Review-based Opinion Mining & Summarization for Products

Akshat Gupta (2021515) Arjit Singh Arora (2021452)
Swati Sharma (2021568) Kumar Aryan Singh (2021468)

ABSTRACT

Web browsing has evolved into more than just consumption, with online shopping being a significant aspect. Users contribute feedback, accumulating a wealth of reviews, potentially containing accurate information. This study focuses on a framework using advanced natural language processing to identify fake reviews and generate precise product descriptions. By leveraging product ratings, review texts, and sellers' descriptions, it aims to improve user decision-making and engagement. Utilizing an Encoder-only architecture with Self and Cross Attention, fake reviews are distinguished from real ones. An alternative product description is generated based on genuine reviews, comparing it with the seller's description to assess accuracy and user experience.

DATASET

1. **Amazon Review Dataset (Reviews without Real/Fake Label):**

- Description: Subset of 5000 samples used; Contains Product metadata (Title, Description, Category, Image, etc) and Review Details (ReviewerID, Review Text, Review Title, Rating, etc)
- Data Sample:

```
{
  "reviewerID": "AUI6WTTT0QZYS",
  "asin": "5120053084",
  "reviewerName": "Abbey",
  "vote": "2",
  "overall": 5.0,
  "reviewText": "I now have 4 of the 5
  available,
  colors of this shirt...",
  "summary": "Comfy, flattering,
  discreet--highly,
  recommended!",
  "reviewTime": "01 1, 2018",
  "verified": true,
  "unixReviewTime": 1514764800,
  "image": ,
  ["https://images-na.ssl-images-amazon.com/
    images/I/71eG75FTJJL._SY88.jpg"],
  "style": {
    "Size:": "Large",
    "Color:": "Charcoal"
  }
}
```

2. **Amazon Reviews Dataset (Real/Fake Label Version):**

- Description: Approximately 21000 samples; Contains Product Metadata (Title, Category), Review Data (Title, Rating, Text), and Real/Fake Labels
- Data Sample:

```
{
  "LABEL": "label1",
  "RATING"": "4",
  "VERIFIED PURCHASE": "N",
  "PRODUCT CATEGORY": "PC",
  "PRODUCT ID": "B00008NG7N",
  "PRODUCT TITLE": "Targus PAUK10U Ultra
  Mini USB Keypad, Black",
  "REVIEW TITLE": "useful",
  "REVIEW TEXT": "When least you think
  so, this product will save the day.
  Just keep it around just in case you
  need it for something."
 }
}
```

METHODOLOGY

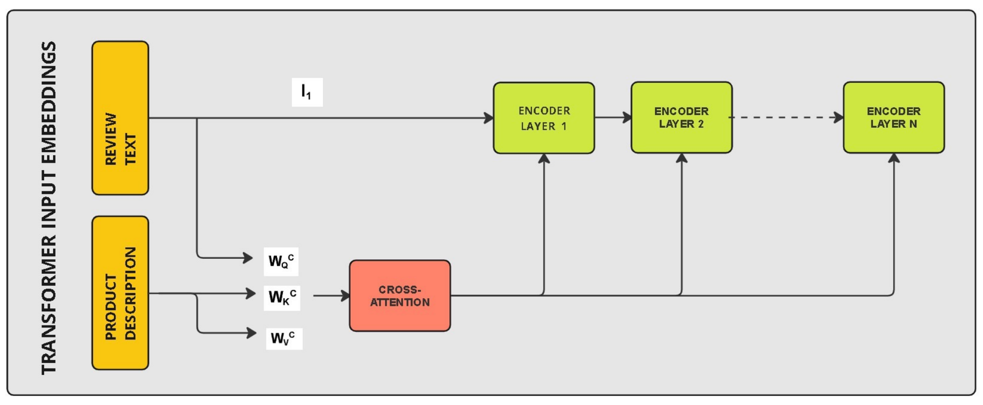
TASK - 1: FAKE REVIEW DETECTION:

- Data Preparation and Preprocessing:

Data is pre-processed by concatenating columns, expanding contractions, removing stop words, and padding/truncating text. Review texts are capped at 512 characters, product descriptions at 82. Tokens like [CLS] and [EOS] are added for classification tasks. Fasttext embeddings capture sentence-level context, while positional encodings maintain sequential nature of text for the Encoder-only Transformer.

