Data Mining

ASPECT-BASED SENTIMENT ANALYSIS (ABSA)

From Customer Reviews in the Smartphone Domain

Advisor: PhD. Hoang Anh

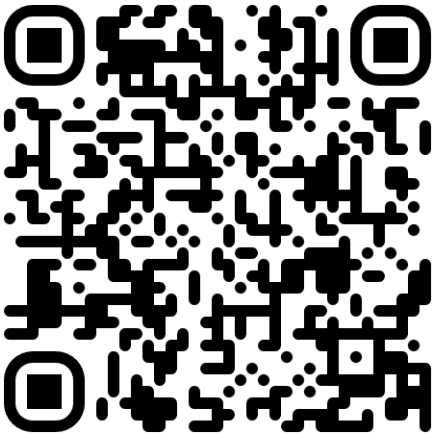
Research Group:

1. Hoàng Lê Minh Nhật - 523H0064
2. Phạm Nguyên Anh - 523H0117
3. Nguyễn Quang Huy - 523H0140

QR Code

Presentation slide

Implementation



Github Resources

Outline

1. Introduction
2. Challenges in Vietnamese NLP
3. Dataset and Data Augmentation
4. Architecture
5. Experimental Setup & Results
6. Future Works
7. Q&A

1. Introduction

The Context

- E-commerce in Vietnam is growing rapidly, making user reviews a valuable resource for business intelligence.
- Current Vietnamese feedback systems (e.g., Shopee, Lazada) mostly use simple star ratings, lacking fine-grained analysis.

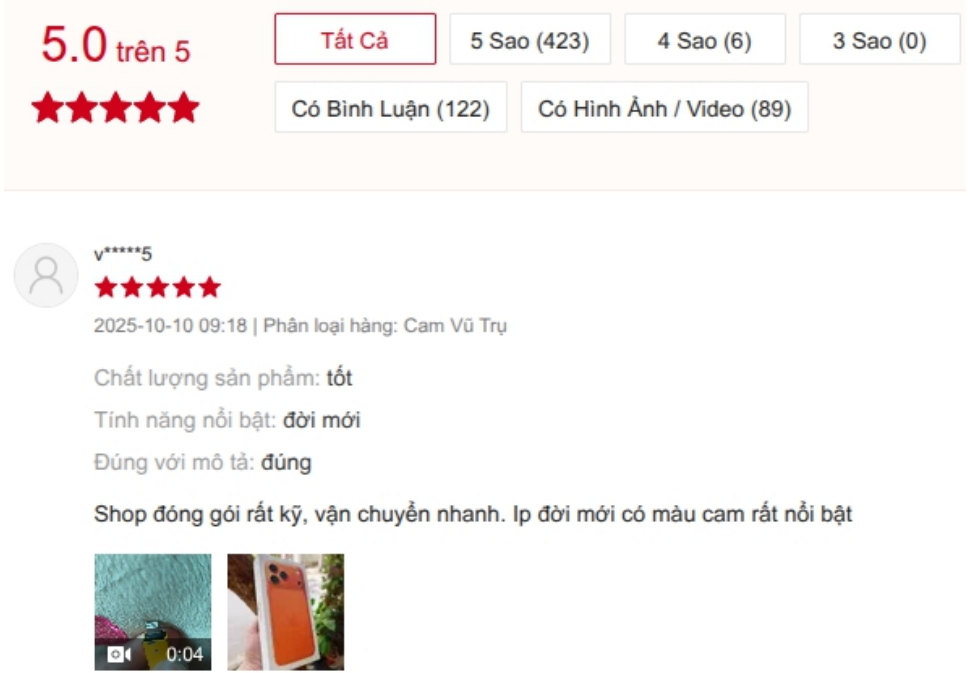
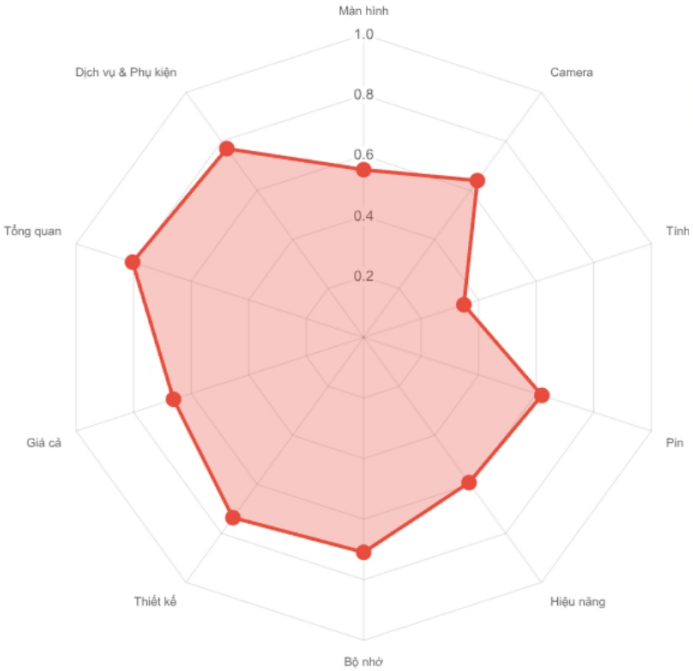


