Cognitive Distortion Classification Pipeline

Gayoung Kim (2024149067)

1 Task Definition

1.1 Task Description

This project performs multi-class text classification on Korean sentences targeting three Cognitive Distortions: All-or-Nothing Thinking, Catastrophizing, and Should Statements. The operational definitions are as follows:

1. **All-or-Nothing Thinking:** Focuses on **evaluating current status** in extremes.
2. **Catastrophizing:** Focuses on **predicting future consequences**, irrationally exaggerating a minor event into a disaster.
3. **Should Statements:** Focuses on **imposing rigid rules** or pressure on oneself/others.

1.2 Motivation

Identifying negative thinking patterns is essential for Cognitive Behavioral Therapy, especially in the digital age where social media often worsens such errors. However, people often find it difficult to recognize these patterns without professional help. While binary classification is technically easier, identifying the specific type of error is more important for behavioral correction. Thus, this pipeline aims to be a tool that improves user self-awareness.

1.3 Input / Output

- **Input:** A Korean sentence (e.g., "다들 짝이 있는데 나만 솔로야, 난 평생 고독사 확정임")
- **Output:** One of the 3 classes (Label 0, 1, or 2).

1.4 Success Criteria

The goal is to go beyond simple keyword detection. The model should understand **context and nuance**, even if specific keywords are absent.

- **Qualitative Goal:**
  1. Understand **metaphorical expressions** and **internet slang**.
  2. Distinguish ambiguous cases based on the criteria defined above.
- **Quantitative Goal:** Achieving a significantly higher accuracy than the naive baseline model.

2 Methods

2.1 Naïve Baseline

- **Method description:**

  - **Approach:** A **Keyword Counting** classifier (Rule-based).
  - **Logic:** It counts specific trigger words (e.g., "망했" → All-or-Nothing, "죽을" → Catastrophizing, "해야" → Should Statements).
- **Why naïve:** It completely ignores context and relies solely on the presence of words.
- **Likely failure modes:**
  - **Failure to Detect Nuance:** It cannot interpret metaphors like *"짐 싸야 하나"* (implying fear of dismissal) or internet slang like *"국룰"* (social norm).
  - **Failure to Grasp Context:** For example, **"면접관이 인상을 썼으니, 난 분명히 떨어졌어"** is a clear case of **Catastrophizing** (predicting a negative future). However, the baseline classifies it as a **Should Statement** simply because of the keyword **"definitely" (unconditionally)**.

2.2 AI Pipeline

- **Models used: 'klue/bert-base'** (Pre-trained on a large Korean corpus).
- **Pipeline stages:**
  1. **Tokenization:** Convert sentences using the BERT tokenizer.
  2. **Embedding:** Generate context-aware vectors via the BERT encoder.
  3. **Classification:** Predict probabilities for the 3 classes using a classification head.
- **Design choices and justification:** To overcome the limitations of keyword matching, I utilized the **Self-Attention mechanism** to capture the semantic relationships within the entire sentence. The model was fine-tuned for 3 epochs on a T4 GPU.

3 Experiments

3.1 Datasets

- **Source:** I referenced **blogs, SNS posts, and comments** to capture realistic tones and expressions used in daily life. (Data augmentation was not applied to maintain data quality.)
- **Total Size: 293 examples** (Balanced: **approximately 97 to 98 examples per class).**
- **Train/Test split:** 80% Training / 20% Testing
- **Preprocessing steps:** Minimal cleaning (whitespace removal). I used the 'klue/bert-base' tokenizer because it handles Korean morphology well, preserving the original nuances.

3.2 Metrics

I used **Accuracy** as the primary metric.

3.3 Results

Report:

Method	Metric 1 (Accuracy)	Note
Naive Baseline	59.32%	Better than random guess (33%)
AI Model (BERT)	**89.83%**	**Significant Improvement (+30.51%p)**

Qualitative Analysis (Comparison Cases)

I analyzed 3 cases where the Baseline failed but the AI Model succeeded.

- **Case 1: Implied Meaning**
  - **Text:** "회식도 업무의 연장이야, 빠지는 건 직무 유기지."
  - **Analysis:** There is no direct "should" keyword. However, the AI understood that **"직무 유기" (dereliction of duty)** implies a strong social pressure or rule, correctly classifying it as a **Should Statement**.
- **Case 2: Limitation of Pre-Defined Keywords**
  - **Text:** "다들 짝이 있는데 나만 솔로야, 난 평생 고독사 확정임"
  - **Analysis: The** baseline failed due to a lack of matching keywords. The AI correctly identified **All-or-Nothing Thinking**, recognizing the speaker's rigid equation of **"솔로" (being single)** with **"평생 고독사" (absolute failure/lonely death)**.
- **Case 3: Informal Language and Lack of Generalization**
  - **Text:** "중간고사를 조졌으니 기말고사 잘 봐봤자 학점 세탁은 불가능해."
  - **Analysis: The** baseline failed to process the informal slang **"조지다" (ruined/bombed)**. The AI, however, correctly identified **All-or-Nothing Thinking** by grasping the context of "조지다" and **"학점 세탁" (grade recovery)**, understanding the binary logic where a single failure renders future success impossible.

4 Reflection and Limitations

4.1 Successes (What worked well?)

- **Dataset Construction Strategy:** This was the most successful aspect. By curating data from real blogs and SNS comments, the model learned to process realistic tones.
- **Metric Validity:** The **30%p increase in accuracy** quantitatively demonstrated the model's quality. Furthermore, the qualitative analysis confirmed that the AI model could detect metaphors that the rule-based approach missed.

4.2 Limitations & Future Work

- **Metric Limitation:** Accuracy was an appropriate metric due to the balanced dataset. However, a single score could not reveal *which* specific classes were being confused. **Therefore, I manually inspected the prediction results** to analyze error patterns. A **Confusion Matrix** would have been a valuable supplementary metric.
- **Ambiguity:** Dealing with sentences where distortions co-occurred was challenging (e.g., "시험을 망치면 내 인생은 끝이야" exhibits traits of both All-or-Nothing and Catastrophizing). This ambiguity likely affected the model's precision in borderline cases.
- **Absence of 'Normal' Class:** The current model is a closed-set classifier, forcing every sentence into a distortion category (even neutral ones).
- **Future Work:** I plan to implement **Multi-label Classification** to better model complex human psychology. Additionally, for real-world deployment, I would introduce a **'Neutral' class** or a **Confidence Threshold** to filter out non-distorted inputs.