CE807-7-SU : Text Analytics - Review Rating Classification Presentation

By Thuan Anh Bui
Student ID: 2412204

Introduction

Impact of tokenization on multi-class text classification
(review ratings 1–5)

Classifier: Logistic Regression

Tokenization Overview

Two techniques compared

- **Traditional**:
  - spaCy Lemmatization
  - Term Frequency-Inverse Document Frequency (TF-IDF)
- **BERT Tokenizer**:
  - WordPiece Segmentation
  - Term Frequency-Inverse Document Frequency (TF-IDF)

Traditional Tokenization

Stages:

Cleaning → Tokenization → Lemmatization → Stopword removal

Example workflow:

“exactly what I needed. works perfectly. Arrived on time.
Thank you”

→ ['exactly', 'need', 'work', 'perfectly', 'arrive', 'time', 'thank']

BERT Tokenizer

BERT Tokenization (bert-base-uncased):
➜ ['exactly', 'what', 'i', 'needed', '.', 'works', 'perfectly', '.', 'arrived', 'on', 'time', '.', 'thank', 'you']

Joined Tokens for TF-IDF Input:
➜ “exactly what i needed . works perfectly . arrived on time . thank you”

Critical Comparison

Same preprocessing for standardized input

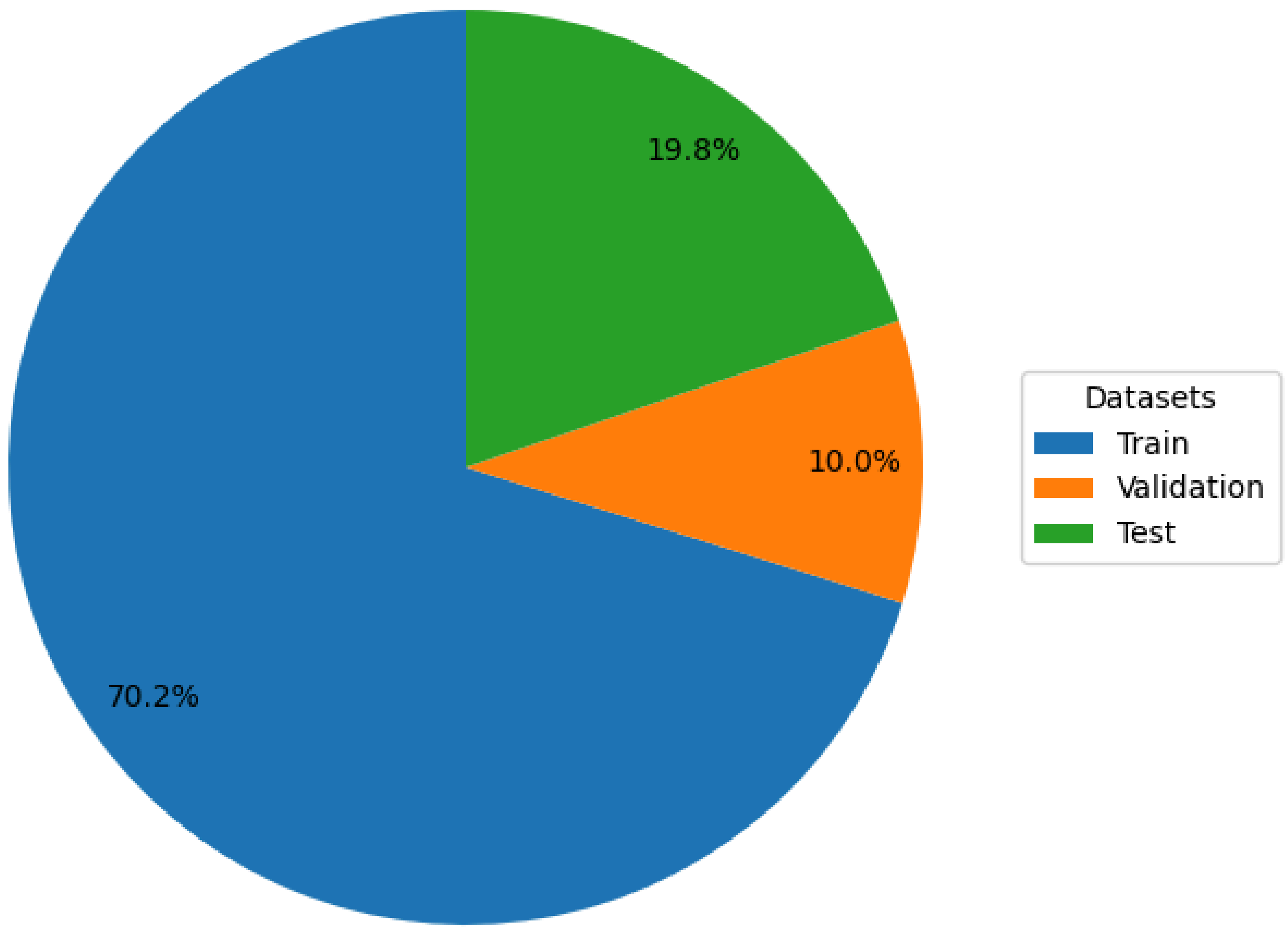
Traditional

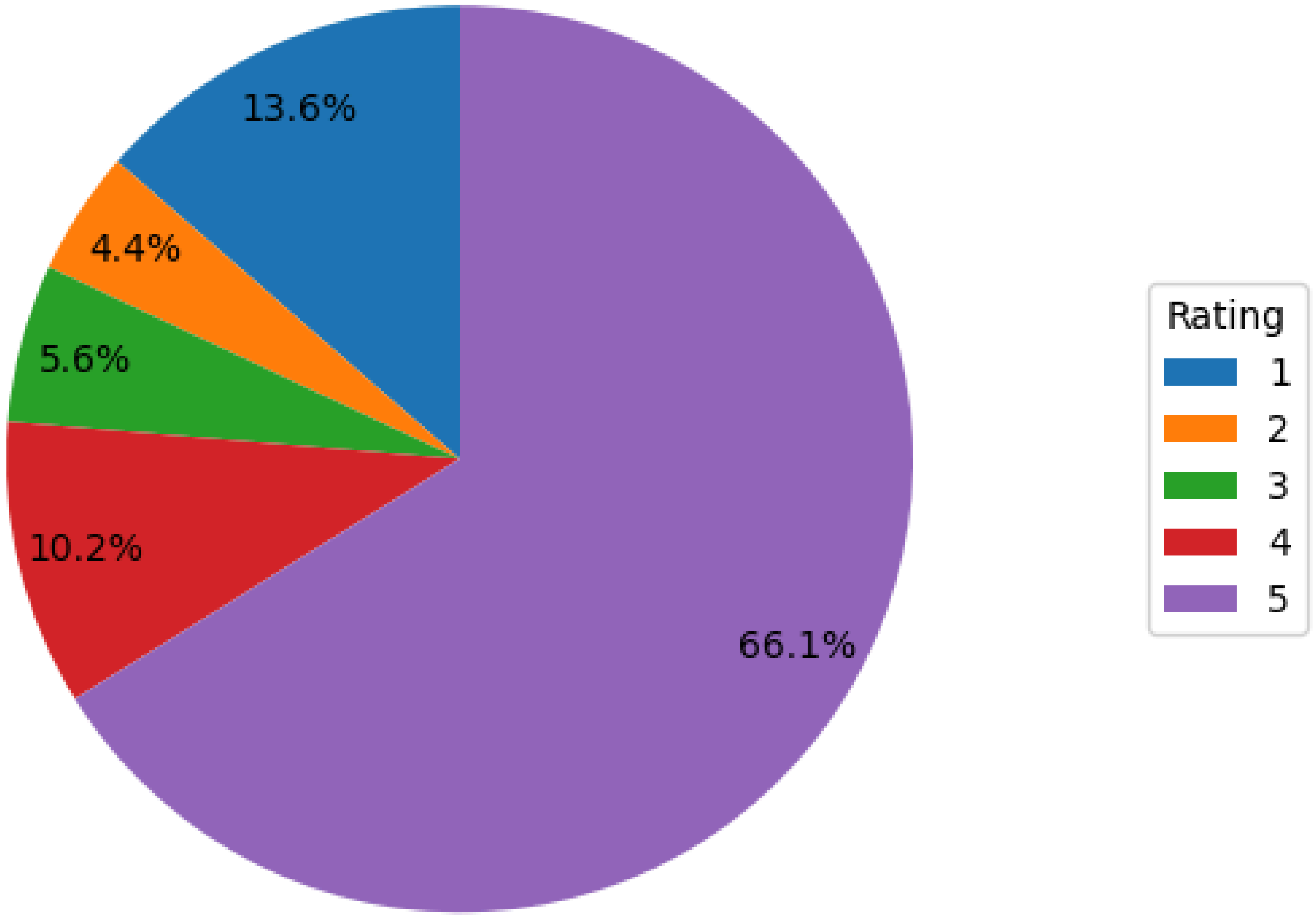
- Punctuations and stopwords removed
- Simplifying input to focus on high-value words

BERT Tokenizer

- Punctuations retained, no stopword removed or lemmatization
- Preserves original grammar and detail

