Task 1: Dialogue Act Recognition:

1. Feature Extraction:

a) Describe your custom feature sets (text-based features and speech-based features), the reasoning behind choosing them and the techniques used to extract them.

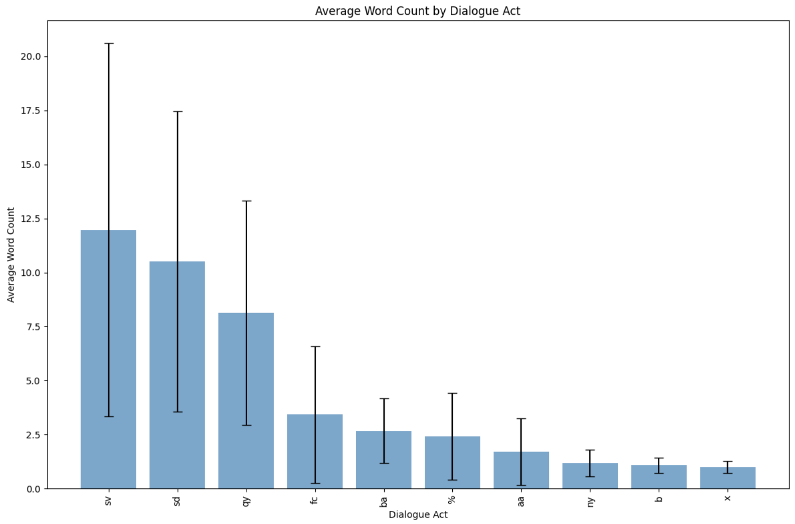
Text based Features:

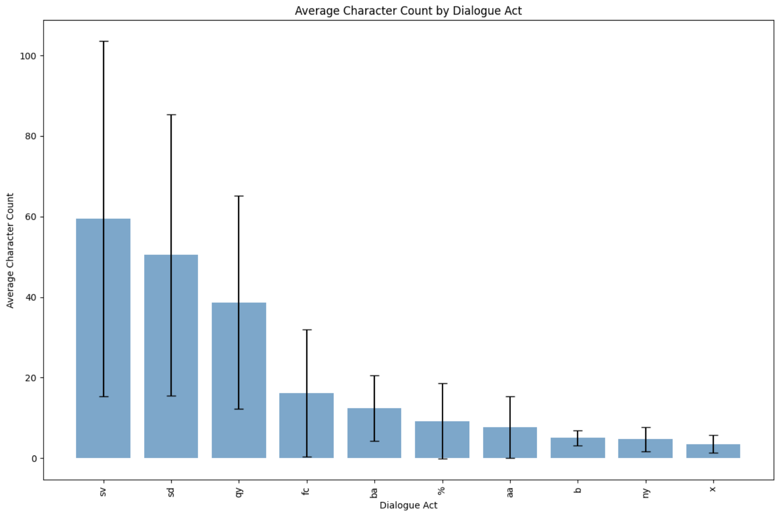
Features included: All LIWC features, word_count, character_count, has_question_mark, has_exclamation_mark, sentence transformer embedding of transcript, Tf-IDF n-grams

I decided to include all LIWC features in my text feature set because they provide a comprehensive range of linguistic indicators. These include structural features (pronouns like "I" and "we"), emotional markers ("anger," "sad," "posemo," "negemo"), and temporal references ("focuspast," "focuspresent," "focusfuture"), which together enable the model to capture structural, psychological, and social dimensions of dialogue.

My analysis of LIWC distributions across dialogue act tags in the training dataset revealed meaningful patterns. Informal language appeared frequently in backchannels ('b'), acknowledgments ('aa'), and yes-answers ('ny'), reflecting the casual nature of these conversational elements. Meanwhile, emotional features like "posemo" and "affect" were predominantly associated with appreciation ('ba') and closing statements ('fc'), which are dialogue acts that typically involve social and emotional engagement. These distinctive patterns suggest that LIWC features are particularly valuable for distinguishing between different dialogue acts.

I also incorporated ‘word_count’ and ‘character_count’ as features after discovering significant length variations across dialogue acts in the training data. Opinion statements ('sv') were notably longer, averaging 12 words and 60 characters, likely because they require more elaborate expression. By contrast, acknowledgments ('aa') averaged just 2 words (approximately 7 characters), while yes-answers ('ny') typically consisted of single words (around 5 characters). These consistent length patterns provide useful signals for dialogue act classification.

