Sentiment Analysis in Amazon Product Reviews

A Comprehensive Exploration of Machine Learning Models and Techniques

Task and Dataset

Task

- Sentiment Analysis on Amazon product reviews
- popular NLP task, useful for business intelligence
- multi-class classification: positive / negative / neutral

Dataset

- "Amazon Reviews: Unlocked Mobile Phones"
- over 400,000 reviews of nearly 4,400 distinct mobile phone models

Brand Name
Product Name
Price
Rating
Review
Review Votes

Project Overview

Models

1. Naive Bayes
2. Logistic Regression
3. SVM
4. biLSTM
5. DistilBERT

Techniques

1. 10-fold cross-validation
2. Hyperparameter tuning
3. Ensemble Learning
4. Local Explanations (LIME)
5. GloVe embeddings
6. Dimensionality Reduction (NMF)
7. Data balancing (SMOTE)
8. Feature engineering and selection
9. Friedman test

Related work

Ezhilarasan et al. (2019)

Sentiment Analysis On Product Reviews: A Survey

Almuayqil et al. (2022)

Sentiment Analysis on Tweet reviews with DistilBERT

Ni and Cao (2020)

Sentiment Analysis based on GloVe and LSTM

Kuhn and Johnson (2019)

Feature Engineering and Selection: A Practical Approach for Predictive Models

Zhou and Long (2018)
Hanane et al. (2021)
Arbane et al. (2023)

Experimental Design: baseline models

1. Naive Bayes, Logistic Regression, SVM: 10-fold CV and hypertuning
2. Dimensionality Reduction, Synthetic Minority Over-sampling
3. Feature extraction, engineering and selection
4. Ensemble Learning, Local Interpretable Model-agnostic Explanations

Experimental Design: state-of-the-art models

5

biLSTM

with pre-trained GloVe embeddings

6

biLSTM

without GloVe embeddings (baseline)

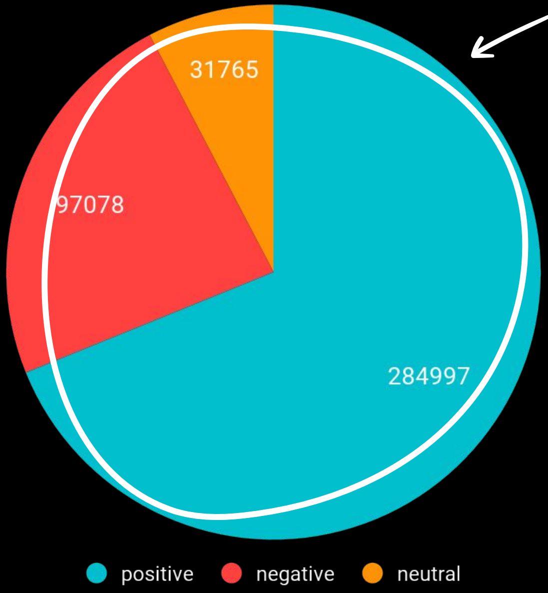
7

DistilBERT

smaller, faster, cheaper version of BERT

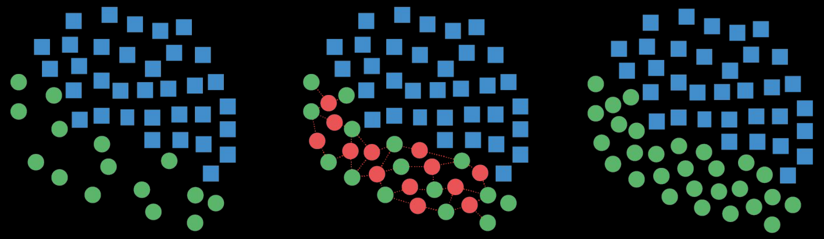
Synthetic Minority Over-sampling

Significant data imbalance!



