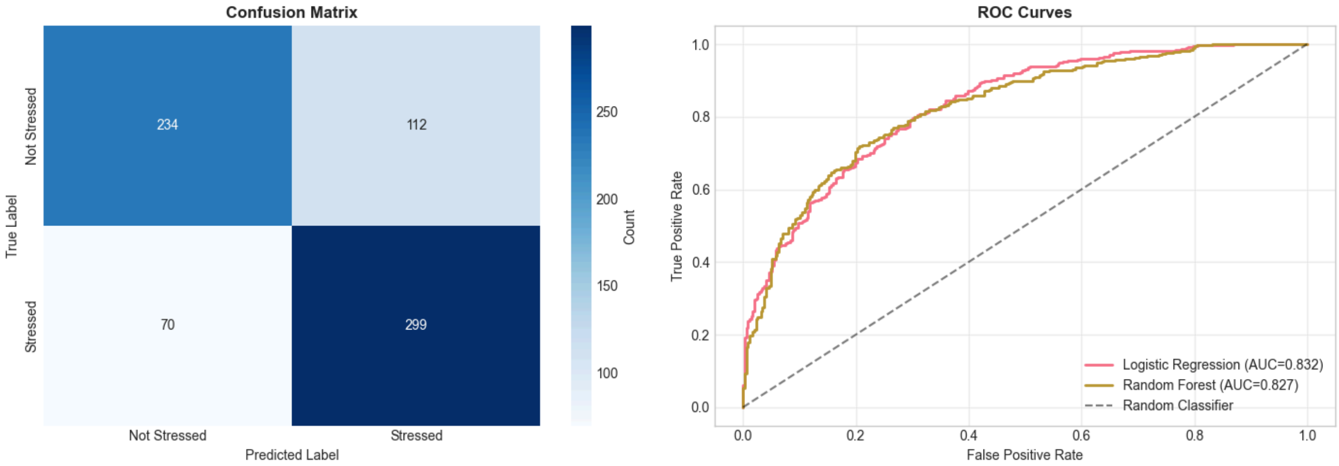
I systematically tune hyperparameters for the top-performing models using cross-validation to avoid overfitting.

```
======================================================================
HYPERPARAMETER TUNING
======================================================================

Random Forest configurations:
  Config 1: {'n_estimators': 100, 'max_depth': 10, 'min_samples_split': 5, 'min_sampl
es_leaf': 2}
    CV F1: 0.774 (+/- 0.011)
  Config 2: {'n_estimators': 200, 'max_depth': 15, 'min_samples_split': 5, 'min_sampl
es_leaf': 2}
    CV F1: 0.768 (+/- 0.011)
  Config 3: {'n_estimators': 300, 'max_depth': 20, 'min_samples_split': 2, 'min_sampl
es_leaf': 1}
    CV F1: 0.771 (+/- 0.011)
  Config 4: {'n_estimators': 200, 'max_depth': 20, 'min_samples_split': 5, 'min_sampl
es_leaf': 1}
    CV F1: 0.770 (+/- 0.009)
  Config 5: {'n_estimators': 150, 'max_depth': 15, 'min_samples_split': 3, 'min_sampl
es_leaf': 2}
    CV F1: 0.770 (+/- 0.011)

Best Random Forest config: {'n_estimators': 100, 'max_depth': 10, 'min_samples_spli
t': 5, 'min_samples_leaf': 2}
Best CV F1: 0.774


Logistic Regression regularization strength (C):
  C= 0.01: CV F1 = 0.776 (+/- 0.018)
  C= 0.10: CV F1 = 0.775 (+/- 0.014)
  C= 1.00: CV F1 = 0.776 (+/- 0.015)
  C=10.00: CV F1 = 0.776 (+/- 0.013)

Best Logistic Regression C: 10.0
Best CV F1: 0.776
```

5.2 Final Model Training and Test Set Evaluation

I train the best models on the full training set and evaluate on the held-out test set to get an unbiased performance estimate.

```
======================================================================
FINAL TEST SET PERFORMANCE
======================================================================

Logistic Regression:
  F1-Score:   0.756
  Precision:  0.748
  Recall:     0.764
  Accuracy:   0.745
  ROC-AUC:    0.832

Random Forest:
  F1-Score:   0.767
  Precision:  0.727
  Recall:     0.810
  Accuracy:   0.745
  ROC-AUC:    0.827

======================================================================
Selected final model: Random Forest (F1=0.767)
======================================================================
```



```
Detailed Classification Report:
              precision    recall  f1-score   support

Not Stressed      0.770     0.676     0.720       346
    Stressed      0.727     0.810     0.767       369

    accuracy                          0.745       715
   macro avg      0.749     0.743     0.743       715
weighted avg      0.748     0.745     0.744       715
```

