Paraphrase Detection in Bangla (Bangladesh National Corpus Paraphrase Model): Bug Identification and Analysis

Submitted By:
Dola Chakraborty
Khadiza Sultana Sayma

1. Introduction

Paraphrase detection is a critical task in natural language processing (NLP) aimed at determining whether two given sentences have the same meaning. In the context of Bangla, this task has significant applications in machine translation, sentiment analysis, question answering, and text summarization. The goal of this report is to evaluate the performance of a paraphrase detection model in Bangla by testing it with various examples and identifying potential bugs.

2. Categories of Test Cases

1. **True Paraphrases:** Sentences that convey the same meaning using different words or structures.
2. **Related but Not Paraphrased:** Sentences that share context or topic but differ in meaning.
3. **Completely Unrelated Pairs:** Sentences with no semantic or lexical relationship.

3. Our Approach

We manually input sentences to test the paraphrase model.

4. Findings

The model exhibited several issues, primarily in detecting non-paraphrased pairs as paraphrases. The identified bugs are detailed below:

Bug 1: Semantic Similarity Confusion

Description:

The model incorrectly identifies pairs with related topics but distinct meanings as paraphrases. This issue arises from an over-reliance on semantic similarity.

Example 1:

- **Sentence 1:** পদ্মা সেতু বাংলাদেশের দীর্ঘতম সেতু।
- **Sentence 2:** পদ্মা সেতু একটি স্থাপত্যিক মাইলফলক।

Analysis: While both sentences discuss the Padma Bridge, the first emphasizes its length, whereas the second highlights its architectural significance. These are not paraphrases but were detected as such.

Example 2:

- **Sentence1 :** কিভাবে আমি একজন সফল গ্রাফিক্স ডিজাইনার হতে পারি?
- **Sentence 2:** দক্ষ গ্রাফিক্স ডিজাইনার হতে হলে কোন সফটওয়্যারগুলো শেখা উচিত?

Analysis: This error occurs because the sentences share similar semantic contexts and keywords (e.g., "গ্রাফিক্স ডিজাইনার" and "হতে") despite having distinct intents. The model may focus on shared vocabulary and thematic similarity, misinterpreting them as paraphrased.

Bug 2: Word Overlap Confusion

Description:

The model struggles with pairs that have high lexical overlap but diverge in meaning. This often leads to false positives.

Example 1:

- **Sentence 1:** বাংলাদেশে ধান চাষ প্রধান কৃষি কার্যক্রম।
- **Sentence 2:** বাংলাদেশে ধানের পাশাপাশি চা উৎপাদনও বিখ্যাত।

Analysis: The sentences share words like “ধান” and “উৎপাদন,” but their meanings differ significantly. One emphasizes rice cultivation, while the other includes tea production.

Example 2:

- **Sentence 1:** কোন সফটওয়্যারগুলো শেখা একজন নতুন গ্রাফিক্স ডিজাইনারের জন্য সবচেয়ে কার্যকর হবে?
- **Sentence 2:** নতুন গ্রাফিক্স ডিজাইনারের জন্য কোন সফটওয়্যারগুলো গুরুত্বপূর্ণ?

Analysis: Although they overlap in discussing software for designers, the focus differs, with one addressing utility and the other emphasizing importance. These are not paraphrases.

Bug 3: Ambiguous or Double Meaning

Description:

The model fails to handle semantically ambiguous pairs, leading to incorrect paraphrase detection.

Example:

- **Sentence 1:** তিনি প্রতিদিন একটি নতুন বই পড়েন।
- **Sentence 2:** তিনি প্রতিদিন একটি বই পড়েন এবং লেখেন।

Analysis: The first sentence focuses solely on reading, while the second includes both reading and writing. The ambiguity of intent causes the model to incorrectly classify these as paraphrases.

Bug 4: Keyword Bias Misclassification

Example 1:

- **Sentence 1:** কোন সফটওয়্যারগুলো শেখা একজন নতুন গ্রাফিক্স ডিজাইনারের জন্য সবচেয়ে কার্যকর হবে?
- **Sentence 2:** সফটওয়্যারগুলো

Example 2:

- **Sentence 1:** কোন সফটওয়্যারগুলো শেখা একজন নতুন গ্রাফিক্স ডিজাইনারের জন্য সবচেয়ে কার্যকর হবে?
- **Sentence 2:** গ্রাফিক্স ডিজাইনার

Analysis: This error arises because the model relies heavily on the presence of shared keywords, such as "সফটওয়্যারগুলো", “ গ্রাফিক্স ডিজাইনার”, “ গ্রাফিক্স” while ignoring the disparity in sentence length and context. The model's inability to discern that one sentence is incomplete leads to the incorrect classification.

