Sentiment Analysis Report — Capstone Project (NLP Applications)

Author: Ntokozo Mazibuko

Introduction

This report summarises a lightweight sentiment analysis pipeline applied to an Amazon product-reviews dataset. Using spaCy with the spaCyTextBlob extension, each review is scored for polarity and subjectivity and then mapped to a simple three-class label (positive, neutral, negative). The goal is to demonstrate a clear, reproducible workflow suitable for a graded environment, highlight the data characteristics, and reflect on practical strengths and limitations of this approach.

Dataset description

- **Source file:**/Users/admin/Documents/HyperionDev/Datafiniti_Amazon_Consumer_Reviews_of_Amazon_Products_May19.csv
- **Target column:** reviews.text (free-text product review content)
- **Records analysed (non-null): 28,332**

Notes:

- The dataset contains consumer reviews for Amazon products; only the review text was used as input.
- There are no human sentiment labels, so the analysis uses a lexicon-style method (spaCyTextBlob).

Preprocessing steps

- Lowercase and trim whitespace.
- Tokenise with spaCy.
- Lemmatise when available; otherwise keep the surface form.
- Keep alphabetic tokens only and remove stopwords using spaCy's vocabulary.
- Score each review with spaCyTextBlob for **polarity** [−1,1] and **subjectivity** [0,1].
- Map polarity to labels with simple thresholds:
    - **polarity ≥ +0.10 → "positive"**
    - **polarity ≤ −0.10 → "negative"**
    - otherwise **"neutral."**

Evaluation of results

Label distribution

- **Positive:** 22,415 (**79.1%**)
- **Neutral:** 4,810 (**17.0%**)

- **Negative:** 1,107 (**3.9%**)

Polarity summary (spaCyTextBlob)

- count: 28,332
- mean: 0.3643
- std: 0.3038
- min: −1.0000
- 25%: 0.1417
- 50% (median): 0.3500
- 75%: 0.5708
- max: 1.0000

Interpretation

- Reviews are strongly skewed positive (≈80%), which is typical for public product feedback.
- Median polarity (-0.35) with moderate spread (std -0.30) indicates varied sentiment strength.
- With no ground-truth labels, evaluation is descriptive (counts/distributions) rather than accuracy/F1.

Strengths and limitations

Strengths

- Simple, explainable pipeline that is easy to reproduce and assess.
- Fast on large corpora and does not require labelled data.
- Transparent thresholds that can be tuned per dataset.

Limitations

- Lexicon methods can miss nuance (sarcasm, complex negation, domain-specific language).
- Positive class imbalance means few negatives under fixed thresholds.
- No quantitative validation without labels; a small human-annotated subset would help.
- Depends on spaCyTextBlob rules; supervised transformer models often perform better when labels are available.

Suggested next steps (optional)

- Tune positive/negative thresholds on a small labelled dev set to target precision/recall.