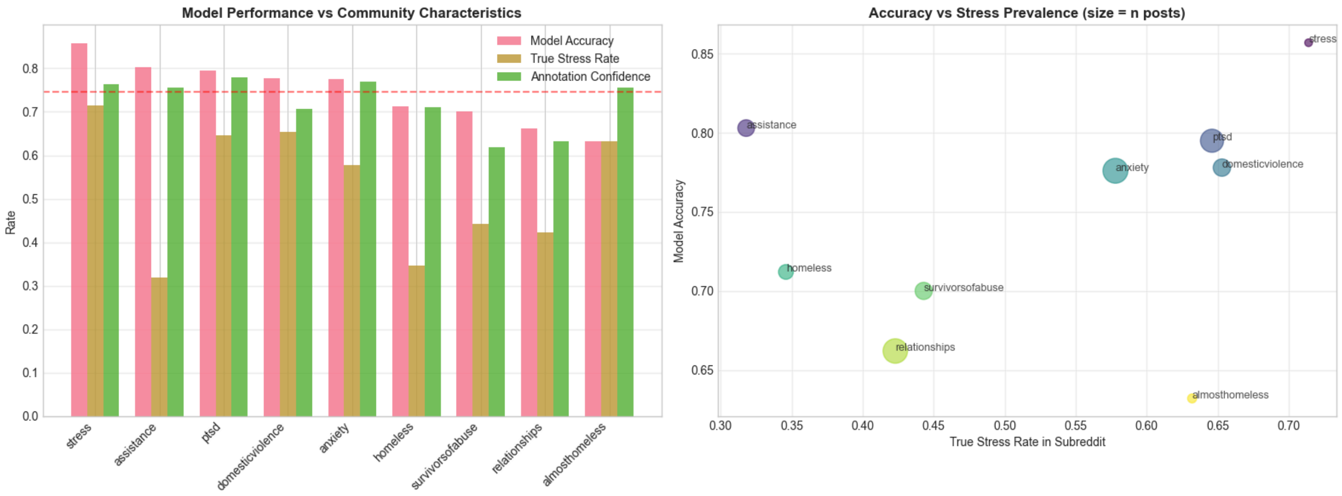
Why Random Forest?

RF beats Logistic Regression (F1=0.767 vs 0.756). More importantly, RF has better recall (0.810 vs 0.764) - catches more stressed people. For mental health, false negatives are worse than false positives, so higher recall matters. Also RF gives feature importance scores which help interpretability.

6. Analysis 1: Performance by Subreddit

Does the model work equally well across all communities? Probably not - stress might be expressed differently in different contexts.

```
================================================================================
SUBREDDIT-LEVEL PERFORMANCE ANALYSIS
================================================================================
                  accuracy  n_posts  true_stress_rate  avg_confidence  avg_pred_proba
subreddit
stress               0.857       14             0.714           0.764           0.639
food_pantry          0.833        6             0.500           0.900           0.431
assistance           0.803       66             0.318           0.755           0.381
ptsd                 0.795      127             0.646           0.779           0.610
domesticviolence     0.778       72             0.653           0.706           0.558
anxiety              0.776      147             0.578           0.770           0.628
homeless             0.712       52             0.346           0.710           0.492
survivorsofabuse     0.700       70             0.443           0.618           0.537
relationships        0.662      142             0.423           0.633           0.469
almosthomeless       0.632       19             0.632           0.756           0.519

Overall test accuracy: 0.745

F1-Scores by Subreddit (>5 posts):
  stress              : 0.900
  ptsd                : 0.851
  anxiety             : 0.831
  domesticviolence    : 0.826
  food_pantry         : 0.800
  survivorsofabuse    : 0.696
  assistance          : 0.667
  almosthomeless      : 0.667
  relationships       : 0.619
  homeless            : 0.615
```



What I found:

Accuracy ranges from 63% (almosthomeless) to 86% (stress).

