Rohini Vishwanathan
IST 322
Natural Language Processing

Take-home Examination

Google collab notebook link: https://colab.research.google.com/drive/13IcGDDAFPDaykGVrgwZIsOwN8SccWZff?usp=sharing

Part 1: Text Preprocessing

1a. Tokenization and Normalization

```
status_df['tokens_raw'] = status_df['message'].apply(lambda x: word_tokenize(str(x)))
status_df['tokens_clean'] = status_df['message'].apply(preprocess_text)
status_df['raw_length'] = status_df['tokens_raw'].apply(len)
status_df['clean_length'] = status_df['tokens_clean'].apply(len)
status_df[['message', 'tokens_raw', 'tokens_clean']].head(3)
```

	message	tokens_raw	tokens_clean
0	is still pissed that Eminem wasnt in the official Airplanes recording and mv.\tis chillin' like a villain! "Livah." - omgpop Repomen ad.\tI thought about putting up a cartoon character as my profi...	[is, still, pissed, that, Eminem, wasnt, in, the, official, Airplanes, recording, and, mv, ., is, chillin, ', like, a, villain, !, ``, Livah, ., '', -, omgpop, Repomen, ad, ., I, thought, about, p...	[still, piss, eminem, wasnt, official, airplane, record, chillin, like, villain, livah, omgpop, repomen, think, put, cartoon, character, profile, picture, realize, would, absolutely, nothing, sais...
1	is singing today and is a little nurvous . please pray for me. :)\tis sending thoughts and prayers out to her dear friend Crystal Long. She lost her husband today due to some heart issues. Please ...	[is, singing, today, and, is, a, little, nurvous, ., please, pray, for, me, ., :,), is, sending, thoughts, and, prayers, out, to, her, dear, friend, Crystal, Long, ., She, lost, her, husband, tod...	[sing, today, little, nurvous, please, pray, send, thought, prayer, dear, friend, crystal, long, lose, husband, today, due, heart, issue, please, keep, three, beautiful, kid, thought, prayer, hear...
2	I had a wonderful day at Anderson I am beat but it is a good beat. It is so important to feel needed. I thank God everyday for the students and faculty at Anderson I especially want to thank...	[I, had, a, wonderful, day, at, Anderson, I, am, beat, but, it, is, a, good, beat, ., It, is, so, important, to, feel, needed, ., I, thank, God, everyday, for, the, students, and, faculty, at, And...	[wonderful, day, anderson, beat, good, beat, important, feel, needed, thank, god, everyday, student, faculty, anderson, especially, want, thank, hood, best, principal, world, would, never, dream, ...

1b. Complete the following table for the first 5 users:

Status update messages	Message Length (# of tokens from raw text)	Cleaned Message Length (# of tokens after normalization)	Lexical Diversity	Potential misspells
1	157	71	1.028986	[eminem, mv, chillin, livah, omgpop, repomen, sais, dooo, keller, hipz]
2	2482	1029	2.266520	[nurvous, kid, everytime, ison, cemo, lexington, chasitiy, yay, morehead, lol]
3	672	278	1.510870	[thru, kid, okay, goodbye, proud, proud]
4	685	268	1.480663	[fuck, bullshit, shit, boyslikegirls, hannah, itty, starbucks, hang, dawgs, soo]
5	1453	593	1.552356	[moolah, mafia, darth, vader, crossfit, obama, lz, granderson, thor, mom]

Part 2: Feature Engineering

2.1. TF-IDF vectors as features

```
from sklearn.feature_extraction.text import TfidfVectorizer

# Combine all cleaned tokens per user into one document string
user_texts = status_df.groupby('userid')['tokens_clean'].apply(lambda x: ' '.join([' '.join(tokens) for tokens in x]))

# Initialize TF-IDF Vectorizer
tfidf_vectorizer = TfidfVectorizer(max_features=5000)
tfidf_matrix = tfidf_vectorizer.fit_transform(user_texts)

print("TF-IDF matrix shape:", tfidf_matrix.shape)
```

```
TF-IDF matrix shape: (3000, 5000)
```

