Report

ABSA LEXICON SENTIMENT FOR VIETNAMESE

Prepared by Van Ha Tran (Tyranno)

AIR – Lab NLP&KD – FIT – Ton Duc Thang University
lt.tdtu.edu.vn
Spring 1, 2026

OVERVIEW

EXECUTIVE SUMMARY

This report presents a Vietnamese lexicon-based sentiment analysis that classifies text as positive, negative, or neutral without model training. The method uses VietSentiWordNet for word/phrase sentiment scores, NegDict to handle negation (polarity inversion within a short context window), and SacThaiDict to adjust intensiry (e.g., “rất”, “có_thể”) via weighting. The pipeline includes text normalization (including Unicode normalization) and Vietnamese tokenization to ensure multi-word terms are matched consistently. The result is a lightweight, interpretable baseline that is fast to deploy, with known limitations on sarcasm, long-range context, and new slang or domain-specific vocabulary.

1. PROJECT OBJECTIVES

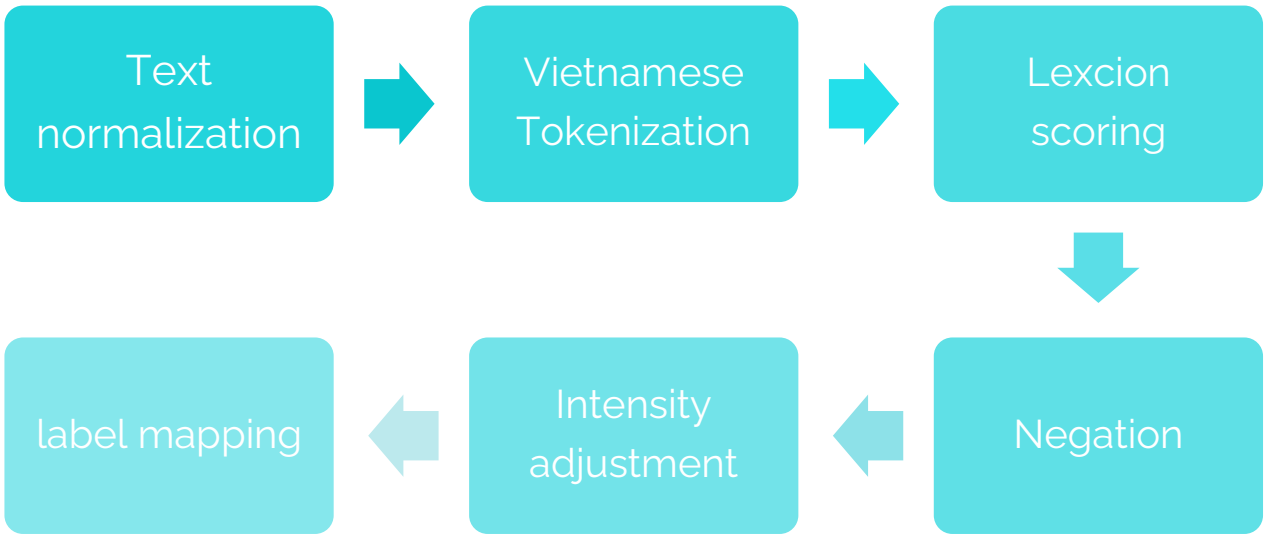
The primary objectives of this project included:

- Building a Vietnamese lexicon-based sentiment classifier (pos/neu/neg).
- Integrating VietSentiWordNet [1] for polarity scoring, NegDict [1] for negation handing, and SacThaiDict [1] for intensity weighting.
- Producing a reproducible notebook workflow that can generate predictions on a CSV dataset export results.

Secondary objectives included ensuring interpretability (rule-base s-coring) and establishing a baseline for future machine learning comp-arisons.

2. APPROACH / METHODOLOGY

The system follows a lightweight rule-based pipeline:



Lexicon scoring: each token receives a sentiment value computed from VietSentiWordNet [1] (PosScore – NegScore).

Negation handling: polarity is inverted when a negation term appears within a short context window.

Intensity handling: sentiment strength is scaled using degree words (e.g., “rất”, “quá”) with predefined weights.

3. PERFORMANCE METRICS & RESULTS

The model was evaluated on data_graded.csv[2] using standard classification metrics.

Accuracy reached 0.5745, and Macro-F1 achieved 0.5761, indicating moderate performance for a fast, training-free baseline.

Results also suggest that errors mainly come from complex linguistic patterns such as longer negation scope, contrastive phrasing, and implicit/sarcastic sentiment.

4. LINK GITHUB SOURCE CODE

https://github.com/vanha2301/AIR-absa-rt/blob/main/lexicon_basic.ipynb

REFERENCES

[1] link github: https://github.com/sonvx/VietSentiWordNet

[2] link github: https://github.com/stopwords/vietnamese-stopwords