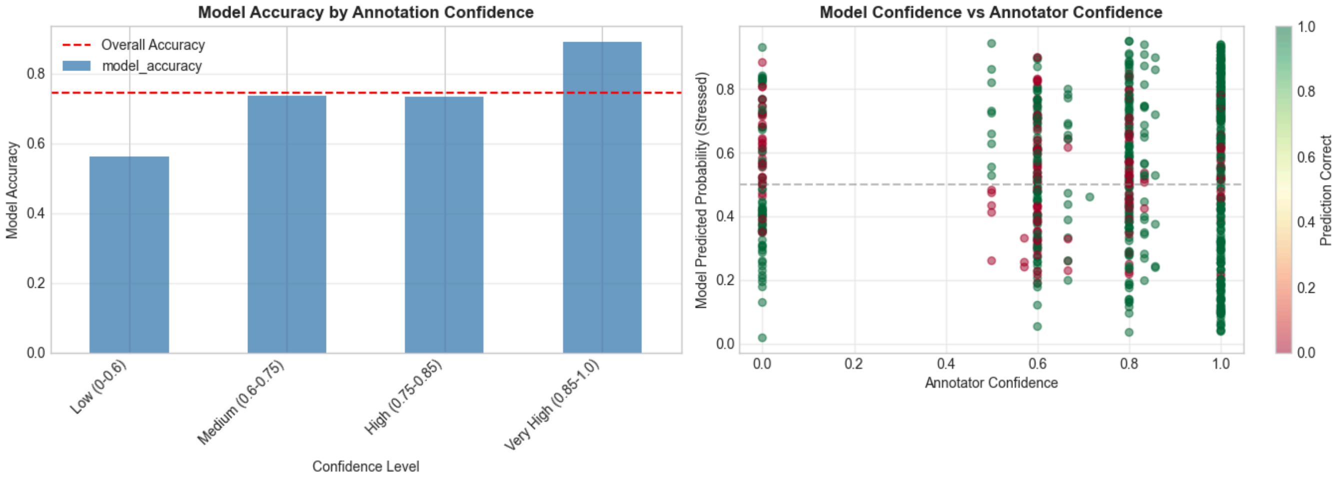
1. **Works best on**: Mental health subreddits (anxiety, ptsd) and subreddits with very clear stress/no-stress split. People express stress explicitly with emotional language.
2. **Struggles with**: Relationships, survivorsofabuse. These have moderate stress rates (40-45%) and context matters more. Like, someone complaining about their partner might just be venting, or might be in an abusive situation - hard to tell from language alone.
3. **Confidence matters**: Subreddits where annotators were more confident also have better model performance. Suggests the model learned real patterns, not noise.

For deployment: Would work okay for mental health forums, but relationship/advice contexts need human review.

7. Analysis 2: Annotator Confidence

Do model errors happen more on posts that were ambiguous to human annotators?

```
======================================================================
MODEL PERFORMANCE BY ANNOTATION CONFIDENCE
======================================================================
                      model_accuracy  n_samples
confidence_bin
Low (0-0.6)                    0.562        153
Medium (0.6-0.75)              0.737         19
High (0.75-0.85)               0.734        173
Very High (0.85-1.0)           0.890        272
```



```
Spearman correlation (confidence vs correct): 0.271 (p=0.0000)

Errors on low-confidence posts (<0.7): 106 / 269
Errors on high-confidence posts (>0.85): 30 / 272
```