Figure 1.1. Review section of Shopee

1. Introduction

Expect



Khía cạnh	Tích cực	Trung tính	Tiêu cực	Cảm xúc chung
Màn hình	512 (53.2%)	41	409 (42.5%)	Tích cực (962 lượt)
Camera	1260 (57.7%)	272	650 (29.8%)	Tích cực (2182 lượt)
Tính năng	922 (33.1%)	99	1767 (63.4%)	Tiêu cực (2788 lượt)
Pin	2300 (58.3%)	287	1356 (34.4%)	Tích cực (3943 lượt)
Hiệu năng	2709 (56.4%)	272	1818 (37.9%)	Tích cực (4799 lượt)
Bộ nhớ	57 (61.3%)	18	18 (19.4%)	Tích cực (93 lượt)
Thiết kế	1078 (72.0%)	47	372 (24.8%)	Tích cực (1497 lượt)
Giá cả	541 (54.4%)	234	220 (22.1%)	Tích cực (995 lượt)
Tổng quan	4398 (78.0%)	263	976 (17.3%)	Tích cực (5637 lượt)
Dịch vụ & Phụ kiện	1426 (76.1%)	36	413 (22.0%)	Tích cực (1875 lượt)

Figure 1.2. Demostration

1. Introduction

The Context

"Máy đẹp, sang, sd thì rất là ok, máy mượt, pin yếu"

- Traditional Sentiment Analysis: Positive
- Aspect-Based Sentiment Analysis:

Máy đẹp, sangDESIGN#POSITIVE,

sd thì rất là okGENERAL#POSITIVE

máy mượtPERFORMANCE#POSITIVE

pin yếuBATTERY#NEGATIVE

1. Introduction

The Task ABSA - Sequence Labeling

- Input: A customer feedback text.
- Goal: Identify specific text spans (opinions), categorize them by Aspect (e.g., CAMERA, BATTERY), and assign Sentiment (Positive, Negative, Neutral).
- Example:

Máy đẹp, sangDESIGN#POSITIVE,
sd thì rất là okGENERAL#POSITIVE
máy mượtPERFORMANCE#POSITIVE.