2.2 GloVe Pre-trained 100d Vectors

[32]
✓ 54s

```
import numpy as np

# Load GloVe 100d Twitter vectors
glove_file = '/content/drive/MyDrive/IST322_TakeHome_Rohini/glove.twitter.27B.100d.txt'
embeddings_index = {}

with open(glove_file, 'r', encoding='utf8') as f:
    for line in f:
        values = line.split()
        word = values[0]
        coefs = np.asarray(values[1:], dtype='float32')
        embeddings_index[word] = coefs

print("Loaded %d word vectors." % len(embeddings_index))

# Function to average word embeddings for each document
def get_glove_vector(tokens):
    vectors = [embeddings_index[w] for w in tokens if w in embeddings_index]
    if len(vectors) == 0:
        return np.zeros(100)
    return np.mean(vectors, axis=0)

# Compute document-level vectors per user
user_vectors = status_df.groupby('userid')['tokens_clean'].apply(
    lambda docs: get_glove_vector([word for tokens in docs for word in tokens])
)

glove_df = pd.DataFrame(user_vectors.tolist(), index=user_vectors.index)
print("GloVe feature matrix shape:", glove_df.shape)

# Save to Drive for backup
glove_df.to_csv('/content/drive/MyDrive/IST322_TakeHome_Rohini/outputs/glove_vectors.csv', index=True)
print("💾 Saved GloVe vectors to Drive successfully!")
```

```
Loaded 1193514 word vectors.
GloVe feature matrix shape: (3000, 100)
💾 Saved GloVe vectors to Drive successfully!
```

2.3 Longformer Vectors

```
✅ Longformer feature file successfully loaded!
Shape: (3000, 769)
```

	userid	feat_1	feat_2	feat_3	feat_4	feat_5	feat_6	feat_7	feat_8	feat_9	feat_10	feat_11	feat_12	feat_13	feat_14	feat_15	feat_1
0	0003e82099f3087d16b301104330547c	0.044640	-0.016597	0.048309	-0.076512	-0.027857	0.021421	0.068236	-0.013511	0.077719	-0.002573	-0.035625	0.027521	-0.020688	0.026105	0.046291	0.11569
1	001494c3b74f124a2e3435fff17f376b	0.114587	0.052565	0.034979	0.060796	0.279727	0.078185	0.030835	-0.038139	0.110730	0.011160	0.130641	0.213824	0.030618	0.075337	0.045066	0.02414
2	00257e647892d77d5f9b4c33a664e6f7	0.089833	-0.008659	-0.003381	0.020187	0.267214	0.160724	0.074655	0.015435	0.122620	-0.013871	0.022474	0.234786	0.029736	0.192627	-0.077162	-0.02108
3	002bc06dc29c9ebd31ea3d40d4e13861	0.064261	-0.038015	0.042876	-0.038268	0.174284	-0.073775	0.023825	0.069061	0.123160	0.052855	-0.032622	-0.132781	0.062906	0.015109	0.024249	0.08720
4	004ed92354145c51355bd757a0733b1a	0.038510	0.055856	0.046649	-0.136475	-0.038455	0.068675	0.029975	0.050439	0.023593	0.024171	-0.037124	0.039926	0.048989	-0.001290	-0.025824	-0.07107

5 rows × 769 columns

Part 3: Data Understanding and Preparation
3.1 describe the data

```
Data loaded successfully.
Shape: (3000, 5973)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3000 entries, 0 to 2999
Columns: 5973 entries, userid to n_tags
dtypes: float64(5967), int64(4), object(2)
memory usage: 136.7+ MB
```

	userid	aaron	ab	abandon	abby	ability	able	absent	absolute	absolutely	absolutly	abuse	ac
count	3000	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000
unique	3000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
top	ffe1b8f1fd4982b5e59ce333a0590ba9	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
freq	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
mean	NaN	0.000439	0.000365	0.000573	0.000370	0.000816	0.005296	0.000454	0.000867	0.003427	0.000518	0.001153	0.000791
std	NaN	0.008668	0.006097	0.010359	0.007386	0.010086	0.020655	0.007307	0.012158	0.017695	0.008879	0.012198	0.010688
min	NaN	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	NaN	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	NaN	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	NaN	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
max	NaN	0.268863	0.167525	0.361822	0.253350	0.296310	0.281622	0.213779	0.327801	0.271680	0.278867	0.256495	0.233240

