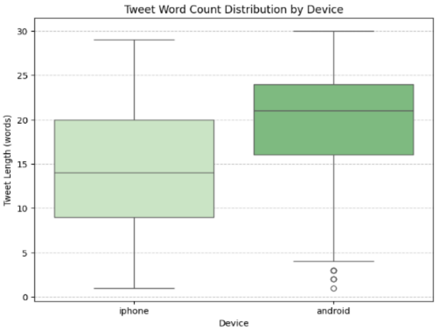
ב"ה

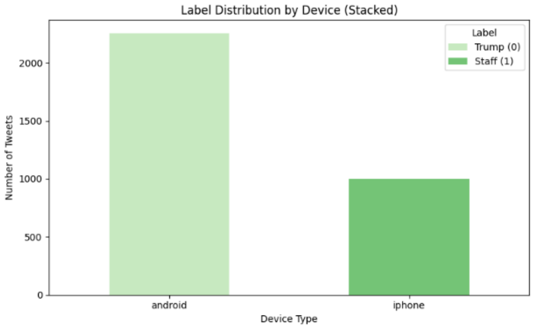
Ayelet Hashahar Cohen 206533895
Tzuf Lahan 208512038

NLP Assignment 3-Text Classification and Authorship Attribution

1. Data

We used the provided tweets.tsv dataset, containing thousands of tweets from Trump and his staff, labeled indirectly through device metadata (Android vs. iPhone). Only tweets from the handle realDonaldTrump were included, and we treated the device field as the binary label (0 for Android, 1 for iPhone).


Tweet Word Count Distribution by Device


Label Distribution by Device (Stacked)

2. Preprocessing steps:

The preprocessing sub-process filters the raw tweet data for the relevant user and device types, then creates a binary target label (NotTrump) indicating iPhone vs. Android. Unused columns are dropped and the classes are balanced via under sampling to ensure equal representation. Finally, the cleaned and engineered features are separated into the feature matrix X and target vector **y**, ready for input into any machine learning model.

Feature Extraction:

To capture stylistic and behavioral cues for device classification, we engineered a diverse set of features grouped into five categories:

- Time-Based Features
  - **Hour of Day, Weekday, Month, Year** – Reflect habitual tweeting times (e.g. early-morning bursts vs. daytime staff posts).
- Raw Text Counts
  - **Word Count & Character-Per-Word** – Trump's tweets tend to be shorter and punchier.
  - **Capital-Letters Ratio** – High usage of ALL-CAPS for emphasis (e.g. "GREAT").
  - **Hashtag, URL & Retweet Counts** – Quantify engagement markers and reuse of links or tags.
  - **Exclamation & Pronoun Counts** – Emphasis and personal style signals.
- Sentiment Features
  - **Polarity (–1 to +1) & Subjectivity (0 to 1)** – Overall tone measured via TextBlob.

  - **Counts of Positive vs. Negative Words** – Capture emotional bias in language.
- **Emotion Features**
  - **Anger, Disgust, Fear, Joy, Neutral, Sadness, Surprise Scores** – Seven-dimensional emotion profile from a transformer pipeline, highlighting nuanced affect.
- **Contextual & Statistical Features**
  - **Text Length & Punctuation Density** – Overall verbosity and intensity of punctuation use.
  - **Presence Flags: URL, Hashtag, Mention** – Binary indicators for metadata elements.
  - **Remapped Sentiment Scores** – Sentiment features re-incorporated for modeling convenience.

Each feature set was chosen to exploit known differences in timing, vocabulary, emotional tone, and typing style between tweets posted from iPhone vs. Android devices.

3. Classifiers Implemented

We employed five models from different model families for the classification task to find the best fit. Each model was trained regularly, except for the RoBERTa model, which was trained on an 80-20 train-test split due to the time-intensive process and limited Colab resources. The hyperparameters were chosen using a grid search. The models used with their respective parameters are as follows:

- **Logistic Regression:** A linear model for binary classification tasks. Parameters: {'C': 0.1, 'penalty': 'l2', 'solver': 'liblinear'}, with a threshold of 0.5.
- **SVM:** A powerful classification model that can use different kernel functions. Both linear and nonlinear kernels were experimented with. The best result was achieved using a linear kernel with {'C': 10}, and a threshold of 0.55.
- **FFNN:** A neural network model with multiple layers for capturing more complex patterns. The chosen architecture includes 2 hidden layers with dropout and ReLU activation, and a sigmoid activation in the output layer. Parameters: {'optimizer': Adam, 'epochs': 100, 'patience': 10}, }, and a threshold of 0.45.
- **XGBoost:** An efficient and scalable implementation of gradient boosting. Parameters: {'colsample_bytree': 0.9, 'gamma': 0.1, 'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 200, 'subsample': 0.9}, with a
- **RoBERTa-Base:** A transformer model pretrained on millions of unlabeled tweets We used word embeddings for the tweet text. Parameters: {'epochs': 5}.

4. Evaluation

To evaluate the performance of our classification models, we used several key metrics to gain a comprehensive understanding of their effectiveness. We assessed accuracy, precision, recall, F1, and ROC-AUC scores.

5. Results

Model	Accuracy	precision	recall	F1-score	ROC-AUC scores
Logistic Regression	0.806	0.884	0.704	0.783	0.881
Linear-SVM	0.801	0.864	0.715	0.782	0.876
FFNN	0.849	0.898	0.788	0.838	0.922
xgboost	0.862	0.883	0.838	0.858	0.942
Twitter RoBERTa-Base	**0.899**	**0.898**	**0.898**	**0.898**	**0.955**

6. Insights and Discussion

Models that lack mechanisms for capturing sequential or contextual dependencies such as Logistic Regression and SVM delivered less impressive results. Despite our comprehensive feature engineering, which accounted for various lexical, temporal, and emotional dimensions of each tweet, these models could not match the performance of the RoBERTa model. RoBERTa-Base proved to be the most effective and consistent classifier, achieving the highest overall accuracy (89.9%) alongside perfectly balanced precision, recall, and F1-score (0.898). While models like FFNN and XGBoost performed well especially when provided with a rich, hand-crafted feature set RoBERTa required no manual feature engineering and still outperformed them. This highlights the power of transformer-based models in capturing deep contextual and stylistic patterns directly from raw text, including subtleties that traditional models often miss.

Observations on Class Behavior:

Tweets authored by Trump (class 0) tended to exhibit more emotional intensity, informal punctuation (e.g., frequent exclamation marks), and were often posted during early morning hours. In contrast, tweets from staff members (class 1) showed a more formal tone and were temporally distributed more evenly throughout the day.