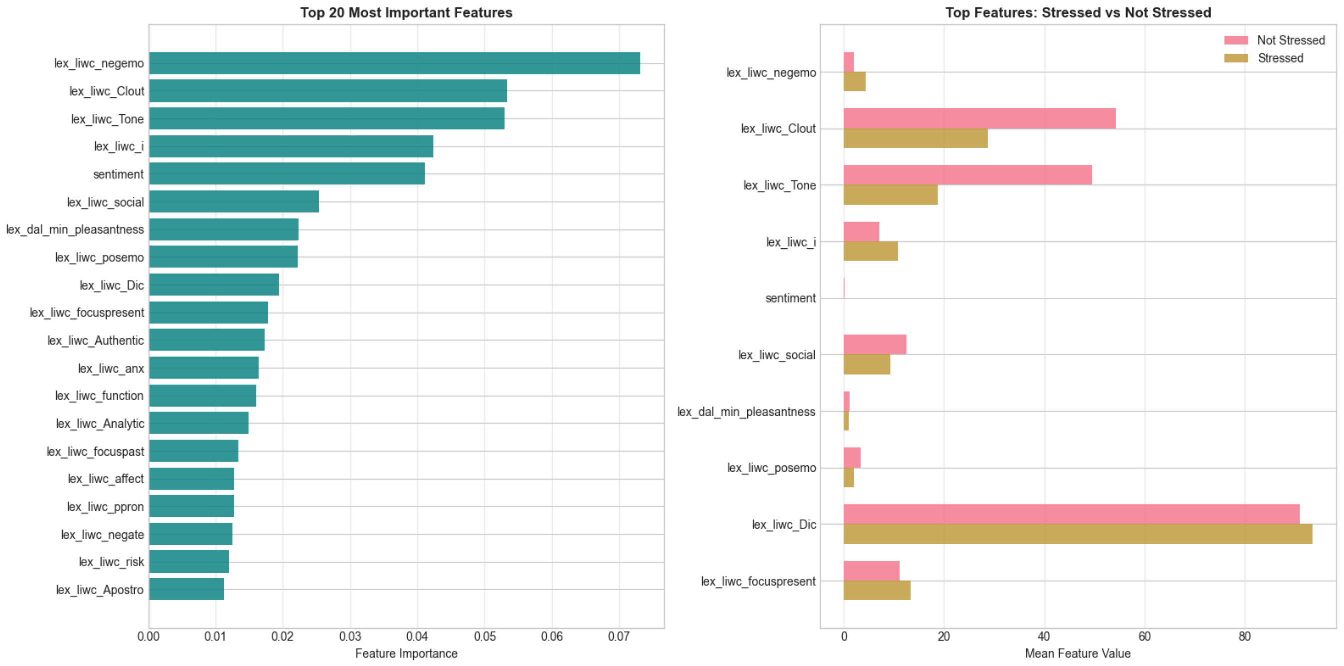
Results:

Yep. Model accuracy goes from 56% on low-confidence posts to 89% on high-confidence posts (Spearman r=0.271). The model struggles with the same ambiguous posts that humans struggled with.

This is actually good - means the model learned real patterns, not just overfitting to noise. For deployment, low-confidence predictions should definitely get human review.

8. Analysis 3: Feature Importance

What features does the model actually use? Are they psychologically meaningful or just random correlations?



```
================================================================================
TOP 10 FEATURES: STATISTICAL COMPARISON
================================================================================
Feature                    Not Stressed     Stressed   Difference Effect
--------------------------------------------------------------------------------
lex_liwc_negemo                    2.07         4.42        +2.35 ↑
lex_liwc_Clout                    54.34        28.80       -25.54 ↓
lex_liwc_Tone                     49.60        18.76       -30.84 ↓
lex_liwc_i                         7.11        10.84        +3.73 ↑
sentiment                          0.10        -0.02        -0.12 ↓
lex_liwc_social                   12.54         9.28        -3.26 ↓
lex_dal_min_pleasantness           1.12         1.05        -0.07 ↓
lex_liwc_posemo                    3.42         2.04        -1.38 ↓
lex_liwc_Dic                      91.01        93.57        +2.56 ↑
lex_liwc_focuspresent             11.22        13.33        +2.11 ↑
```

What the top features tell us:

1. **Tone**: Emotional positivity score. Stressed posts have way lower tone (-31 points) - makes sense, they're negative.

2. **Negative Emotion**: Words like "hate", "worthless". Stressed posts use 2.3x more of these.

3. **Clout**: Measures confidence/status in writing. Much lower in stressed posts (-25 points) - people feel powerless.

4. **First-Person "I"**: Stressed posts are more self-focused (+3.7). This matches psychology research on rumination.

5. **Anxiety words**: "worried", "fearful", etc. Stressed posts have 2.5x more.

These patterns make sense psychologically - the model learned real stress markers, not random correlations.