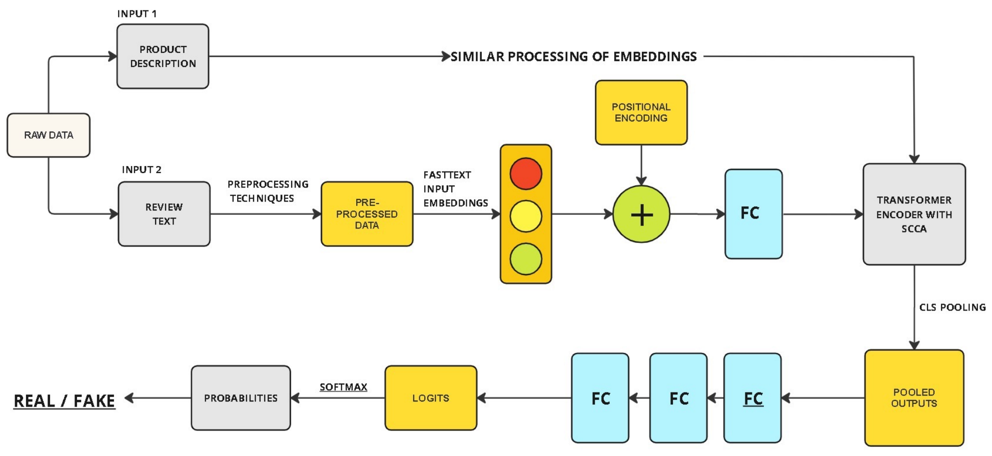
- SCCA - Self & Cross-Categorical Attention:

The model implements custom attention with Self and Cross Attention. Self Attention captures intra-dependencies within the review text, while Cross Attention evaluates inter-dependencies between reviews and product descriptions. Multi-Head SCCA mechanism in each encoder layer computes both self and cross attention. This process ensures relevance between reviews and products, aiding in recognizing their correlation.



- Encoder-only Transformer:

It comprises of multiple Encoder layers with SCCA. Each layer takes input from the previous layer along with computed cross-attention values and attention masks. The final layer extracts the [CLS] token output for sequence classification. Multiple feed-forward layers reduce dimensionality to 2 for logits, with softmax activation for probabilities. Training employs Cross Entropy Loss with the Adam optimizer.