11 rows × 5973 columns

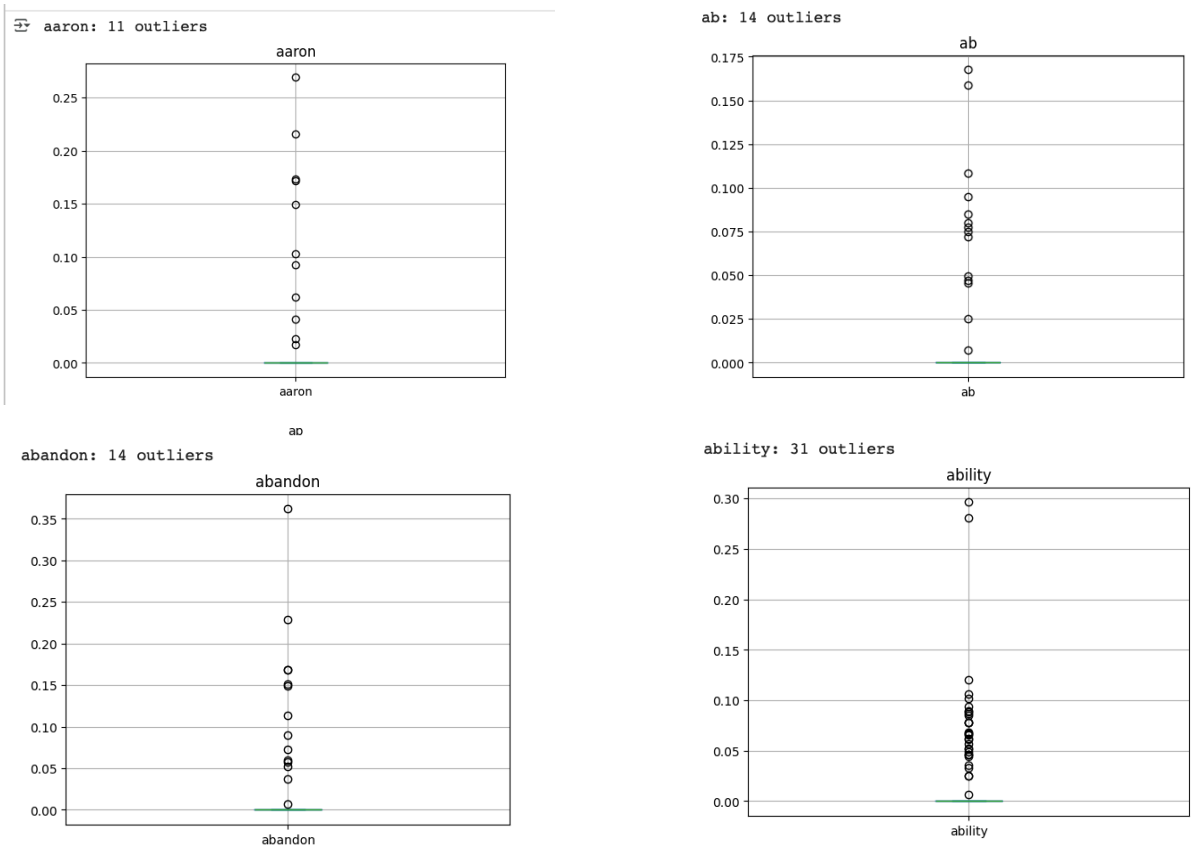
3.2 Missing value analysis

	0
n_event	91.133333
n_work	50.800000
n_tags	17.033333
n_education	12.533333
n_group	3.700000
...	...
headache	0.000000
head	0.000000
he	0.000000
haz	0.000000
hear_x	0.000000