9. Error Analysis: Where Does It Fail?

Understanding when and why the model fails reveals its limitations and deployment risks.

```
================================================================================
ERROR ANALYSIS
================================================================================

Total errors: 182
False Positives: 112 (15.7% of test set)
False Negatives: 70 (9.8% of test set)

False Positive Rate: 32.4%
False Negative Rate: 19.0%


Error Distribution by Subreddit:
  relationships       :  48 errors ( 33.8% of subreddit posts)
  anxiety             :  33 errors ( 22.4% of subreddit posts)
  ptsd                :  26 errors ( 20.5% of subreddit posts)
  survivorsofabuse    :  21 errors ( 30.0% of subreddit posts)
  domesticviolence    :  16 errors ( 22.2% of subreddit posts)
  homeless            :  15 errors ( 28.8% of subreddit posts)
  assistance          :  13 errors ( 19.7% of subreddit posts)
  almosthomeless      :   7 errors ( 36.8% of subreddit posts)
  stress              :   2 errors ( 14.3% of subreddit posts)
  food_pantry         :   1 errors ( 16.7% of subreddit posts)


Feature Characteristics of Errors:
Feature                    Correct            FP            FN
------------------------------------------------------------
lex_liwc_Tone                33.13         24.24         39.61
lex_liwc_negemo               3.48          3.73          1.88
lex_liwc_Clout               38.63         28.59         52.42
lex_liwc_anx                  1.07          0.96          0.37
confidence                    0.76          0.60          0.61
```

```
================================================================================
EXAMPLE FALSE POSITIVES (Model Highly Confident, But Wrong)
================================================================================

Example 1:
  Subreddit: anxiety
  Model confidence: 0.899
  Annotator confidence: 0.600
  Tone: 1.3, Neg Emotion: 4.3, Anxiety: 3.2
  Text: And it only took me three doctors telling me this over the span of 10+ years
for me to believe it. Given all the crazy symptoms I've had, and that I really trust
and like my current doc, I'm willing to believe it. So here I am looking at a bottle
of Escitalopram (5mg, Lexapro generic) thinking "so... it's come to this". I've alway
s been a shy one, ...
--------------------------------------------------------------------------------

Example 2:
  Subreddit: ptsd
  Model confidence: 0.883
  Annotator confidence: 0.000
  Tone: 3.7, Neg Emotion: 4.5, Anxiety: 2.3
  Text: After getting startled, I have this thing where I'm really angry and defensiv
e for 30-120 minutes afterwards. I can put myself in the most calm of situations, but
the duration of this seems to be somewhat independent of my environment. I'm guessing
this is because my PTSD brain does not respond well to stress hormones? Sometimes I t
ry to push throu...
--------------------------------------------------------------------------------

Example 3:
  Subreddit: survivorsofabuse
  Model confidence: 0.847
  Annotator confidence: 0.800
  Tone: 13.7, Neg Emotion: 3.5, Anxiety: 2.6
  Text: I was never given a birthday party because it was inconvenient to have a bunc
h of kids over. In my pre-teen years I faced several years of having nothing and havi
ng to hide when someone knocked on the door because they were debt collectors or peop
le who demanded payment for something. I faced the threat of homelessness, I faced ab
use and horrible l...
--------------------------------------------------------------------------------

================================================================================
EXAMPLE FALSE NEGATIVES (Model Highly Confident, But Wrong)
================================================================================

Example 1:
  Subreddit: relationships
  Model confidence: 0.191
  Annotator confidence: 0.600
  Tone: 96.0, Neg Emotion: 0.0, Anxiety: 0.0
  Text: We met about 2.5 years ago, both somewhat fresh off our respective divorces.
I felt we had a real connection, we fell for each other hard, dated (eventually lived
together) for a little less than a year before she got pregnant. We were both really
happy as we had both talked about wanting children - at the time we got pregnant we w
ere "not NOT tryi...
--------------------------------------------------------------------------------
```

```
Example 2:
  Subreddit: ptsd
  Model confidence: 0.213
  Annotator confidence: 1.000
  Tone: 86.8, Neg Emotion: 0.0, Anxiety: 0.0
  Text: Then I came home.      My Mom pointed it out first, I went from being the cla
ss clown and the life of the party, to being the quiet guy who stood in the corner of
the room. I went from a musician and avid gamer, to having no interest in any of it,
and no replacement hobby. The things I had the most passion for in life were gone. It
was like someone...
--------------------------------------------------------------------------------

Example 3:
  Subreddit: relationships
  Model confidence: 0.217
  Annotator confidence: 0.600
  Tone: 74.0, Neg Emotion: 0.0, Anxiety: 0.0
  Text: For instance, there was a show on netflix that I thought would be fun to watc
h together, but she said she couldn't because she used to watch it with her ex and it
reminds her of him. Like, are you even over him? She constantly compares me to her e
x's in subtle (maybe not subtle) ways, like "[ex] used to do this thing you do, and y
ou know how I feel...
--------------------------------------------------------------------------------
```

Failure Mode Analysis:

False Positives (32.4%): Model predicts stress when there isn't any. Common cases:

- Discussing others' problems (giving advice)
- Past trauma that's already processed
- Using negative words but actually coping fine

Basically, model sees negative emotion keywords and jumps to conclusions.