1. Introduction

The Idea

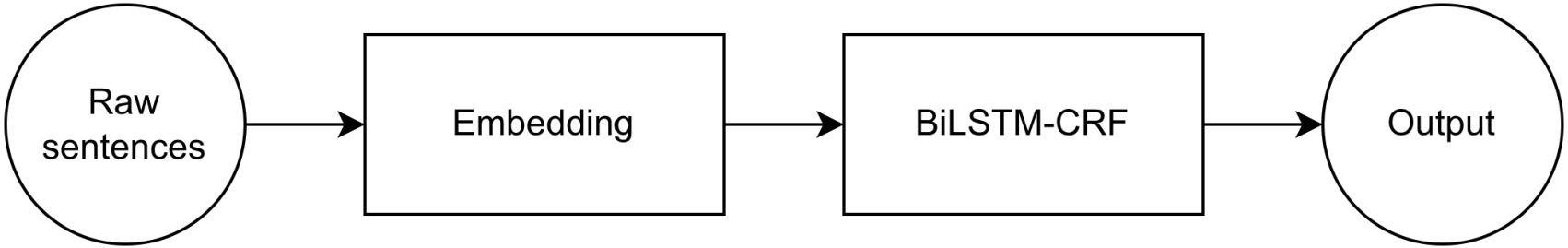


Figure 1.2. Idea Architecture

2. Challenge in Vietnamese NLP

Linguistic Characteristics

- Loanwords: Technology reviews contain many English loanwords which models may misclassify.

  “Chơi game người ta bắn chết lâu rồi mới quay được lại bắn trả”

- Implied Subjects: Vietnamese sentences often omit the subject, making dependency parsing difficult.

  “<subject???>Chơi free fire nó cứ out ra hoài chán”

2. Challenge in Vietnamese NLP

Label Complexity

- The task requires strict boundary detection. The span is correct only if it exactly matches the gold standard indices.

“<start>Camera của điện thoại này chụp xấu vl,<end> <start>bù lại pin trâu.<end>”

  - Gold: [0, 38, “CAMERA#NEGATIVE], [39, 54, “BATTERY#POSITIVE]
  - Pred: [0, 49, “CAMERA#NEGATIVE], [50, 54, “GENERAL#POSITIVE]

- Imbalance: A significant challenge is the dominance of "Positive" labels compared to "Neutral" or "Negative".

3. Dataset and Data Augmentation

Statistics

- Source: 11,122 smartphone feedback comments manually annotated.
- Total Spans: 35,396 annotated spans.
- Aspect:

SCREEN	CAMERA	FEATURES	BATTERY	PERFORMANCE
STORAGE	DESIGN	PRICE	GENERAL	SER&ACC

3. Dataset and Data Augmentation

Statistics

Review Length

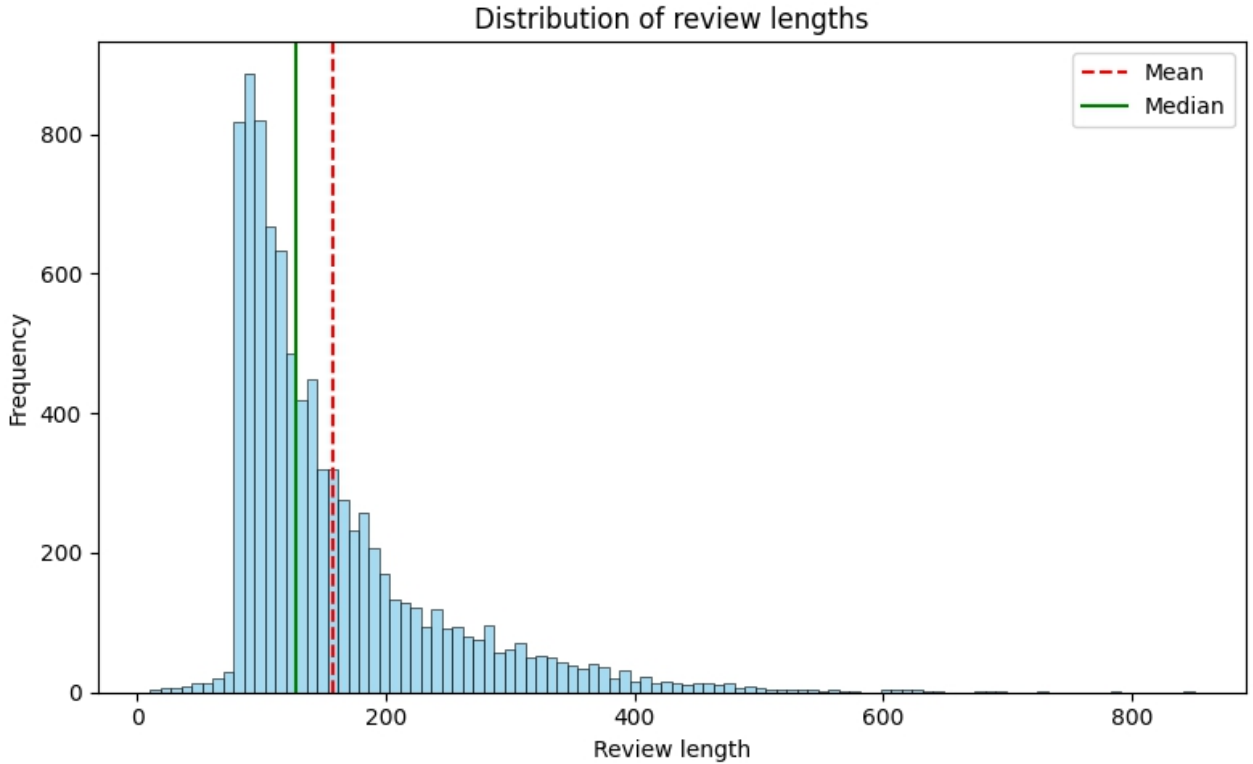


Figure 3.1. Review length distribution

3. Dataset and Data Augmentation

Statistics Review Length

- Short: CNN, GRU, ...
- Medium: LSTM, BiLSTM, ...
- Long: Transformer

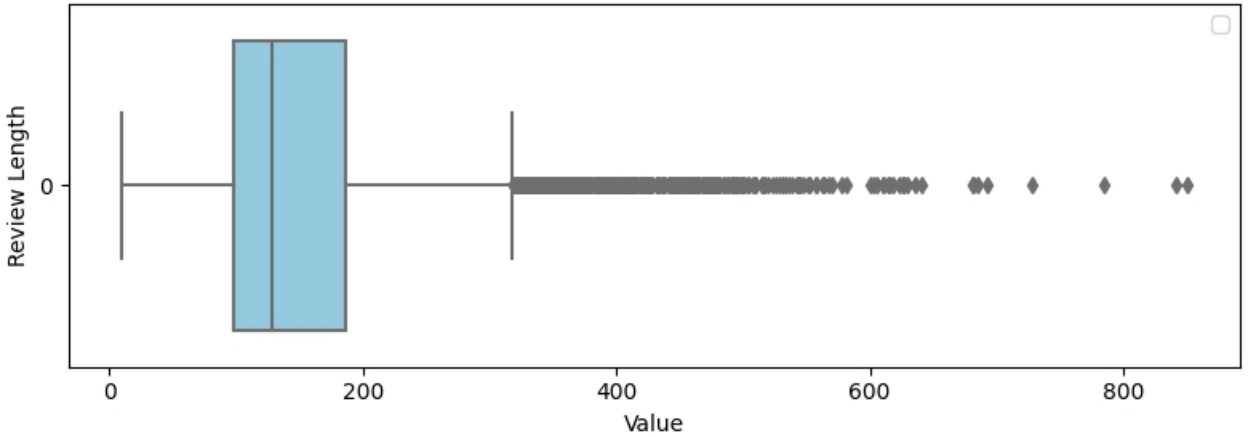


Figure 3.2. Boxplot of Review Length

3. Dataset and Data Augmentation

The Problem Minority Class

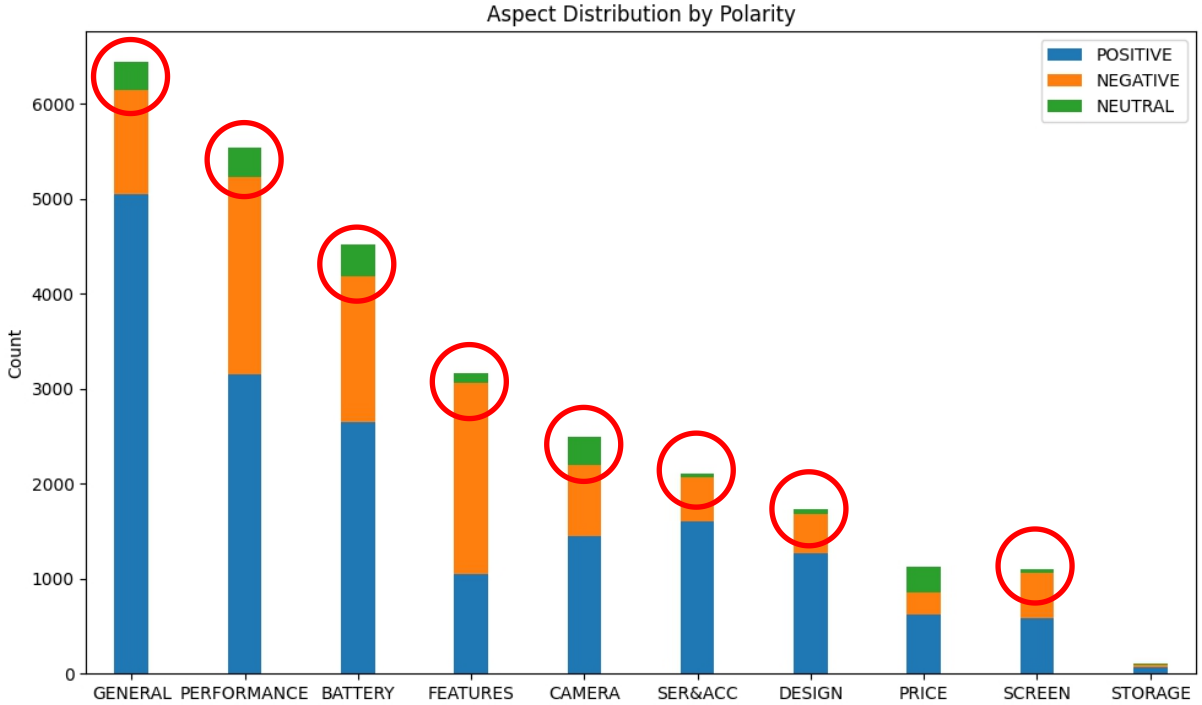


Figure 3.3. Review length distribution

3. Dataset and Data Augmentation

The Problem Minority Aspect

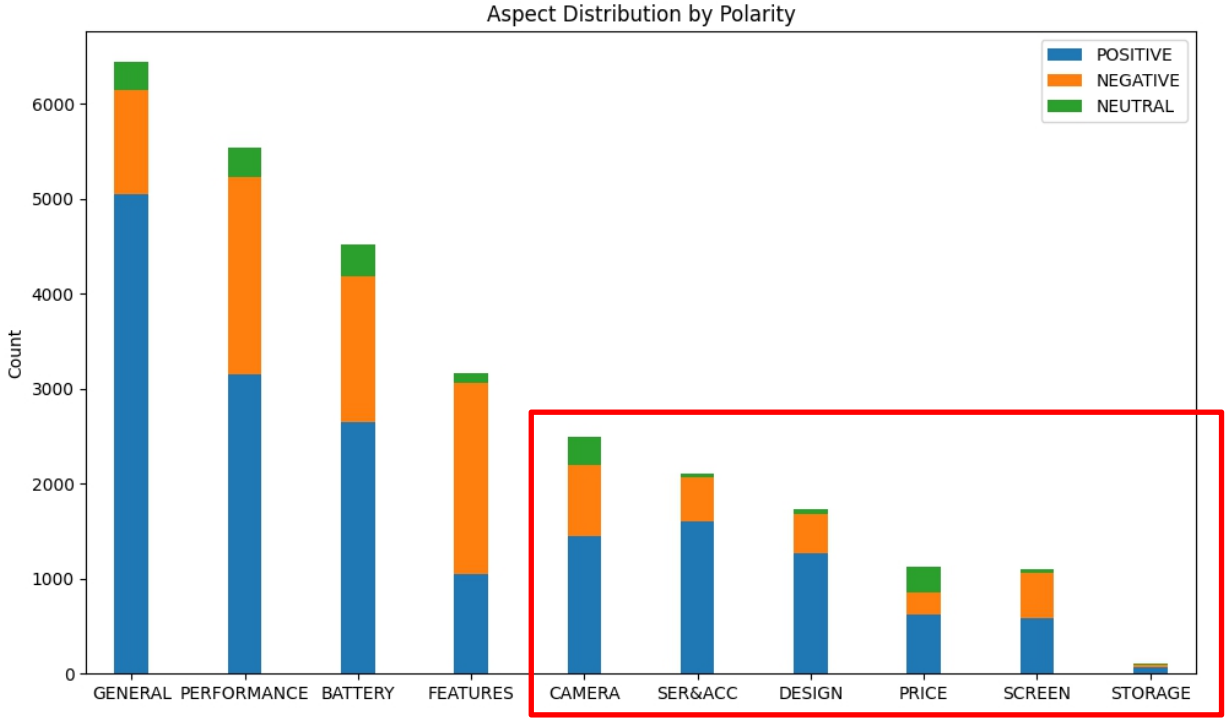


Figure 3.4. Review length distribution

3. Dataset and Data Augmentation

Weight Random Sampler

- Mechanism: Assigns a weight of 5.0 to samples containing minority aspects (STORAGE, PRICE, SER&ACC) to force the data loader to pick them more frequently.
