Subtask 2: Baseline Model Design and Experimental Framework Sarcasm Detection in English-Hindi Code-Mixed Tweets using Random Forest

Jash Shah
202201016@dau.ac.in
Dhirubhai Ambani University

November 8, 2025

1. Introduction

This subtask focuses on implementing a reproducible and interpretable baseline model for sarcasm detection in English-Hindi code-mixed tweets. We build upon the dataset introduced by **Swami et al. (2018)**, which consists of 5,250 annotated tweets, with 504 labeled as sarcastic and 4,746 as non-sarcastic. A **Random Forest classifier** is implemented as the baseline model to establish an interpretable and effective starting point. All results presented here are based on the real experimental outputs obtained from the trained model.

2. Dataset Description and Preparation

Source: Swami et al. (2018), "A Corpus of English-Hindi Code-Mixed Tweets for Sarcasm Detection."

The dataset contains three columns: `tweet_id`, `text`, and `label`. Tweets were preprocessed to remove noise such as URLs, mentions, punctuation, and hashtags.

Label distribution:

- Non-sarcastic (0): 4,746 tweets

- Sarcastic (1): 504 tweets

To handle class imbalance, we used the original dataset for training but monitored F1-score to ensure fair performance across classes.

3. Experimental Setup

3.1 Data Splitting

The dataset was split using an 80-20 ratio:

- **Training set:** 4,200 tweets
- **Test set:** 1,050 tweets

3.2 Feature Extraction

Text features were extracted using the TF-IDF vectorizer:

- Vocabulary size: 5,000 most frequent terms
- N-gram range: (1, 2)
- Minimum document frequency: 5

This resulted in:

- TF-IDF matrix (train): (4,200, 5,000)
- TF-IDF matrix (test): (1,050, 5,000)

3.3 Model Configuration

Model: Random Forest Classifier

Parameters:

- `n_estimators` = 100

- `max_depth` = None
- `criterion` = "gini"
- `random_state` = 42

4. Experimental Results

4.1 Accuracy and F1-Score

- **Training Accuracy:** 1.0000
- **Test Accuracy:** 0.9419
- **Training F1 Score:** 1.0000
- **Test F1 Score:** 0.9367

The model performs well overall, achieving strong generalization on unseen test data.

4.2 Detailed Classification Report

Class	Precision	Recall	F1-Score	Support
Not Sarcastic (0)	0.95	0.99	0.97	949
Sarcastic (1)	0.79	0.53	0.64	101
Accuracy			**0.94**	1050
Macro Avg.	0.87	0.76	0.80	1050
Weighted Avg.	0.94	0.94	0.94	1050

4.3 Confusion Matrix

$$\begin{bmatrix} 935 & 14 \\ 47 & 54 \end{bmatrix}$$

Interpretation:

- True Negatives (Non-Sarcastic correctly classified): 935

- False Positives: 14
- False Negatives: 47
- True Positives (Sarcastic correctly classified): 54

The model achieves high accuracy for non-sarcastic tweets, but performance on sarcastic tweets remains moderate due to class imbalance.

4.4 Top 20 Most Important Features

The most influential TF-IDF features learned by the Random Forest model are listed below:

Rank	Feature	Importance
1	irony	0.1274
2	sarcasm	0.0704
3	politics	0.0621
4	cricket	0.0476
5	triple	0.0325
6	talaq	0.0311
7	triple talaq	0.0300
8	bollywood	0.0207
9	hai irony	0.0108
10	hai sarcasm	0.0086
11	ye	0.0077
12	hain	0.0076
13	hai	0.0063
14	aur	0.0061
15	kuch	0.0054
16	hain irony	0.0048
17	me	0.0047
18	to	0.0047
19	bhi	0.0047
20	ki	0.0044

These features reflect the linguistic mix of Hindi-English expressions, domain-specific keywords (politics, cricket, bollywood), and explicit sarcasm indicators (irony, sarcasm).

5. Discussion and Observations

- The model achieves high accuracy (94.2%) and F1-score (93.6%) overall.
- It demonstrates strong detection of non-sarcastic tweets but moderate recall for sarcastic tweets due to class imbalance.

- Top features indicate clear topic and sentiment cues related to sarcasm-heavy domains such as politics and entertainment.
- Training accuracy of 100% suggests overfitting, although test performance remains strong.

6. Output Structure and Reproducibility

- **Predictions:** Stored as `predictions.csv` with columns: `tweet_id`, `text`, `true_label`, `predicted_label`, `confidence_score`.
- **Logs:** Training and testing metrics stored in `results_log.csv`.
- **Artifacts:** TF-IDF vectorizer and trained model serialized for reuse.
- **Visualization:** Confusion matrix and feature importance plotted for interpretability.

7. Conclusion and Next Steps

The Random Forest baseline on the Swami et al. (2018) dataset provides a strong starting point for sarcasm detection in English-Hindi code-mixed tweets. The model achieves over 94% test accuracy and successfully identifies linguistic markers associated with sarcasm. Future improvements will include:

- Addressing class imbalance using SMOTE or data augmentation.
- Exploring deep learning approaches such as Bi-LSTM and attention-based models.
- Incorporating semantic and contextual embeddings (e.g., FastText, mBERT).

References

- Swami, S., Khandelwal, A., Singh, V., Akhtar, S. S., & Shrivastava, M. (2018). *A Corpus of English-Hindi Code-Mixed Tweets for Sarcasm Detection.* arXiv:1805.11869.

- Joshi, A., Bhattacharyya, P., & Carman, M. J. (2017). *Automatic Sarcasm Detection: A Survey.* ACM Computing Surveys.
- Bouazizi, M., & Ohtsuki, T. (2016). *A Pattern-Based Approach for Sarcasm Detection on Twitter.* IEEE Access.