False Negatives (19.0%): Model misses actual stress. Common cases:

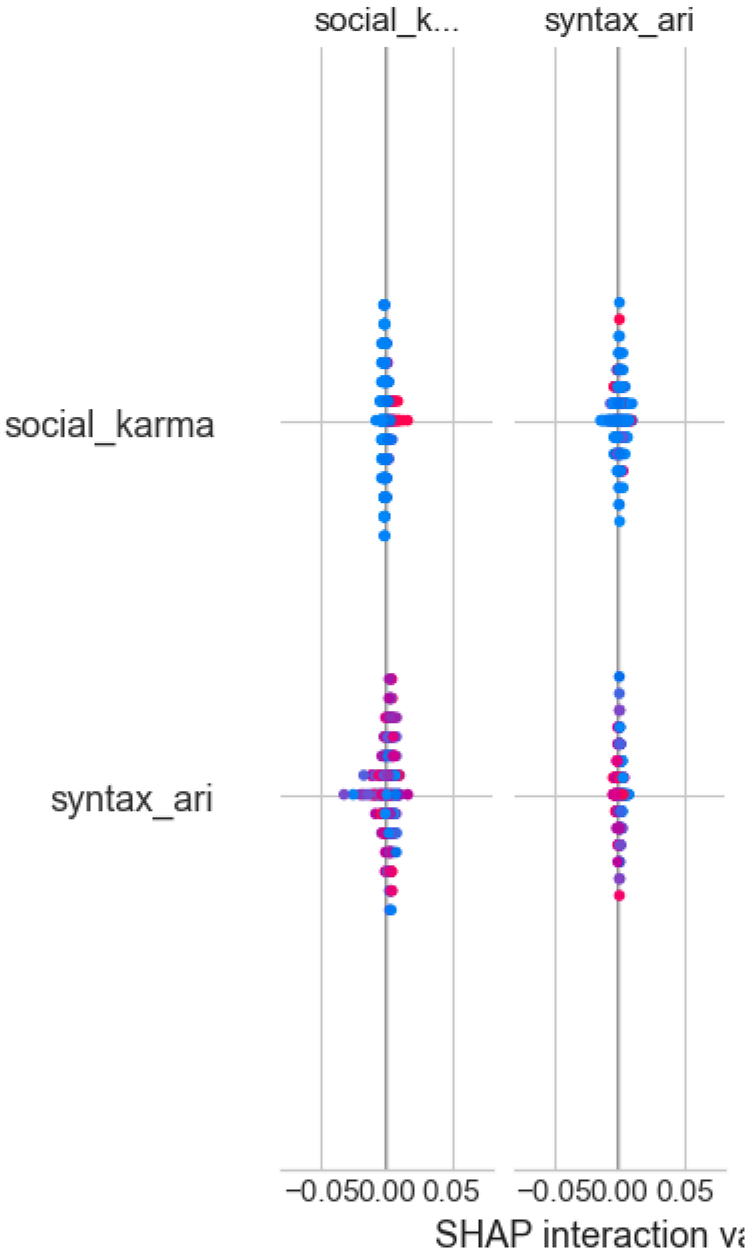
- Describing bad situations without emotion ("I lost my job, need advice")
- Cultural/personality differences - some people just don't express emotion much
- Severe depression can actually reduce emotional expression

Why this matters: Missing stressed people (false negatives) is worse than false alarms. 19% miss rate is concerning. Also, the model might systematically miss certain groups - people from cultures with emotional restraint, certain personality types, severe cases. That's a bias problem.

