Milestone 1: Dataset Identification and Preliminary Data Modeling Plan for Sarcasm Detection in Hindi-English Code-Mixed Texts

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

October 29, 2025

1. Introduction

Sarcasm detection is a complex sub-task within sentiment analysis and social media text mining. The challenge becomes more pronounced when dealing with *code-mixed* data, where speakers alternate between languages, commonly Hindi and English (Hinglish) in Indian social media contexts. Traditional natural language processing (NLP) methods often struggle with such data due to irregular grammar, transliteration, and non-standard spelling patterns.

Our project aims to design and implement a sarcasm detection system using bilingual word embeddings and deep learning architectures, drawing on the approaches proposed by Aggarwal et al. (2020) and Swami et al. (2018). This milestone establishes our data backbone, identifying suitable datasets, defining preprocessing strategies, and summarizing relevant research.

2. Dataset Description

2.1 Primary Dataset

The primary dataset is the **Hindi-English Code-Mixed Sarcasm Corpus** introduced by **Aggarwal et al. (2020)** in their EMNLP W-NUT paper titled *"Did you really mean what you said?"*.

This dataset contains approximately **106,899 tweets**, balanced between sarcastic (52,587) and non-sarcastic (54,312) examples. The tweets were collected using the *TwitterScraper API* with hashtags such as #sarcasm, #irony, #humor, #bollywood, and #cricket. Each tweet was annotated for the presence of sarcasm, with additional quality control performed to minimize noise.

The dataset is designed specifically for deep learning models and supports the creation of bilingual word embeddings for Hindi-English text.

2.2 Supporting Dataset

As a secondary resource, we employ the **A Corpus of English-Hindi Code-Mixed Tweets for Sarcasm Detection** (Swami et al., 2018). This dataset comprises **5,250 tweets**, of which 504 are sarcastic. It provides both tweet-level sarcasm labels and token-level language annotations (`en`, `hi`, `rest`).

This dataset complements our primary dataset by offering well-annotated smaller-scale data useful for baseline comparisons, language identification, and early-stage feature analysis.

3. Data Processing and Preparation

3.1 Preprocessing Pipeline

The datasets will undergo extensive preprocessing to handle social media noise:

- Removal of URLs, mentions, hashtags, and punctuation marks.
- Normalization of text (handling transliteration inconsistencies).
- Tokenization and filtering of rare words (occurrence < 10).
- Balanced sampling to avoid bias toward non-sarcastic classes.
- Hashtag decomposition (e.g., #ILoveIndia $\rightarrow$ `I Love India`).

3.2 Data Representation

Following Aggarwal et al. (2020), we will create **bilingual word embeddings** using both Word2Vec and FastText approaches:

- Word2Vec embeddings trained on 427k Hinglish tweets + 300k English tweets.
- FastText embeddings trained on subword units to handle spelling variations.
- Embedding dimension: 300; window size: 10; negative sampling enabled.

The embeddings will be used to initialize deep learning models for sarcasm classification.

4. Preliminary Data Modeling Plan

4.1 Model Architectures

We plan to reproduce and extend the three neural architectures explored by Aggarwal et al. (2020):

1. LSTM
2. Bi-directional LSTM
3. Attention-based Bi-directional LSTM

Each model takes the bilingual word embeddings as input and outputs a binary classification label (sarcastic or non-sarcastic). The best performance reported in the literature is an accuracy of **78.49%** using the Attention Bi-LSTM model with Word2Vec embeddings.

4.2 Baseline and Comparison Models

For benchmarking, we will include traditional machine learning classifiers from Swami et al. (2018):

- Random Forest
- Support Vector Machine (RBF Kernel)

- Linear SVM

These baselines achieved F-scores around **78.4%** on smaller code-mixed datasets.

5. Literature Review

- **Aggarwal et al. (2020):** Proposed deep learning models with bilingual word embeddings for sarcasm detection in Hinglish tweets. Demonstrated the effectiveness of Bi-LSTM with attention in capturing contextual nuances.
- **Swami et al. (2018):** Introduced the first English-Hindi code-mixed sarcasm corpus, annotated at both tweet and token levels. Established baseline results using traditional ML models.
- **Joshi et al. (2017):** Provided a comprehensive survey on automatic sarcasm detection methods, emphasizing the need for context-aware modeling.
- **Bouazizi & Ohtsuki (2016):** Introduced a pattern-based sarcasm detection method using syntactic and lexical cues.
- **Wang et al. (2016):** Developed attention-based LSTM for sentiment analysis, later adapted for sarcasm detection in multilingual contexts.

6. Expected Outcomes

By the end of this phase, we expect to have:

- A cleaned and structured Hinglish sarcasm dataset ready for modeling.
- Pretrained bilingual embeddings capturing Hindi-English semantics.
- A baseline evaluation of deep learning vs. traditional models.
- An informed roadmap for model optimization and feature exploration.

7. References

- Aggarwal, A., Wadhawan, A., Chaudhary, A., & Maurya, K. (2020). *"Did you really mean what you said?": Sarcasm Detection in Hindi-English Code-Mixed Data using Bilingual Word Embeddings.* EMNLP W-NUT.

- 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.

- Wang, Y., Huang, M., Zhu, X., & Zhao, L. (2016). *Attention-based LSTM for Aspect-Level Sentiment Classification.* EMNLP.