- Goal: To artificially balance the batch distribution during training.

3. Dataset and Data Augmentation

Preprocessing

- Tokenization: Uses VnCoreNLP for word segmentation.
- Span Alignment: Map character-level indices to token-level IOB tags required for the CRF layer.
    - Input: “Camera của điện thoại này chụp xấu vl
    - Tokenized: [“Camera”, “của”, “điện thoại”, “này”, “chụp”, “xấu”, “vl”]
    - Output:
        - “Camera”: 0
        - “của”: 7
        - “điện thoại”: 11
        - ...

3. Dataset and Data Augmentation

IOB

- I - inside: Token inside an entity/span
- O - outside: Token no belongs to any entity/span
- B - beginning: Token start a new entity/span

- Review: "Máy có camera đẹp"
- Label: CAMERA#POSITIVE
  - Máy → O (not an aspect)
  - có → O
  - camera → B-CAMERA#POSITIVE (start span for camera)
  - đẹp → I-CAMERA#POSITIVE (inside camera span)

4. System Architecture

- Core Model: BiLSTM-CRF (Bidirectional LSTM with Conditional Random Field).
- Input: The Embedding Fusion concatenated by:
  1. Syllable Embeddings: Learned lookups for Vietnamese syllables.
  2. Character Embeddings (CharCNN-LSTM): Capture morphological features.
  3. Contextual Embeddings (XLM-R): providing deep contextual information
- Sequence Labeling:
  - The BiLSTM captures forward and backward context.
  - The CRF layer models the dependency between adjacent tags to output the valid tag sequence.

    e.g. I-CAMERA must follow B-CAMERA

4. System Architecture

- Syllable Embeddings
  - Máy → [0.12, -0.05, 0.33, …]
  - camera → [0.45, 0.10, -0.22, …]
- Character Embeddings
  - Token: camera → LSTM over [c, a, m, e, r, a] → [0.21, -0.11, 0.35, …]
  - Token: đẹp → LSTM over [đ, ẹ, p] → [0.05, 0.40, -0.12, …]
- Contextual Embeddings
  - Câu: "Máy có camera đẹp"
    - camera → [0.12, -0.08, 0.25, …]
  - Câu khác: "Camera của máy này rất xịn"
    - camera → [0.30, 0.05, -0.15, …]