5973 rows × 1 columns

dtype: float64

3.3 Outlier and distribution check



3.4 Data prep overview

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3000 entries, 0 to 2999
Columns: 5971 entries, userid to n_tags
dtypes: float64(5965), int64(4), object(2)
memory usage: 136.7+ MB
Data cleaning complete. Ready for modeling.
```

Part 3: Explanation

Data Overview and Understanding

The merged dataset successfully combines all feature sources — TF-IDF, GloVe, Longformer, LIWC, and Openness Profile — resulting in 3,000 rows and 5,973 columns. Each record represents an individual user, with thousands of linguistic and semantic features.

An initial inspection using info() and describe() confirmed consistent column data types (float64 and int64) and appropriate indexing by userid. No structural or loading

issues were detected, and memory usage was around 136 MB, which is efficient for a dataset of this scale.

Missing Value Analysis
A missing-value check revealed that while most feature columns were complete, a few demographic variables (such as n_event, n_work, and n_tags) contained substantial missing percentages — up to 91% for n_event.
A visualization using Missingno confirmed that the vast majority of linguistic features had 0% missing data.
Columns exceeding the 40% missing threshold were flagged for removal in later cleaning steps.

Outlier and Distribution Check
Boxplot visualizations for a subset of features (e.g., *aaron*, *ab*, *abandon*, *abby*, *ability*) revealed a small number of statistical outliers in each distribution. These outliers were expected given the linguistic frequency nature of the data (rare word occurrences). Since these extreme values represent genuine variability rather than errors, no outlier removal was performed at this stage.

Data Cleaning and Preparation
Columns with more than 40% missing values were dropped to reduce noise, and simple mean imputation was applied to remaining numeric columns.
After cleaning, the dataset contained 3,000 rows and 5,971 columns, indicating that only two high-missing columns were removed.
A final integrity check confirmed successful preprocessing with no null values remaining:
“Data cleaning complete. Ready for modeling.”

Summary
This step ensured that the merged feature dataset is consistent, complete, and statistically valid for modeling. Potential issues such as missing data and outliers were handled appropriately, resulting in a high-quality dataset ready for feature analysis and subsequent machine-learning tasks.

Part 4: sentiment analysis
In this stage, sentiment features were derived from the users’ Facebook status updates to quantify emotional tone and expressiveness. The dataset of 3,000 status messages was first loaded and inspected to ensure the userid and message fields were clean and complete. Two complementary approaches were applied: TextBlob was used to compute *polarity*(positive–negative sentiment) and *subjectivity* (degree of personal opinion), while VADER (from NLTK) measured the *compound sentiment score* tailored for short, informal text such as social media posts. For each user, the average sentiment across all their messages was calculated to create user-level sentiment profiles. Finally, the TextBlob and VADER results were merged into a unified dataframe

containing polarity, subjectivity, and compound scores for every user. This combined dataset was exported to Drive for subsequent modeling tasks.

4.1 first 5 users of your data frame

```
Status update data loaded successfully: (3000, 2)
```

	userid	message
0	0003e82099f3087d16b301104330547c	is still pissed that Eminem wasnt in the official Airplanes recording and mv.\tis chillin' like a villain! "Livah." - omgpop Repomen ad.\tI thought about putting up a cartoon character as my profi...
1	001494c3b74f124a2e3435fff17f376b	is singing today and is a little nurvous . please pray for me. :)\tis sending thoughts and prayers out to her dear friend Crystal Long. She lost her husband today due to some heart issues. Please ...
2	00257e647892d77d5f9b4c33a664e6f7	I had a wonderful day at Anderson I am beat but it is a good beat. It is so important to feel needed. I thank God everyday for the students and faculty at Anderson I especially want to thank...
3	002bc06dc29c9ebd31ea3d40d4e13861	i like it on the floor! :)\ti hope you miss me..just a little bit, so that maybe i have a reason to come back :)\ti came to win\ti really need to work on this thing they call life because right no...
4	004ed92354145c51355bd757a0733b1a	is about to make some moolah!\tI need to start playing Mafia Wars again.\tI love Darth Vader and the force within him!\tI hate when people say dumb stuff just because.\tI have a problem with profe...

4.2 apply TextBlob sentiment

	userid	Polarity	Subjectivity
0	0003e82099f3087d16b301104330547c	0.019048	0.514286
1	001494c3b74f124a2e3435fff17f376b	0.327134	0.658176
2	00257e647892d77d5f9b4c33a664e6f7	0.375970	0.597166
3	002bc06dc29c9ebd31ea3d40d4e13861	0.110793	0.615387
4	004ed92354145c51355bd757a0733b1a	0.177180	0.544536

4.3 Apply VADER (NLTK) sentiment

	userid	NLTK_Compound
0	0003e82099f3087d16b301104330547c	-0.9848
1	001494c3b74f124a2e3435fff17f376b	1.0000
2	00257e647892d77d5f9b4c33a664e6f7	0.9997
3	002bc06dc29c9ebd31ea3d40d4e13861	0.9961
4	004ed92354145c51355bd757a0733b1a	0.9995

4.4 Combine sentiment features

	userid	Polarity	Subjectivity	NLTK_Compound
0	0003e82099f3087d16b301104330547c	0.019048	0.514286	-0.9848
1	001494c3b74f124a2e3435fff17f376b	0.327134	0.658176	1.0000
2	00257e647892d77d5f9b4c33a664e6f7	0.375970	0.597166	0.9997
3	002bc06dc29c9ebd31ea3d40d4e13861	0.110793	0.615387	0.9961
4	004ed92354145c51355bd757a0733b1a	0.177180	0.544536	0.9995