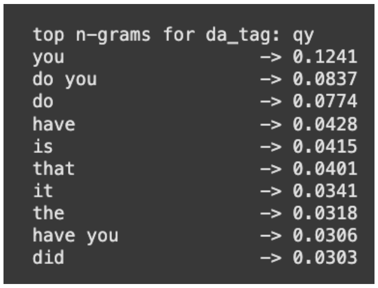




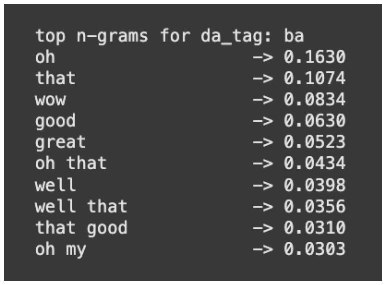
Punctuation features (‘has_question_mark’ and ‘has_exclamation’) were also included. Question marks appeared exclusively in yes-no questions ('qy'), while exclamation marks occurred primarily in appreciations ('ba') with rare instances in backchannels ('b') and declarative statements ('sd'). These punctuation patterns offer straightforward indicators for identifying specific dialogue acts.

For semantic representation, I employed the 'all-mpnet-base-v2' model from the *sentence_transformers* library to generate embeddings that capture contextual meaning. Since dialogue act classification fundamentally depends on understanding utterance intent and meaning, these transformer-based embeddings provide crucial semantic information that complements the surface level features.

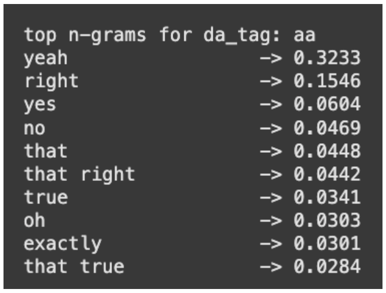
Finally, I applied scikit-learn's *TfidfVectorizer* with trigram capabilities (n-gram range of 1-3) and a 3000-feature limit. This choice was motivated by my analysis of characteristic phrases for each dialogue act. Yes-no questions ('qy') frequently featured interrogative patterns like "do you," "have you," and "did you" while appreciations ('ba') contained expressive interjections such as "oh," "wow," "good," and "great." The TF-IDF weighting scheme effectively emphasizes these class-specific n-grams while downweighting common filler words, making it particularly suited for capturing dialogue act signatures.



```
top n-grams for da_tag: qy
you                  -> 0.1241
do you               -> 0.0837
do                   -> 0.0774
have                 -> 0.0428
is                   -> 0.0415
that                 -> 0.0401
it                   -> 0.0341
the                  -> 0.0318
have you             -> 0.0306
did                  -> 0.0303
```



```
top n-grams for da_tag: ba
oh                   -> 0.1630
that                 -> 0.1074
wow                  -> 0.0834
good                 -> 0.0630
great                -> 0.0523
oh that              -> 0.0434
well                 -> 0.0398
well that            -> 0.0356
that good            -> 0.0310
oh my                -> 0.0303
```



```
top n-grams for da_tag: aa
yeah                 -> 0.3233
right                -> 0.1546
yes                  -> 0.0604
no                   -> 0.0469
that                 -> 0.0448
that right           -> 0.0442
true                 -> 0.0341
oh                   -> 0.0303
exactly              -> 0.0301
that true            -> 0.0284
```

Speech based Features:

Speech feature columns:
['Pitch_Min', 'Pitch_Max', 'Pitch_Mean', 'Pitch_SD', 'Intensity_Min', 'Intensity_Max', 'Intensity_Mean', 'Intensity_SD', 'Jitter', 'Shimmer', 'HNR', 'Speaking_Rate']

I incorporated various acoustic features extracted with *Parselmouth* into my speech-based feature set, including pitch statistics (minimum, maximum, mean, standard deviation), intensity measures (minimum, maximum, mean, standard deviation), jitter, shimmer, and harmonics-to-noise ratio (HNR). These features capture prosodic variations that characterize different dialogue acts.

To measure speaking rate, I implemented an algorithm inspired by de Jong & Wempe's (2009) syllable nuclei detection method adapted to work with pre-extracted acoustic features rather than raw audio. This approach estimates syllable counts by analyzing intensity variations in the speech signal, which reflect the natural energy fluctuations associated with syllable boundaries. For voiced segments, I estimated syllables using a combination of duration (assuming an average syllable duration of 225ms) and intensity variation, where higher standard deviation in intensity indicates more syllabic transitions. For unvoiced segments, I applied a more conservative estimation of one syllable per 300ms. After calculating the total syllable count, I computed speaking rate by dividing this count by utterance duration then applied bounds of 1.0 to 8.0 syllables per second to handle outliers.

These acoustic features provide distinct cues for dialogue act classification. Rising pitch contours typically characterize questions while flatter pitch patterns are associated with statements and closing remarks. Intensity variations capture emphasis and emotional engagement with more expressive and excited dialogue acts exhibiting higher intensity and pitch values. Jitter and shimmer measurements detect voice quality characteristics such as breathiness and tension while HNR distinguishes clearer, harmonic speech from noisier productions, potentially differentiating dialogue acts with contrasting interaction styles. Speaking rate may allow the detection of slower, more thoughtful statements or wrap-ups, and faster speaking rate might indicate excitement or assertiveness. Combined with the other acoustic features, these prosodic cues help in training the classifier model to recognize dialogue acts.

