Sarcasm Detector: Project Report

Team: [Your Name / Team Name] **Date:** November 9, 2025

---

1. Project Approach

Our goal was to build a machine learning model to detect sarcasm in headlines, as per the "Sarcasm Detector" problem statement. The dataset was relatively small (~28,000 headlines), so our strategy focused heavily on **Feature Engineering** rather than complex deep learning models.

Our approach consisted of a hybrid feature set:

1. **Linguistic Features (TF-IDF):** To capture *what* is being said.
2. **Meta-Features:** To capture *how* it's being said (e.g., exaggeration, punctuation).

We used a `scikit-learn Pipeline` to combine these features, train two candidate models (Logistic Regression and SVM), and select the best-performing one using 5-fold cross-validation.

---

2. Features Used

Our final model was trained on a combination of two feature sets:

A. Text Features (TF-IDF)

- **Vectorizer:** `TfidfVectorizer`
- **N-grams:** `(1, 2)` - This captures both single words (e.g., "great") and two-word phrases (e.g., "yeah right"), which is crucial for understanding sarcastic context.
- **Parameters:** We removed English "stop words" and limited the vocabulary to the top 5,000 features to prevent overfitting.

B. Engineered Meta-Features We engineered four new features to quantify the *style* of the headline:

- `exclamation_count`: The number of exclamation marks, often a sign of emphasis or mock excitement.
- `question_count`: The number of question marks, often used in rhetorical sarcastic questions.
- `all_caps_count`: The number of words written in ALL CAPS (e.g., "WOW," "GREAT"), a strong indicator of exaggeration.
- `word_count`: The total number of words in the headline.

These numerical features were scaled using `StandardScaler` to be on the same level as the TF-IDF features.

---

3. Model Choice

We evaluated two models suitable for this type of classification task: **Logistic Regression** and **Linear Support Vector Machine (LinearSVC)**.

To get a reliable performance estimate, we used 5-fold cross-validation on our training data.

Cross-Validation Results:

- **Logistic Regression Mean Accuracy:** 0.8872
- **Linear SVM Mean Accuracy:** 0.9015

The **Linear SVM** consistently outperformed Logistic Regression in cross-validation. Therefore, we selected the `LinearSVC` as our final model.

The final, trained model achieved an accuracy of **[Check your output from Cell 7, e.g., 0.9034]** on the held-out test set.