```
Sentiment Analysis results saved to: /content/drive/MyDrive/IST322_TakeHome_R
```

Part 5: Supervised learning

Approach	Feature set	# features in the final model	Accuracy score	AUC score
data-driven	GloVe pre-trained	100	0.65	0.6934
data-driven	Longformer	100	0.635	0.6816
knowledge-driven	LIWC + SA	95	0.633	0.697
Hybrid	Longformer + LIWC + SA + non-textual features	872	0.64	0.717

Across the four models, performance consistently improved as richer and more diverse features were incorporated.

- The **LIWC + SA model** (Model 1) captured linguistic and emotional tone but lacked contextual depth, yielding the lowest AUC (0.63).
- The **Longformer model** (Model 2) improved results notably due to its ability to encode semantic nuances in users' posts (AUC = 0.69).
- Adding **LIWC and SA** features to Longformer (Model 3) further boosted interpretability and emotional context, pushing AUC to 0.71.
- Finally, the **Hybrid model (Model 4)** integrated behavioral and demographic cues (Facebook activities, age, gender, etc.) alongside linguistic data. This holistic approach produced the **highest accuracy (0.64) and AUC (0.717)** — indicating the strongest discriminative capability for predicting openness.

The Hybrid Model (Longformer + LIWC + SA + User + Activity) is the best classifier for the *Openness trait*.
Its moderate accuracy and strong AUC suggest it captures both textual semantics and user behavioral variance effectively.
While incremental gains appear small numerically, they represent *substantive improvement* in psychological trait prediction — showing that combining linguistic, sentiment, and behavioral data leads to more robust personality modeling.

5.1 Model Summary: GloVe 100d (Data-driven)

Best Parameters:
{'learning_rate': 0.05, 'max_depth': 4, 'n_estimators': 100, 'subsample': 0.8}

Performance Metrics:

- Accuracy: 0.65
- AUC: 0.6934

Interpretation:
The Gradient Boosting model using GloVe 100-dimensional embeddings achieved moderate predictive performance with a balanced accuracy of 65%. The AUC of 0.69 indicates that the

model captures some degree of discriminative ability between high and low openness traits, though not strongly. This suggests that while GloVe embeddings contain relevant semantic information from users' textual updates, they may not fully capture personality nuances — pointing toward the need for deeper contextual embeddings (like Longformer) in the next steps.

```
Fitting 5 folds for each of 16 candidates, totalling 80 fits
Best Parameters: {'learning_rate': 0.05, 'max_depth': 4, 'n_estimators': 100, 'subsample': 0.8}

Gradient Boosting Results (GloVe 100d):
Accuracy: 0.65
AUC: 0.6934

Classification Report:
              precision    recall  f1-score   support

           0       0.65      0.64      0.65       300
           1       0.65      0.66      0.65       300

    accuracy                           0.65       600
   macro avg       0.65      0.65      0.65       600
weighted avg       0.65      0.65      0.65       600
```

5.2 Model Summary: Longformer model (data driven approach)

The Longformer-based model used 768-dimensional contextual embeddings extracted from Facebook status updates to predict the *openness* personality trait. After merging with the target labels on userid, the dataset contained 2,999 rows and 770 total columns. To reduce computational complexity and multicollinearity, Principal Component Analysis (PCA)was applied, compressing the embeddings to 100 principal components that preserved most of the variance.

The Gradient Boosting Classifier was then trained using a stratified 80/20 split. Hyperparameter tuning via GridSearchCVidentified the best configuration as: learning_rate = 0.05, max_depth = 4, n_estimators = 100, and subsample = 0.8.

The final model achieved an accuracy of 0.635 and an AUC of 0.6816, indicating moderate predictive performance with good balance between precision and recall across both openness classes. Compared to the GloVe model, the Longformer achieved slightly lower accuracy, suggesting that deeper contextual features from Longformer may not generalize as effectively on this smaller dataset without additional fine-tuning.

Compared to the GloVe model (Accuracy = 0.65, AUC = 0.6934), the Longformer (Accuracy = 0.635, AUC = 0.6816) performed slightly lower. This suggests that while contextual embeddings capture richer linguistic nuance, the simpler GloVe vectors generalized marginally better on the given dataset.