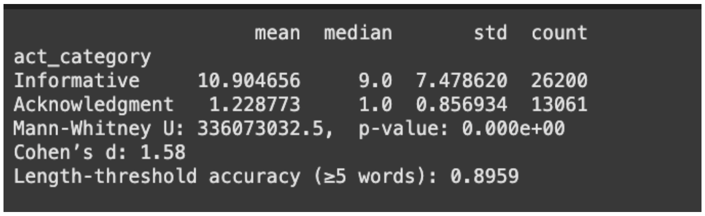
2. Feature Analysis:

a. For each custom feature set (text-feature set and speech feature set), formulate and test a hypothesis about the features (visually or statistically). Observe if the results are in accordance with your hypothesis or not. Give an explanation about your thinking behind the observed behavior.

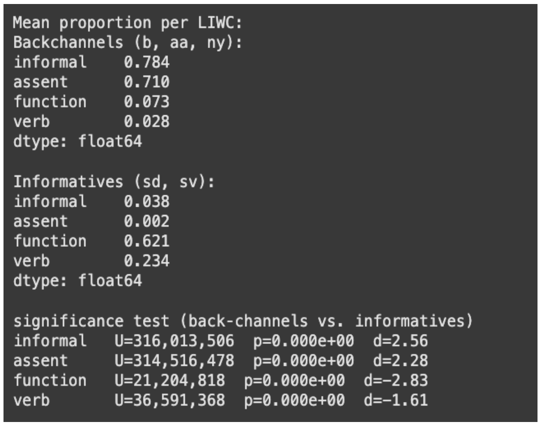
Text feature set:

I hypothesized that informative acts like 'sv' (statement-opinion) and 'sd' (statement-non-opinion) would contain more words than acknowledgment/backchannel acts like 'b' (backchannel) and 'aa' (agree/accept). This hypothesis was confirmed. Informative acts 'sv' and 'sd' averaged approximately 11 words while acknowledgment acts 'b' and 'ba' averaged only 1 word.

The Mann-Whitney U test (U = 336,073,025.5, p < 0.001, d = 1.58) confirmed that this difference is highly significant. When I applied a simple classification using a five-word threshold, the two categories were distinguished with 89.6% accuracy. This high accuracy reflects the fundamental difference between informative statements, which convey substantial content, and backchannels, which typically consist of brief utterances like "yeah."



```
                    mean  median       std  count
act_category
Informative     10.904656     9.0  7.478620  26200
Acknowledgment   1.228773     1.0  0.856934  13061
Mann-Whitney U: 336073032.5,  p-value: 0.000e+00
Cohen’s d: 1.58
Length-threshold accuracy (≥5 words): 0.8959
```



```
Mean proportion per LIWC:
Backchannels (b, aa, ny):
informal    0.784
assent      0.710
function    0.073
verb        0.028
dtype: float64

Informatives (sd, sv):
informal    0.038
assent      0.002
function    0.621
verb        0.234
dtype: float64

significance test (back-channels vs. informatives)
informal   U=316,013,506  p=0.000e+00  d=2.56
assent     U=314,516,478  p=0.000e+00  d=2.28
function   U=21,204,818  p=0.000e+00  d=-2.83
verb       U=36,591,368  p=0.000e+00  d=-1.61
```

I also hypothesized that backchannel acts 'b', 'aa', and 'ny' (yes answer) have more "informal" (mean ≈ 0.78) and "assent" (mean ≈ 0.71) tokens, while the informative acts ('sd', 'sv') would show almost none of these tokens (mean ≈ 0.04 and ≈ 0.00). Mann-Whitney tests confirm that the gap is highly significant for informal tokens (U = 316,013,506, p < 0.001, d ≈ 2.6) and assent tokens (U = 314,516,478, p < 0.001, d ≈ 2.3).