Bug 5: Incomplete Paraphrases

Description:

Pairs where one sentence is more detailed than the other are wrongly classified as paraphrases.

Example:

- **Sentence 1:** পদ্মা সেতু বাংলাদেশের দীর্ঘতম সেতু।
- **Sentence 2:** পদ্মা সেতু বাংলাদেশের দীর্ঘতম সেতু, যা ২০২২ সালে উদ্বোধন করা হয়।

Analysis: The additional detail in the second sentence (inauguration year) makes it distinct from the first. However, the model considers them paraphrases due to shared core content.

Bug 6: Structural Similarity Misclassification

Example :

- **Sentence 1:** একজন ওয়েব ডেভেলপার হতে গেলে কোন প্রোগ্রামিং ভাষা শেখা উচিত?

- **Sentence 2:** একজন ওয়েব ডেভেলপার হয়ে কিভাবে ফ্রিল্যান্সিং শুরু করবেন?

Analysis: The error occurs because the model focuses on shared phrases like "একজন ওয়েব ডেভেলপার," overlooking the difference in intent (learning programming languages vs. starting freelancing). This happens due to over-reliance on lexical similarity and insufficient contextual understanding.

Bug 7: Fragment vs. Full Sentence Misclassification

Example:

- **Sentence 1:** গিট এবং গিটহাবের মধ্যে পার্থক্য কী?
- **Sentence 2:** পার্থক্য

Analysis: The error occurs because the model heavily relies on overlapping keywords like "পার্থক্য," ignoring the fact that one sentence is a complete query while the other is an isolated fragment. This misclassification happens due to insufficient context awareness and an inability to differentiate between fragments and full sentences.

Bug 8: Negation Misclassification

Example:

- **Sentence 1:** আপনি কি মনে করেন প্রোগ্রামিং শেখা কঠিন?
- **Sentence 2:** আপনি কি মনে করেন প্রোগ্রামিং শেখা সহজ?

Analysis: The model misclassifies these sentences as paraphrased because it overlooks the subtle semantic difference introduced by the opposing words কঠিন (difficult) and সহজ (easy). It relies too heavily on the structural similarity and shared context (আপনি কি মনে করেন প্রোগ্রামিং শেখা), failing to account for negation or contradictory meanings.

Bug 9: Contextual Misclassification

Example:

- **Sentence 1:** ডেটা সায়েন্স শেখার সহজ উপায় কী?
- **Sentence 2:** ডেটা সায়েন্স কি শেখা সহজ?

Analysis: The model misclassifies these sentences as paraphrased because it focuses on shared phrases like ডেটা সায়েন্স and সহজ, while failing to grasp the distinct intents. The first sentence inquires about methods to learn data science, while the second questions the general ease of learning it. This happens due to over-reliance on lexical overlap and insufficient understanding of sentence-level context.

Bug 10:

Example:

- **Sentence 1:** ডেটা সায়েন্স কি শেখা সহজ?
- **Sentence 2:** শেখার সহজ উপায় কী?

The two sentences differ completely in both context and semantics. The model still predicts the first sentence as a paraphrase of the second.

---

5. Root Cause Analysis

Key Factors Behind the Bugs

1. **Over-reliance on Lexical Similarity:**
   - The model tends to equate high word overlap or semantic similarity with paraphrasing, leading to false positives.
2. **Insufficient Contextual Understanding:**
   - The model struggles to account for nuanced differences in meaning or context, particularly in cases of ambiguity or incomplete information.
3. **Lack of Training Data for Complex Variations:**
   - The model's training data likely underrepresents cases involving structural variation, idiomatic expressions, or semantic nuance, causing these scenarios to be misclassified.

6. Conclusion

This analysis highlights critical weaknesses in the paraphrase detection model for Bangla, including issues with semantic similarity confusion, word overlap, ambiguity, and incomplete paraphrases. Addressing these issues through improved training data, contextual understanding, and model architecture enhancements will significantly improve the model's accuracy and reliability.