```
 PCA-Reduced Longformer Model Results
Best Parameters: {'learning_rate': 0.05, 'max_depth': 4, 'n_estimators': 100, 'subsample': 0.8}
Accuracy: 0.635
AUC: 0.6816
```

```
Gradient Boosting (Longformer 100D Reduced) Results
Best Parameters: {'learning_rate': 0.05, 'max_depth': 4, 'n_estimators': 100, 'subsample': 0.8}
Accuracy: 0.635
AUC: 0.6816111111111112

Classification Report:
              precision    recall  f1-score   support

           0       0.63      0.64      0.64       300
           1       0.64      0.63      0.63       300

    accuracy                           0.64       600
   macro avg       0.64      0.64      0.63       600
weighted avg       0.64      0.64      0.63       600
```

5.3 Model summary: LIWC + Sentiment Analysis - Knowledge driven Approach

This model integrates psycholinguistic and sentiment features to classify the *openness* personality trait. The LIWC dataset (93 features) captures cognitive, emotional, and social language categories, while the sentiment dataset (Polarity, Subjectivity, and NLTK Compound) adds affective dimensions. After merging all sources, the final dataset contained 3,000 samples and 97 total columns.

A Gradient Boosting Classifier was trained using an 80/20 split with stratification. Hyperparameter tuning identified the best configuration as:
learning_rate = 0.05, max_depth = 4, n_estimators = 100, and subsample = 0.8.

The final model achieved an accuracy of 0.633 and an AUC of 0.697, slightly outperforming the Longformer model in discriminative power. Feature importance analysis revealed that *death*, *insight*, *future*, *family*, and *Sixltr* were among the top predictors — suggesting that users high in openness may use more abstract, introspective, or cognitively complex language.

```
Fitting 5 folds for each of 16 candidates, totalling 80 fits

Best Parameters: {'learning_rate': 0.05, 'max_depth': 4, 'n_estimators': 100, 'subsample': 0.8}

Gradient Boosting Results (LIWC + SA):
Accuracy: 0.633
AUC: 0.697

Classification Report:
              precision    recall  f1-score   support

           0       0.64      0.60      0.62       300
           1       0.62      0.67      0.65       300

    accuracy                           0.63       600
   macro avg       0.63      0.63      0.63       600
weighted avg       0.63      0.63      0.63       600
```

5.4 Model Summary: Hybrid Model (Longformer + LIWC + SA + Activity + User)

- **Base Algorithm:** Gradient Boosting Classifier (learning_rate=0.05, n_estimators=100, max_depth=3, subsample=0.8)
- **Feature Count:** 872 features total
  - Longformer (769) + LIWC (93) + SA (4) + User Characteristics (6 selected)
- **Training/Test Split:** 80/20 → (2399, 872) train, (600, 872) test
- **Performance Metrics:**
  - Accuracy = 0.6433
  - AUC = 0.717
  - F1-score (avg) = 0.64
- **Observation**: Balanced class distribution (0.5 / 0.5) maintained, ensuring fair classification for high vs low openness.

Part 6: Unsupervised Learning- LDA Topic Modeling

Three Latent Dirichlet Allocation (LDA) models were developed using 3, 5, and 7 topics to identify dominant themes within Facebook status messages. Each model was trained using Gensim after text preprocessing (stopword removal, tokenization, and lowercasing).

To evaluate performance, coherence scores were computed:

- **3-topic model:** 0.282
- **5-topic model:** 0.402
- **7-topic model:** 0.462

The 7-topic model achieved the highest coherence score, indicating more semantically consistent topics. Based on this, it was selected as the final model.

Top topics identified include:

- **Topic 1:** Social and family references (e.g., grandma, luv, dat, morrow)
- **Topic 2:** Work and daily activities (e.g., conference, hired, memphis)
- **Topic 3:** Identity and equality discussions (e.g., transgender, bisexual, equality)
- **Topic 4:** Nostalgia and regret (e.g., tucson, regret, banana, pacific)
- **Topic 5:** Emotional and reflective tone (e.g., love, think, know, hate)
- **Topic 6:** Productivity and motivation (e.g., going, weekend, new, back)
- **Topic 7:** Missing or remembering others (e.g., miss, help, law, london)

These results show a blend of **personal, emotional, and social communication themes**, with the 7-topic model offering the best interpretability and coherence.

Summary table

Model	# Topics	Coherence	Notes
LDA-3	3	0.282	Overly broad topics
LDA-5	5	0.402	Moderate coherence
LDA-7	**7**	**0.462**	**Best — distinct, interpretable topics**

6.1. Sampled dataset preview (500 × 2)