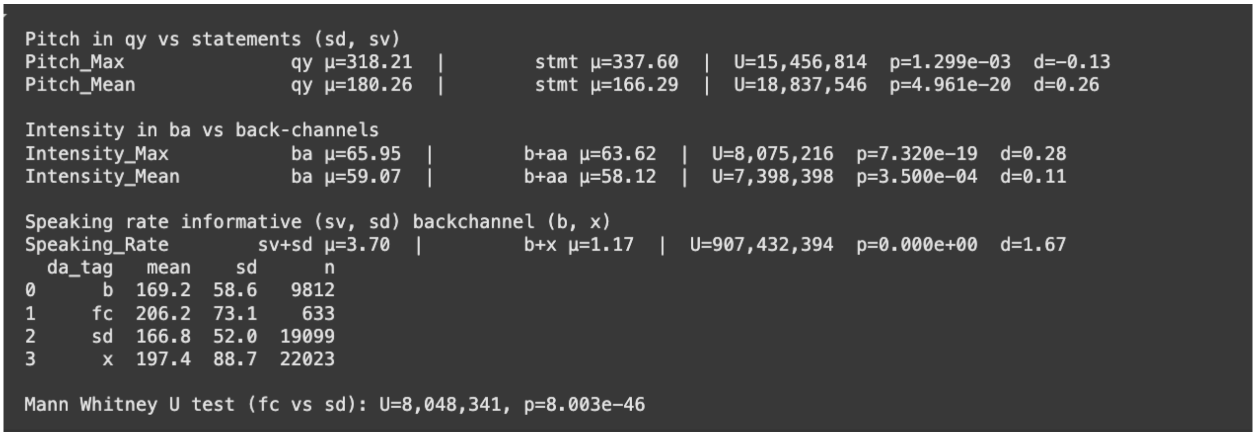
On the other hand, informative acts contain more function words and content verbs. Informative acts average 0.62 function words and 0.23 verbs, whereas backchannels average around 0.07 and 0.03. Mann-Whitney U tests show large effects in the opposite direction (function: U = 21,204,818, d ≈ –2.8; verb: U = 36,591,368, d ≈ –1.6; p < 0.001 in both cases). These results support that backchannels are mostly concise acknowledgments while informative acts are mainly fully articulated statements.

Speech feature set:

I hypothesized that yes-no questions ('qy') would have higher pitch than statement acts ('sd', 'sv'). The test results largely support this. Mean pitch is significantly higher in 'qy' (≈180 Hz) than in statements (≈166 Hz) (U = 18,837,546, p < 0.001, d ≈ 0.26), indicating that speakers raise their voices when asking questions. However, maximum pitch is slightly lower in 'qy' (318 Hz) than in statements (338 Hz), though the difference is small (U = 15,456,814, p ≈ 0.001, d ≈ –0.13). This suggests that questions are characterized by an overall pitch elevation rather than isolated pitch peaks.

I also hypothesized that appreciation acts ('ba') would be louder than ordinary backchannels ('b', 'aa'). Intensity tests confirm this. 'ba' peaks at approximately 66 dB versus 64 dB for 'b' and 'aa' (U = 8,075,216, p < 0.001, d ≈ 0.28). Average loudness follows the same pattern though with a smaller effect (U = 7,398,398, p ≈ 0.001, d ≈ 0.11), indicating that appreciation acts are moderately louder than ordinary backchannels.

I further hypothesized that content-heavy acts ('sv', 'sd') would be spoken more slowly than backchannel acts ('b', 'aa'). Speaking rate analysis strongly supports this. Informative statements ('sv', 'sd') average 3.7 syllables/second, while backchannels average 1.17 syllables/second (U = 907,432,394, p < 0.001, d ≈ 1.67). This confirms that content-rich acts require more time to articulate whereas brief backchannels like "yeah" are delivered rapidly.



```
Pitch in qy vs statements (sd, sv)
Pitch_Max              qy µ=318.21  |        stmt µ=337.60  |  U=15,456,814  p=1.299e-03  d=-0.13
Pitch_Mean             qy µ=180.26  |        stmt µ=166.29  |  U=18,837,546  p=4.961e-20  d=0.26

Intensity in ba vs back-channels
Intensity_Max          ba µ=65.95  |        b+aa µ=63.62  |  U=8,075,216  p=7.320e-19  d=0.28
Intensity_Mean         ba µ=59.07  |        b+aa µ=58.12  |  U=7,398,398  p=3.500e-04  d=0.11

Speaking rate informative (sv, sd) backchannel (b, x)
Speaking_Rate        sv+sd µ=3.70  |        b+x µ=1.17  |  U=907,432,394  p=0.000e+00  d=1.67
  da_tag   mean    sd      n
0      b  169.2  58.6   9812
1     fc  206.2  73.1    633
2     sd  166.8  52.0  19099
3      x  197.4  88.7  22023

Mann Whitney U test (fc vs sd): U=8,048,341, p=8.003e-46
```

3. Classification and Error Analysis:

Using the feature sets, train machine learning classifiers to identify/predict the 10 most frequent dialogue acts (Use the top 10 from section 3 here). Train 3 models: (1) speech features only (2) text features only (3) speech + text features. Describe the models used.

a-1) Description of the models used:

1) Speech features model: 1-dimensional CNN + 2-layer Transformer + SoftMax classifier

The speech classification model I trained uses a hybrid CNN-transformer architecture. The model begins with a one-dimensional convolutional neural network that processes speech features and extracts acoustic patterns then passes them to the transformer component. The transformer contains two encoder layers with multiple attention heads that capture contextual relationships between utterances using a sliding window of six consecutive utterances. The architecture incorporates learnable positional encodings to maintain sequence information and a classification head (layer normalization + linear layer) that outputs probabilities for the 10 dialogue act categories.

I performed hyperparameter tuning using *Optuna*, which identified the optimal configuration as: hidden dimension size = 128, number of attention heads = 4, and learning rate ≈ 0.00075. The model was trained for 25 epochs with early stopping based on validation performance, achieving a macro F1 score of 0.2729 and an accuracy of 0.6715 on the validation set.