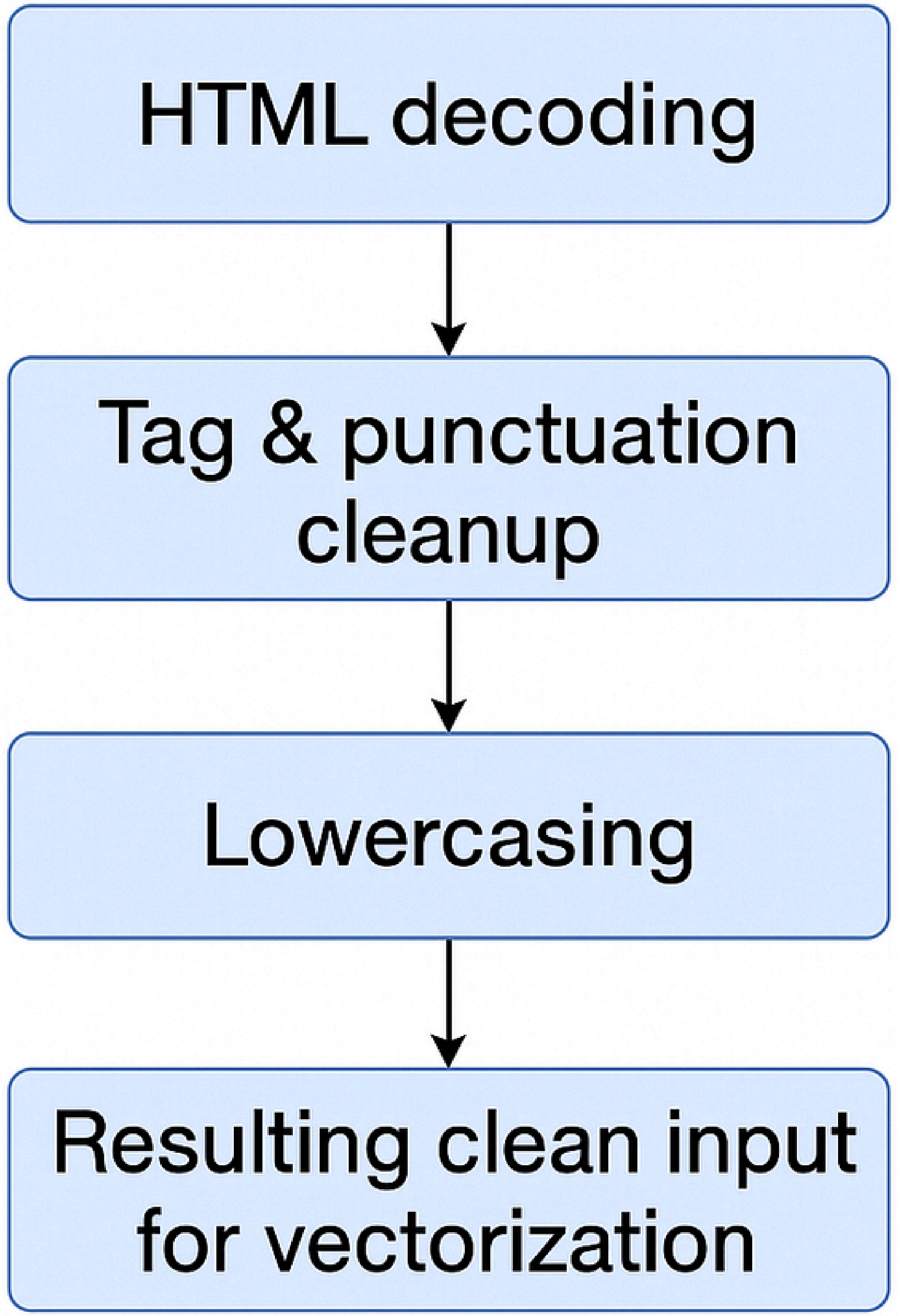
Dataset





Training and Validation

Preprocessing Workflow



Vectorization Details

Traditional Tokenization

- TF-IDF with unigrams-bigrams
- TF-IDF with unigrams–trigrams

BERT Tokenization

- TF-IDF on subword-joined tokens
- 20,000 max features
- Sublinear term frequency

Results

NOTE: The accuracy rate for BERT is 72.71% but I had issue with Canva formatting

Model	Accuracy	F1 Score
Traditional	71,57	6.316
BERT Tokenizer	7.271	6.445

Model Performance on Selected Examples

- BERT better at nuanced, context-rich reviews
- Traditional performs well when sentiment is explicit
- Both models struggle with mixed tone, structural noise

ID	**Text** (Full text in Appendix A)	True Rating	Model 1	Model 2
302654	I read several of the reviews. . . [R1]	5	5	5
56099	Didn't purchase from Amazon but posting. . . [R2]	1	1	5
503538	I'll be honest, I was very. . . [R3]	5	5	5
574685	What Amazon's product page says: \br / \br. . . [R4]	4	5	5
136885	[[VIDEOID:4fd222938153f33c6f93079d83e0720d]] I'm so glad I bought. . . [R5]	5	4	4

Feature & Algorithm Choices

Token-level representations (unigram, bigram, trigram) were critically analyzed to enhance sentiment understanding in review texts.

- **Unigrams** like “happy” and “perfect” reflect *strong* positive sentiment.
- Negated or intensified expressions are *only* captured with **bigrams/trigrams**.

Adjective	Frequency
great	889
perfect	304
happy	170
nice	131
clean	106
better	137
best	97
excellent	81

Phrase	Frequency
disappointed	33
expensive	21
very disappointed	21
not worth	18
too much	10
not happy	4

Comparison to SoTA

Logistic Regression:

- Simplicity and interpretability
- Goal: Isolate the effect of tokenization, not model complexity

State-of-the-art models like BERT, RoBERTa or DistilBERT

- Contextual embeddings
- End-to-end deep learning

BERT tokenizer already improves performance ... even without full transformer-based classification

Lesson Learned

- Preprocessing Matters
- N-grams Add Context
- BERT Tokenizer Helps
- Model-Tokenizer Fit

Thank You