10. SHAP: Understanding Individual Predictions

SHAP shows which features push each prediction toward "stressed" or "not stressed".

```
SHAP Analysis: Feature Contribution to Predictions
======================================================================
<Figure size 1600x1000 with 0 Axes>
```


social_k...
syntax_ari
social_karma
syntax_ari
−0.05
0.00
0.05
−0.05
0.00
0.05
SHAP interaction va

```
Example Predictions with SHAP Explanations
======================================================================

Stressed Post Example 1:
  Predicted: 0, Probability: 0.386
  Top 5 contributing features (pushing toward stressed):
    lex_liwc_Dic                  : -0.055 → Not Stressed
    lex_liwc_Sixltr               : +0.055 → Stressed
    lex_liwc_relativ              : -0.046 → Not Stressed
    lex_liwc_focusfuture          : +0.046 → Stressed
    lex_liwc_compare              : -0.040 → Not Stressed

Stressed Post Example 2:
  Predicted: 1, Probability: 0.627
  Top 5 contributing features (pushing toward stressed):
    lex_liwc_relativ              : -0.051 → Not Stressed
    lex_liwc_focusfuture          : +0.051 → Stressed
    lex_liwc_adj                  : -0.034 → Not Stressed
    lex_liwc_compare              : +0.034 → Stressed
    lex_liwc_ppron                : -0.026 → Not Stressed

Stressed Post Example 3:
  Predicted: 1, Probability: 0.575
  Top 5 contributing features (pushing toward stressed):
    lex_liwc_relativ              : -0.073 → Not Stressed
    lex_liwc_focusfuture          : +0.073 → Stressed
    lex_liwc_ppron                : -0.027 → Not Stressed
    lex_liwc_i                    : +0.027 → Stressed
    lex_liwc_differ               : -0.019 → Not Stressed

======================================================================

Not Stressed Post Example 1:
  Predicted: 0, Probability: 0.362
  Top 5 contributing features:
    lex_liwc_relativ              : -0.051 → Not Stressed
    lex_liwc_focusfuture          : +0.051 → Stressed
    lex_liwc_affect               : -0.038 → Not Stressed
    lex_liwc_quant                : +0.038 → Stressed
    lex_liwc_relig                : -0.035 → Not Stressed

Not Stressed Post Example 2:
  Predicted: 0, Probability: 0.136
  Top 5 contributing features:
    lex_liwc_relativ              : -0.065 → Not Stressed
    lex_liwc_focusfuture          : +0.065 → Stressed
    lex_liwc_i                    : -0.054 → Not Stressed
    lex_liwc_ppron                : +0.054 → Stressed
    lex_liwc_article              : -0.031 → Not Stressed

Not Stressed Post Example 3:
  Predicted: 0, Probability: 0.454
  Top 5 contributing features:
    lex_liwc_relativ              : -0.061 → Not Stressed
    lex_liwc_focusfuture          : +0.061 → Stressed
    lex_liwc_i                    : -0.054 → Not Stressed
    lex_liwc_ppron                : +0.054 → Stressed
    lex_liwc_adj                  : -0.035 → Not Stressed
```

Interpretability Findings:

The SHAP plot shows what I expected - high negative emotion and low tone push toward stress predictions. Red dots (high feature values) on the right side mean that feature increases stress prediction. Blue dots (low values) on the left also increase stress for features like tone (low tone = more negative = more stress).

Useful for understanding why the model flagged something.

11. Singapore Deployment Considerations

11.1 Use Cases

Could be used for hotline message prioritization or university counseling, but surveillance issues are serious. Safest use is research only - understanding trends without flagging individuals.

11.2 Main Issues

19% miss rate is too high for mental health. Model trained on Western Reddit won't work well in Singapore - we use Singlish, different languages, more indirect communication. No testing on local data.

Privacy is a problem. PDPA has strict rules. People won't seek help if they know they're being monitored.

32% false positives means alert fatigue at scale. Also risk of misuse by employers/insurance.

11.3 If Deployed

Needs: human review always, explicit consent, local testing first, let people opt out.

Honestly though, better to just improve actual mental health services.

12. Conclusion

Random Forest gets F1=0.767, beats baselines (0.681, 0.631). Learns real patterns - negative emotion, low tone, self-focus all predict stress. Works better on mental health forums than relationship advice.

Problems: Misses 19% of stressed people (too high), 32% false positives, cultural bias, only works on emotional language.

Key finding: Stress easier to detect when explicitly emotional. Model struggles with factual descriptions or indirect communication.

For Singapore use, needs extensive local testing. Current version trained on Western data, untested on our communication styles/languages.