4. System Architecture

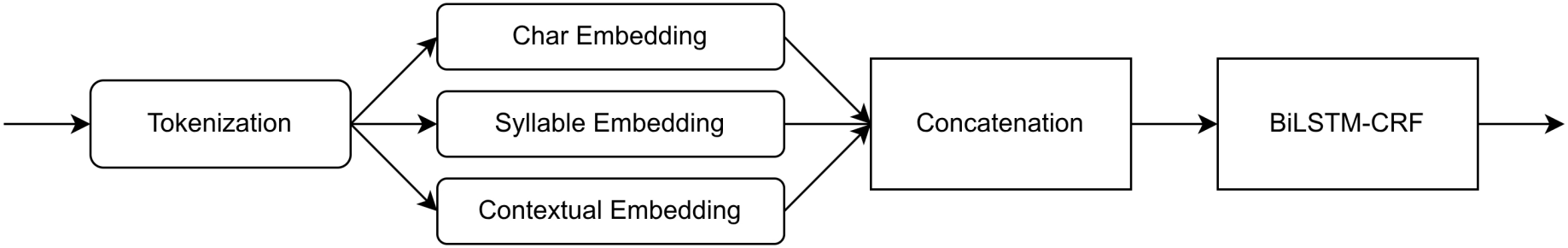


Figure 4.1. Architecture Overview

5. Experimental Setup & Results

Hyperparemeters

- CONTEXT_DIM = 1024
- SYLLABLE_EMB_DIM = 100
- CHAR_EMB_DIM = 100
- CHAR_HIDDEN_DIM = 50
- LSTM_HIDDEN_DIM = 400 # 400 in paper
- LSTM_DROPOUT = 0.33