2) Text features model: Stacked Ensemble model (LinearSVC + LightGBM classifier)

The text classification model I trained is an ensemble that combines *LinearSVC* and *LightGBM* classifiers through probability stacking. The *LinearSVC* model, which effectively identifies distinct word usage patterns, processes the TF-IDF features. I wrapped it with *CalibratedClassifierCV* to convert decision scores into probability estimates and configured it with balanced class weights to handle class imbalance.

The *LightGBM* model, which excels at capturing complex semantic relationships, processes the transcript sentence embeddings (768-dimensional vectors) along with LIWC features. It's configured with optimized hyperparameters including 400 estimators and a learning rate of 0.05.

The final classification is determined by averaging the probability outputs from both models with equal weights (50% *LinearSVC*, 50% *LightGBM*), selecting the class with the highest combined probability. This ensemble approach achieved a macro F1 score of 0.68 and an accuracy of 0.86 on the validation set.

3) Combined Speech + Text Model:

For the combined model, I used an ensemble approach where a meta-learner (Logistic Regression) combines predictions from the speech and text models. The ensemble includes a CNN-Transformer that processes acoustic features like pitch, intensity, and jitter; a *LinearSVC* model that handles TF-IDF text features; and a *LightGBM* model that processes sentence embeddings and linguistic indicators. The model was trained using cross-validation to generate out-of-fold predictions, ensuring the meta-learner learns from unbiased base model outputs.

b-1) Performance on validation set:

Model	Accuracy	F1
Speech	0. 6715	0.2729
Text	0.8564	0. 6803
Speech + Text	0.8559	0.6702

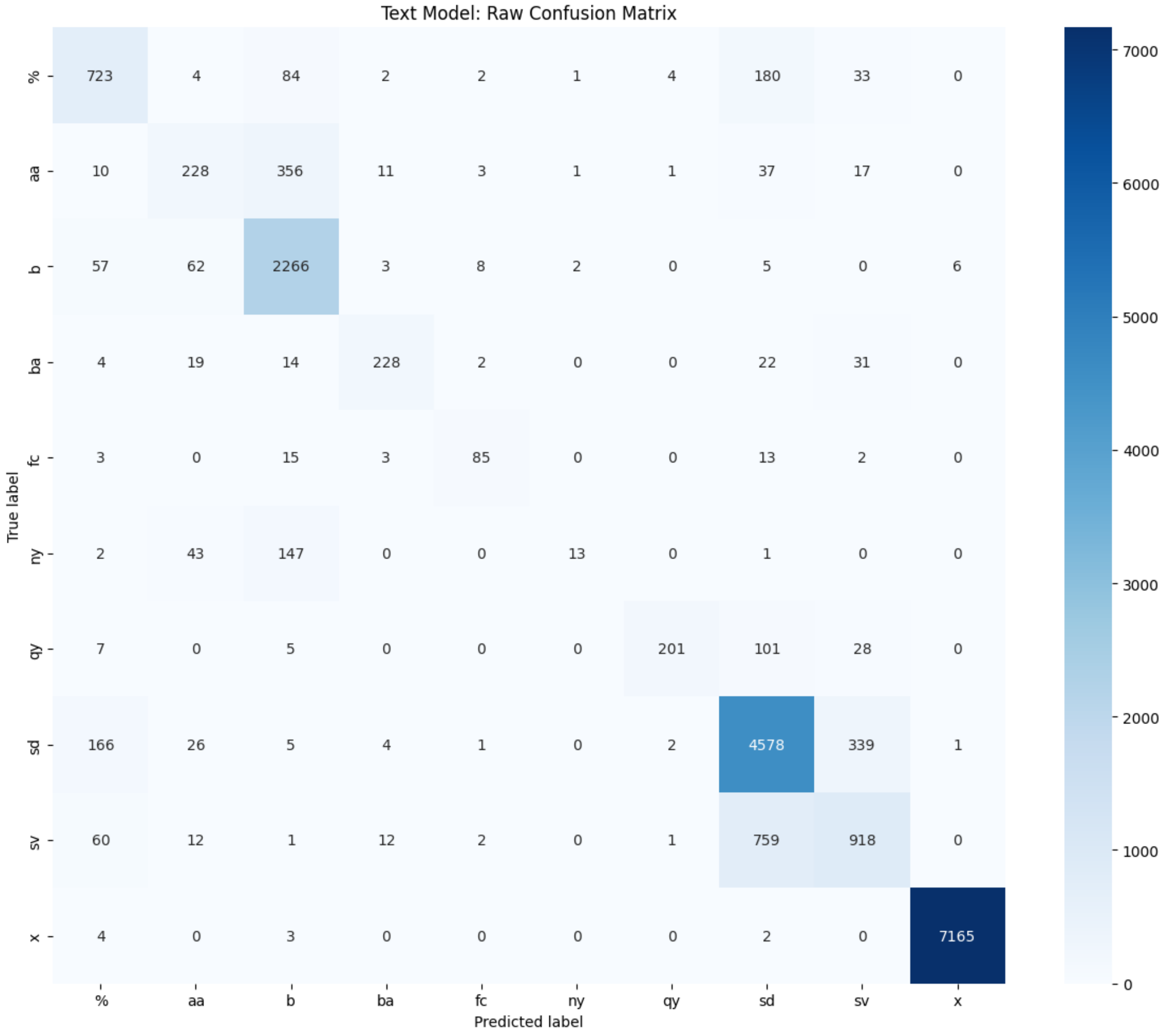
b-2) Best performing model:

The text-only model performed best. I believe this is due to the comprehensive linguistic information in the text feature set, including LIWC components, transcript embeddings, and TF-IDF features, which captured syntax, word choice, and sentence structure, enabling effective prediction of dialogue act tags. Additionally, the robust ensemble approach (combining LinearSVC and LightGBM) handled each feature type effectively.

The speech model performed comparatively worse for several reasons. First, the extracted speech features were less diverse than text features. Second, missing values in the speech features required imputation, potentially introducing noise. While the combined speech and text model provided complementary information, it may have introduced confusion when prosodic patterns didn't align with typical dialogue act patterns, essentially diluting the stronger text features with weaker speech features.

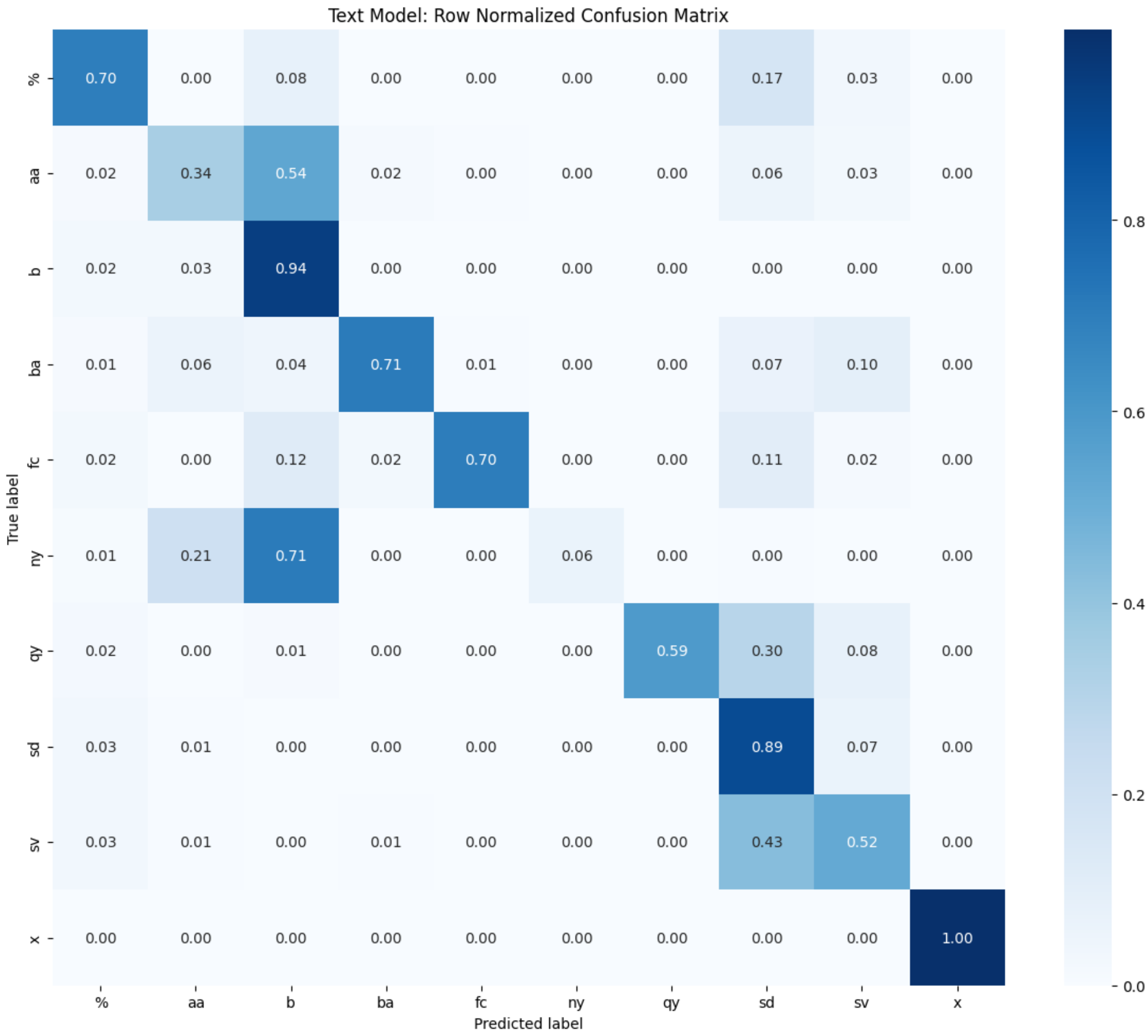
c) Error analysis (Text only model):

c-1) Confusion Matrix:



