NLP: Solutions for Real-World Problems

COMP8460: ARTIFICIAL INTELLIGENCE FOR NATURAL LANGUAGE PROCESSING APPLICATIONS

Group F

Use Case 6: E-commerce Intelligent System

Presented by:
Name: Sadat Hossain, ID: 48738085
Name: Muxin Qiao, ID: 48983888
Name: Niharika Puthalapattu, ID: 48942642

Use Case: E-commerce Intelligent System

GOAL OF THE SYSTEM

Goal:
To build an intelligent E-commerce review analysis system that can:

- Extract key product aspects
- Determine sentiment expressed towards each aspect
- Analyze and summarize customer reviews at scale
- Detect fake or unreliable reviews to improve trust
- Generate concise product descriptions automatically
- Provide actionable business insights for product improvement and marketing decisions

Speaker: Niharika Puthalapattu

Dataset

MACQUARIE University SYDNEY·AUSTRALIA

AMAZON PRODUCT REVIEWS (KAGGLE)

- **Source:** Kaggle – Consumer reviews of Amazon products
- Combined 3 CSV files into one dataset
- Total reviews before cleaning: 67,992
- Dataset included around 15–20 attributes
- **Key fields used:** `product_name`, `review_text`, `rating`, `review_date`

Speaker: Niharika Puthalapattu

Data processing

MACQUARIE University SYDNEY·AUSTRALIA

DATA CLEANING AND PRE-PROCESSING WORKFLOW

1. **Data Cleaning**
   - Removed duplicates and missing entries
   - Standardized key columns (`product_name`, `review_text`, `rating`, `review_date`)
   - Corrected inconsistent text formatting and casing
   - Final cleaned dataset size: ***64,037*** **reviews**

2. **Text Pre-processing**
   - Removed punctuation, numbers, and stop-words
   - Converted all text to lowercase
   - Tokenized and lemmatized using ***spaCy*** *`(en_core_web_sm)`*

3. **Feature Preparation**
   - Encoded sentiment labels: ***0 = Negative, 1 = Positive, 2 = Neutral***
   - Label distribution: Positive: ***91.9%,*** Neutral: ***4.3%,*** Negative: ***3.8%***

Speaker: Niharika Puthalapattu

Basic Technique: Product Aspect Extraction (NER)

SPACY CUSTOM NAMED ENTITY RECOGNITION

Goal:

Identify key product features *(e.g., battery, display, sound, price)* mentioned in customer reviews.

Method:

- Created aspect keyword patterns using regex
- Trained a custom **spaCy NER** model using these aspect labels
- Evaluated performance using **Precision, Recall, and F1-Score**

Process Steps:

1. Loaded the preprocessed review dataset
2. Defined aspect terms and patterns (battery, display, sound, price, quality)
3. Trained and fine-tuned spaCy NER model
4. Visualized extracted aspects using `spacy.displacy.render()`
5. Saved final model for downstream sentiment and summarization tasks

Results:

- **F1-Score:** 0.75 – 0.85 (varies by aspect)
- Successfully extracts the most relevant product attributes from user reviews

Speaker: Sadat Hossain

Basic Technique: Sentiment Classification

MACQUARIE University SYDNEY · AUSTRALIA

TF-IDF FEATURES + LOGISTIC REGRESSION

Goal:

Classify each review as **Positive, Negative, or Neutral**.

Method:

- Converted review text into TF-IDF weighted features (unigrams + bigrams)
- Trained a **Logistic Regression** classifier with class balancing
- Evaluated using **Accuracy, Precision, Recall, and Confusion Matrix**

Process Steps:

1. Vectorized cleaned text using TF-IDF
2. Split dataset into train and test sets
3. Trained Logistic Regression model
4. Identified top predictive words for each sentiment class
5. Saved the trained model and vectorizer for reuse

Results:

- **Accuracy:** 0.88 – 0.92 on test set
- Model highlights sentiment-defining terms (e.g., *great*, *poor*, *expensive*)

Speaker: Sadat Hossain

Advanced Techniques : RAG + LLM-Based Summarization

SECTION 1: RAG PIPELINE (RETRIEVAL-AUGMENTED GENERATION)

Goal:
Retrieve the most relevant reviews to answer product-related queries.

Method:
- Encoded reviews using ***SentenceTransformer*** embeddings
- Built a **FAISS** index for fast similarity search
- Retrieved **top-K** semantically similar reviews for each query
- Injected retrieved text into the **LLM prompt** to ground responses in real data

Speaker: Muxin Qiao

Advanced Techniques : RAG + LLM-Based Summarization

SECTION 2: LLM SUMMARIZATION & EVALUATION

Approach:

- Used **few-shot prompting** with structured templates
- Generated **aspect-wise summaries** (e.g., Battery, Screen, Price) with sentiment
- Model used: ***Microsoft Phi-2***

Output:

- Final summaries and product descriptions **with cited review snippets**

Quality Check:

- Evaluated summaries using ***ROUGE-L*** to measure factual consistency and improvement

Results:

- Rough-L=0.288

Speaker: Muxin Qiao

Thank You