- KFOLD_SPLIT = 5
- BATCH_SIZE = 16 # 5000 in paper
- EPOCHS = 30
- LR = 1e-4
- PATIENCE = 5

5. Experimental Setup & Results

Baseline Results

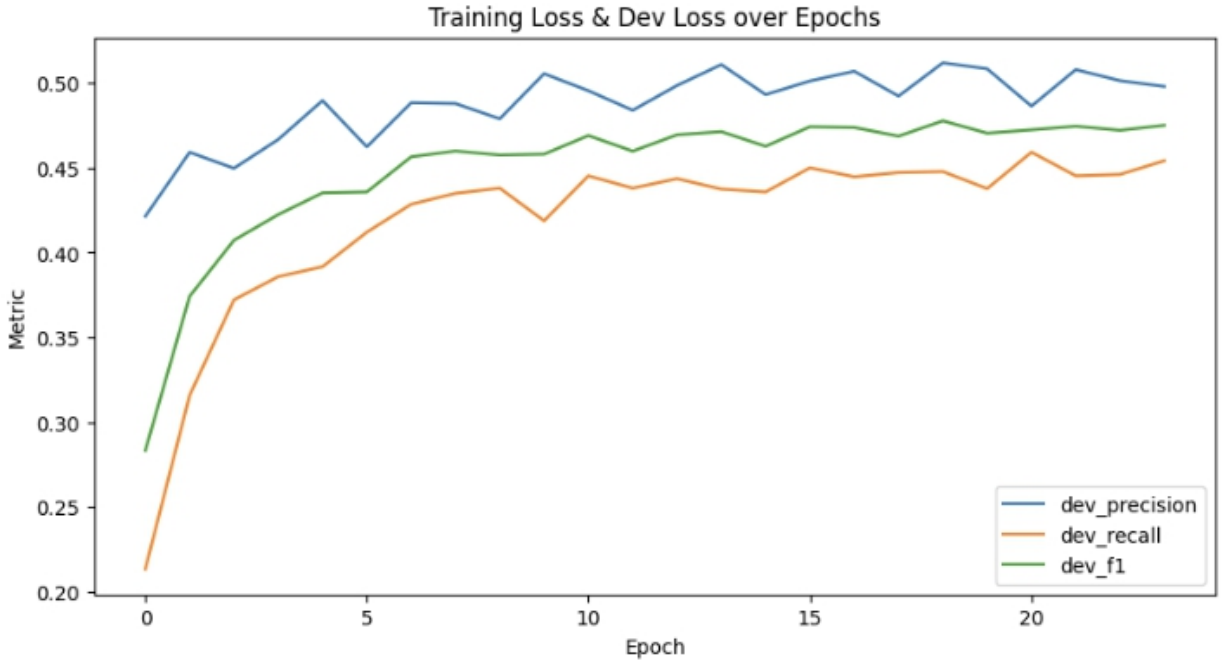


Figure 5.1. Evaluaiton on Dev (Eval) Dataset

5. Experimental Setup & Results

Baseline Results

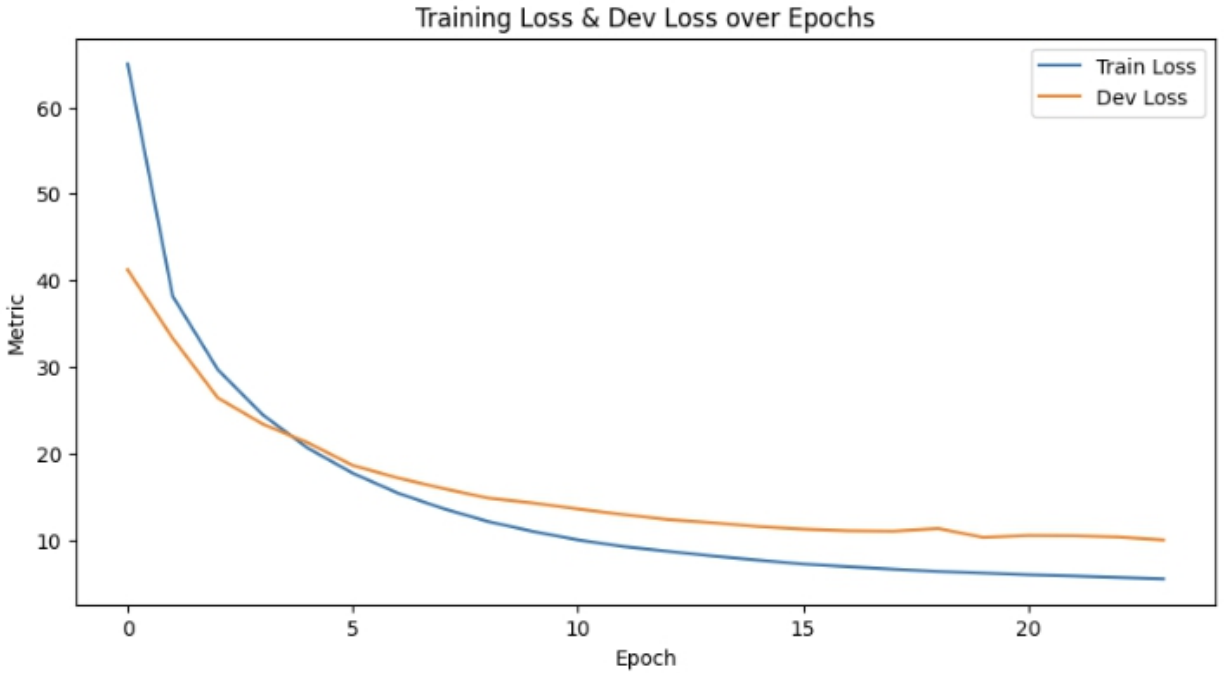


Figure 5.2. Evaluaiton on Dev (Eval) Dataset

5. Experimental Setup & Results

Baseline Results

Task	F1_macro	F1_macro (origin paper)
aspect	no exp	62.76%
polarity	no exp	49.77%
aspect#polarity	50.53%	45.70%

Difference Analysis

1. Impact of Batch Size on Model Training
2. Dataset Characteristics
3. Learning Dynamics
4. Empirical Evidence (from UIT-ViSD4SA paper)

5. Experimental Setup & Results

Difference Analysis

1. Impact of Batch Size on Model Training
2. Dataset Characteristics
3. Learning Dynamics
4. Empirical Evidence (from UIT-ViSD4SA paper)

6. Future Workds & Improvement

- Explore Machine Reading Comprehension (MRC) frameworks to treat span detection as a Question-Answering task.
- Utilize larger multilingual pre-trained models to handle loanwords better.
- Fine-tuning XLM-R: The current baseline freezes XLM-R; fully fine-tuning it could yield better representations.
- Advanced Sampling: Tune the WeightedRandomSampler weights further or try Focal Loss to penalize easy examples (Positive class).

Question & Answer

References

- PyTorch implementations of GANs
- A list of all named GANs
- Vanilla GANs paper
- WGAN paper
- Fréchet Inception Distance
- A mix of GAN implementations including progressive growing
- GAN-play
- Style GAN