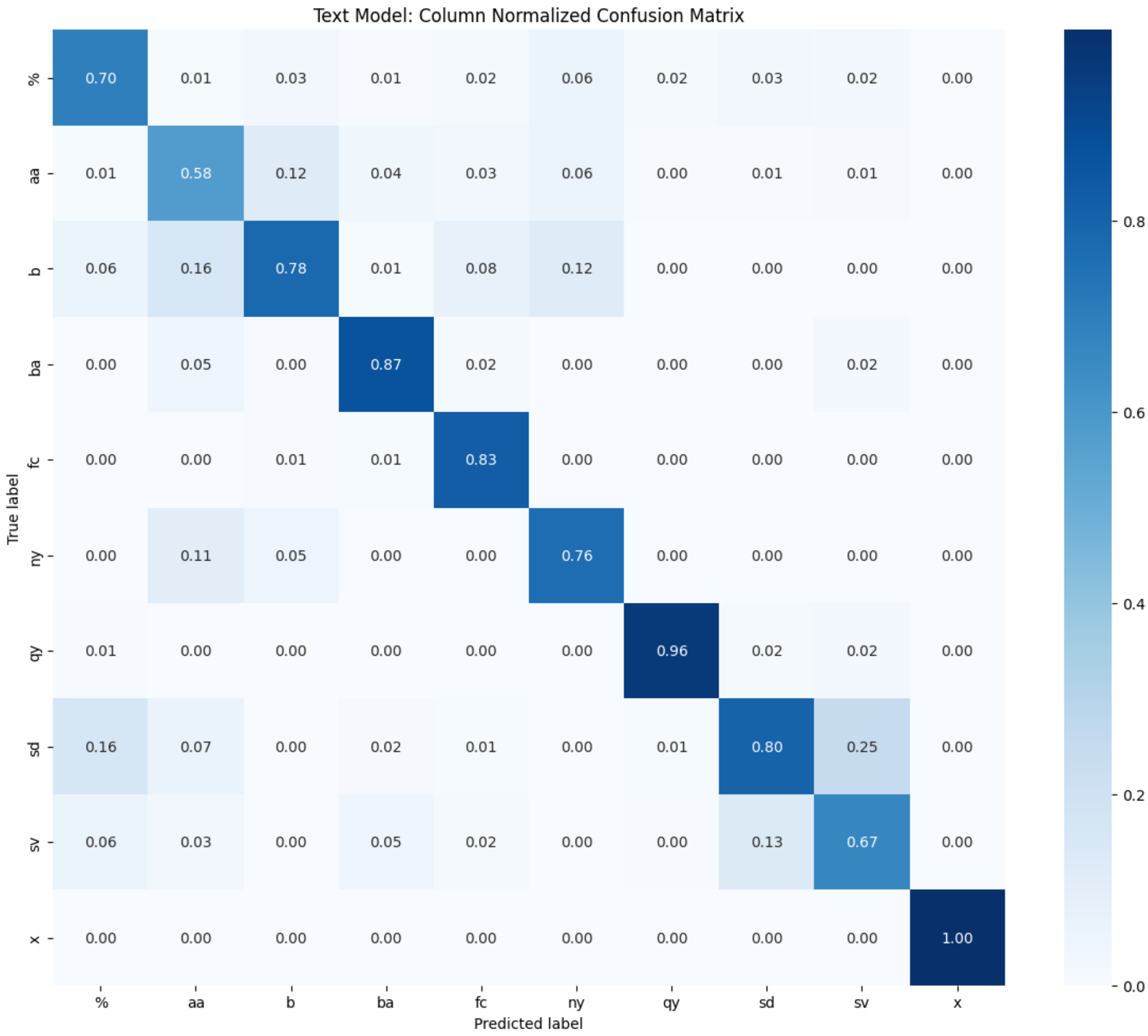
Text Model: Raw Confusion Matrix

True label \ Predicted label	%	aa	b	ba	fc	ny	qy	sd	sv	x
%	723	4	84	2	2	1	4	180	33	0
aa	10	228	356	11	3	1	1	37	17	0
b	57	62	2266	3	8	2	0	5	0	6
ba	4	19	14	228	2	0	0	22	31	0
fc	3	0	15	3	85	0	0	13	2	0
ny	2	43	147	0	0	13	0	1	0	0
qy	7	0	5	0	0	0	201	101	28	0
sd	166	26	5	4	1	0	2	4578	339	1
sv	60	12	1	12	2	0	1	759	918	0
x	4	0	3	0	0	0	0	2	0	7165