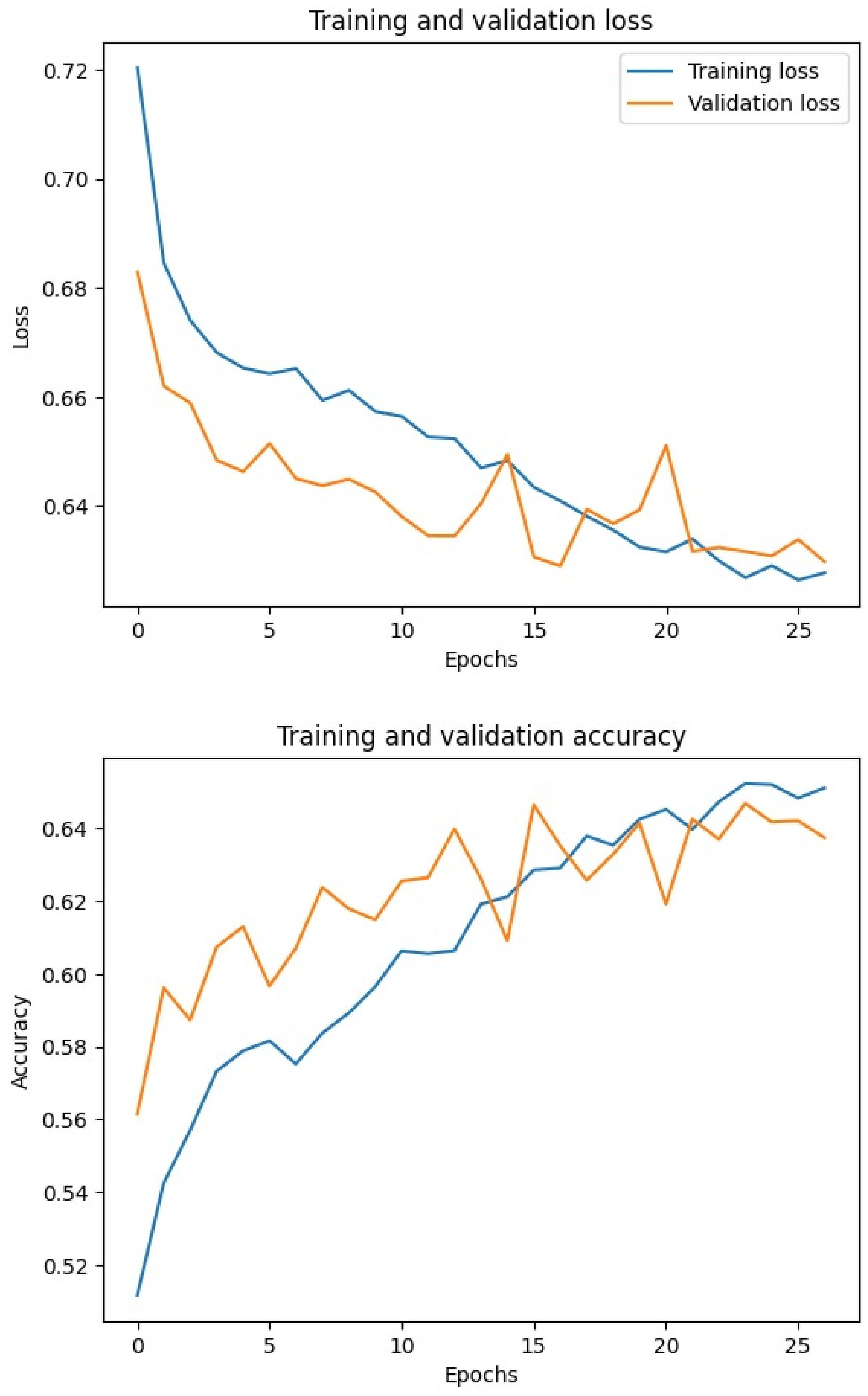
TASK - 2: OPINION MINING THROUGH TEXT SUMMARISATION:

This task generates product description summaries from identified real reviews and evaluates seller descriptions using Semantic Text Similarity. It utilizes a fine-tuned version of BART for the review summarization. Sentence similarity is measured using cosine similarity with sentences converted to SentenceBERT Embeddings. A Human Evaluation (of randomly sampled elements from the generated summaries) against the product descriptions confirms its capability to summarize reviews effectively.

RESULTS

TASK - 1:

- The Fake Review Classifier achieves a confidence level of 75% in identifying fake reviews based on the context provided by the product description.
- Training Accuracy → 75.98%
- Validation Accuracy → 75.63%
- Loss & Accuracy v/s Epoch Trends:



TASK - 2:

- Generated Summary: “Joaquin Phoenix magical role If like The Big Lebowski love this”
- Product Description: “In Los Angeles at the turn of the 1970s, drug-fueled detective Larry...”
- Semantic Similarity: 0.1493

This is an example of a movie; as can be clearly seen, the description is not relevant to the reviews which produces a low similarity score, but the generated summary shows reviews were mostly good about the product.

Future Objectives

- Enhance fake review detection and product description generation.
- Experiment with diverse datasets and multi-modal approaches.
- Incorporate active learning to address emerging challenges.

References

1. A. Vaswani et al., “Attention is All You Need,” in Advances in Neural Information Processing Systems 30 (NIPS 2017), NeurIPS Proceedings.
2. P. Novgorodov and I. Guy, ”Generating Product Descriptions from User Reviews,” in Proc. 27th ACM Int. Conf. Inf. Knowl. Manag., 2019, pp. 2455–2458
3. L. Dong and S. Huang, ”Mining of Product Reviews at aspect level,” in Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguist.: Volume 2, Short Papers, 2017, pp. 560–566. 46, no. 1, pp. 155–184, 2016