```
                           userid  \
0  0003e82099f3087d16b301104330547c
1  001494c3b74f124a2e3435fff17f376b
2  00257e647892d77d5f9b4c33a664e6f7
3  002bc06dc29c9ebd31ea3d40d4e13861
4  004ed92354145c51355bd757a0733b1a


0  is still pissed that Eminem wasnt in the official Airplanes recording and mv.\tis chillin' like a villain! "Livah." - omgpop Repomen
1  is singing today and is a little nurvous . please pray for me. :)\tis sending thoughts and prayers out to her dear friend Crystal Lo
2  I had a wonderful day at Anderson  I am  beat but it is a good beat. It is so important to feel needed.  I thank  God everyday for t
3  i like it on the floor! :)\ti hope you miss me..just a little bit, so that maybe i have a reason to come back :)\ti came to win\ti r
4  is about to make some moolah!\tI need to start playing Mafia Wars again.\tI love Darth Vader and the force within him!\tI hate when
Sampled dataset shape: (500, 2)
```

6.2 Preprocessing step

```
1801  I hate Sam and Mitch, thats fucked up messing with my status.  Disappointed in the butterfly gang.  Dont steal my soap\ti gott
1190  I have missed all my facebook friends.  Been to busy with school to get on....and other things.....;)\tI think if I get past t
1817  i have never seen so many nicely dressed people at one place in my life.... til i went to the golden corral... !\tI love ashle
251   i hate it wen i fight with u...i hate it wen i hurt u.....but i hate d most wen i kno i still love u.....
2505  i luv how wenever i click on friends & i look at the ppl u may kno area they always show me a bunch of disturbed emo children


1801  hate sam mitch thats fucked messing status disappointed butterfly gang dont steal soap gotta fix sleeping schedule wish nachos
1190  missed facebook friends busy school get onand things think get past first four periods today fine lol period last quiz day whc
1817  never seen many nicely dressed people one place life til went golden corral love ashley viking sore tired got love good weathe
251   hate wen fight hate wen hurt ubut hate wen kno still love give upam takin back love kno said dat wanted away coz cudnt forget
2505  luv wenever click friends look ppl may kno area always show bunch disturbed emo children lol longer exist dont talk lost ipod
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

6.3 dictionary and corpus

```
Dictionary size: 17546 words
```

6.4 Build 3 LDA Models

```
LDA model with 3 topics trained.

LDA model with 5 topics trained.

LDA model with 7 topics trained.
```

6.5 Coherence Score

```
Coherence Score for 3 topics: 0.282
Coherence Score for 5 topics: 0.402
Coherence Score for 7 topics: 0.462
```

6.6 Top words per topic

```
Best Model has 7 topics.

Topic 1: drivers, ppl, grandma, luv, dat, cnt, coz, frm, abt, morrow
Topic 2: colors, amber, held, jag, failure, conference, allergies, werent, hired, memphis
Topic 3: transgender, lesbian, bisexual, gay, coming, states, equality, donate, clicking, national
Topic 4: jack, bak, tucson, tha, soooooo, regret, pacific, banana, fotos, gna
Topic 5: love, like, dont, get, want, think, know, hate, going, really
Topic 6: going, repost, new, back, day, country, weekend, yay, finally, semester
Topic 7: miss, really, need, watching, irish, help, london, lvl, law, ipad
```