Text Model: Row Normalized Confusion Matrix

True label \ Predicted label	%	aa	b	ba	fc	ny	qy	sd	sv	x
%	0.70	0.00	0.08	0.00	0.00	0.00	0.00	0.17	0.03	0.00
aa	0.02	0.34	0.54	0.02	0.00	0.00	0.00	0.06	0.03	0.00
b	0.02	0.03	0.94	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ba	0.01	0.06	0.04	0.71	0.01	0.00	0.00	0.07	0.10	0.00
fc	0.02	0.00	0.12	0.02	0.70	0.00	0.00	0.11	0.02	0.00
ny	0.01	0.21	0.71	0.00	0.00	0.06	0.00	0.00	0.00	0.00
qy	0.02	0.00	0.01	0.00	0.00	0.00	0.59	0.30	0.08	0.00
sd	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.89	0.07	0.00
sv	0.03	0.01	0.00	0.01	0.00	0.00	0.00	0.43	0.52	0.00
x	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00



c-2) Easiest to predict classes:

The 'x' (non-verbal) act was easiest to predict, achieving perfect recall (1.00) and precision (1.00) for an F1 score of 1.00. This is likely because the 'x' act has distinctive linguistic features that clearly separate it from other dialogue acts, combined with the largest representation in the dataset (7,174 samples). The linguistic patterns for this class appear highly consistent and unique, making it easily identifiable.

The 'b' (backchannel) class also performed well with high recall (0.94) and good precision (0.78), yielding an F1 score of 0.85. This strong performance likely stems from both its substantial representation in the dataset (2,409 samples) and the distinctive linguistic patterns of backchannel responses, particularly short affirming phrases that the model could consistently recognize. Similarly, the 'sd' class achieved strong results with high recall (0.89) and good precision (0.80) for an F1 score of 0.85, likely due to its large sample size (5,122 instances) and the clear syntactic structures characteristic of statement declarations.

c-3) Most difficult to predict classes:

The 'ny' (yes-answers) class was the most challenging to predict, with extremely low recall (0.06) and moderate precision (0.76), resulting in the lowest F1 score (0.12). This poor performance can be attributed to its small representation (206 samples) and the fact that 71.36% of 'ny' instances were misclassified as 'b'. Yes-answers appear to lack distinctive textual features that would separate them from backchannel responses.

The 'aa' (agree/accept) class also proved challenging with low recall (0.34) and moderate precision (0.58), yielding an F1 score of 0.43. More than half (53.61%) of 'aa' instances were misclassified as 'b', suggesting that agreement responses share similar lexical features with backchannels.

The 'sv' (statement-opinion) class showed moderate performance with recall of 0.52 and precision of 0.67, producing an F1 score of 0.59. The primary source of error was confusion with 'sd' (statement-non-opinion), with 43% of 'sv' instances misclassified as 'sd'. This highlights how distinguishing between opinion and non-opinion statements requires deeper semantic understanding beyond surface-level features.

c-4) Easily confused classes:

The most confusion occurred between 'sv' (statement-opinion) and 'sd' (statement non-opinion), with 759 instances (43% of the 'sv' class) misclassified. This confusion stems from both being statements with similar syntactic structures, where the difference is in expressing an opinion versus stating a fact. As I mentioned above distinction between these two probably requires deeper semantic understanding and contextual awareness than the current model is able to.

Accept/Agree ('aa') and Backchannel ('b') classes also showed high instances of confusion, with 356 instances (53.61% of the 'aa' class) misclassified. Both typically use similar short affirmative responses and their difference depends on conversational context rather than lexical form. The distinction lies in intent (whether the user wants to express agreement or acknowledgment), which is difficult to distinguish without more contextual features.

'ny' and Backchannel 'b' classes were also frequently confused, with 147 instances (71.36% of the 'ny' class) misclassified. Both involve short responses with similar structures, which probably led the model to misclassify.

Yes-No-questions ('qy') were often misclassified as statement non opinion ('sd'), with 101 instances (29.53%), probably because questions can sometimes appear similar to statements without intonation markers from speech analysis.

c-5) Further improvements:

To improve the text classifier, feature engineering could target problematic classes by creating specialized features that capture subtle differences. For instance, distinguishing between statement-opinions and statement-non-opinions or developing features that differentiate accept/agree responses from backchannels based on their functional differences rather than lexical similarities.

Addressing class imbalance through techniques like class weighting or oversampling for underrepresented classes such as 'ny' would help improve performance on minority classes. Additionally, incorporating conversational context by including previous and subsequent utterances as features could provide valuable contextual information. Sequence models capable of capturing longer-range dependencies may also prove beneficial.

The text ensemble model could be enhanced through several strategies, such as implementing specialized binary classifiers for commonly confused pairs, adjusting ensemble weights to emphasize models that perform better on difficult classes, and incorporating additional linguistic features such as syntactic dependency analysis to better distinguish questions from statements. These targeted improvements would address the specific challenges identified in the current model's performance.