SMOTE creates synthetic examples of the minority class

1. selects **random** instances of the minority class
2. identifies their k-nearest **neighbors**
3. generates **synthetic** instances along connecting line segments



Local Interpretable Explanations (LIME)

★★★☆☆ **Review**

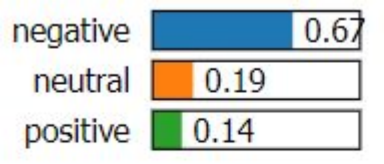
Verified Purchase

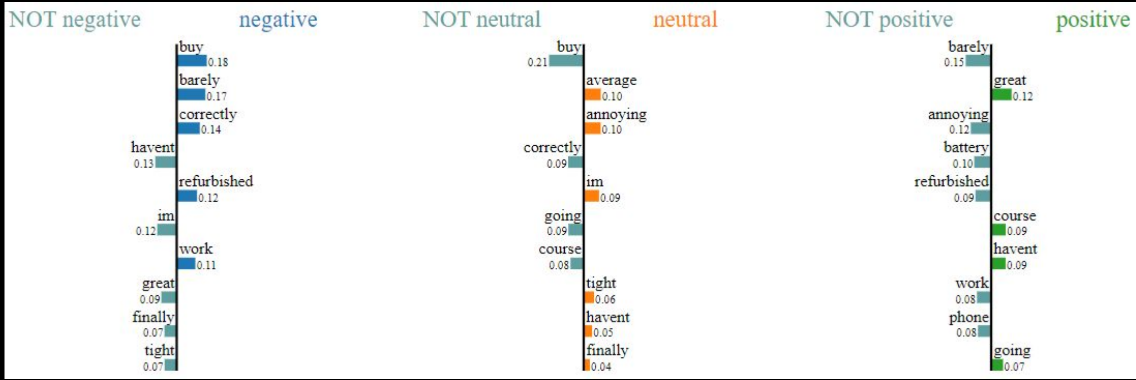
Well first off I was sent the wrong phone which put me in a tight.. Once I finally received the correct phone it seemed to work like a dream . It had pretty much all the memory to be refurbished which was great. Then I noticed that the battery was no good.. it could barely last over an hour with average use so now I'm going to have to buy a battery. Then (of course I haven't dropped it) the screen is not in its frame.. it's popping up and so I'm going to have to pay for someone to put it in correctly .. this is very hectic and annoying.

2 people found this helpful

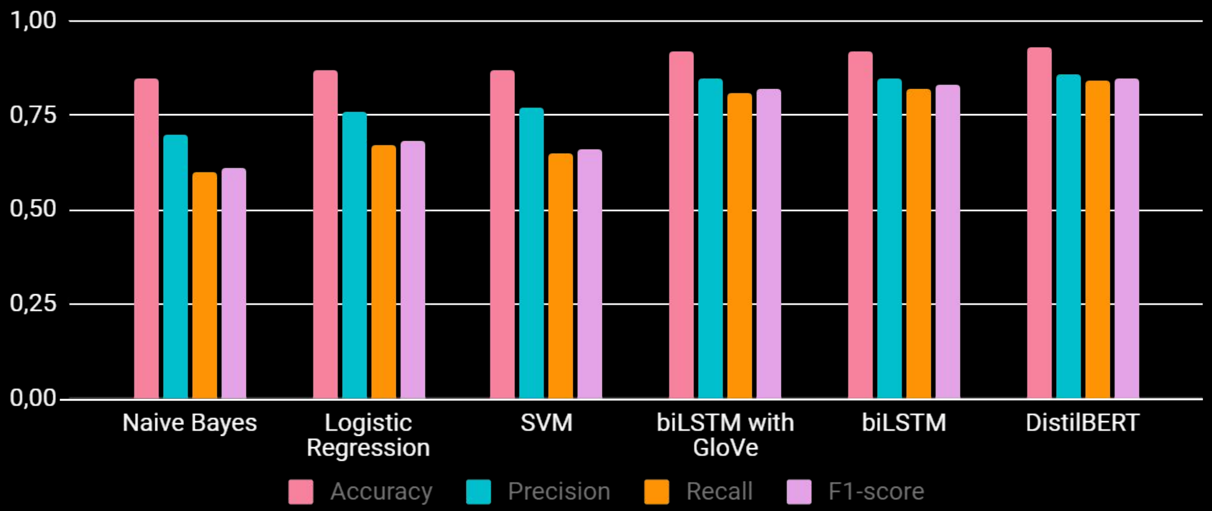
Helpful | Report

well first sent wrong phone put tight finally received correct phone seemed work like dream pretty much memory refurbished great noticed battery good could barely last hour average use im going buy battery course havent dropped screen frame popping im going pay someone put correctly hectic annoying





Experimental Results

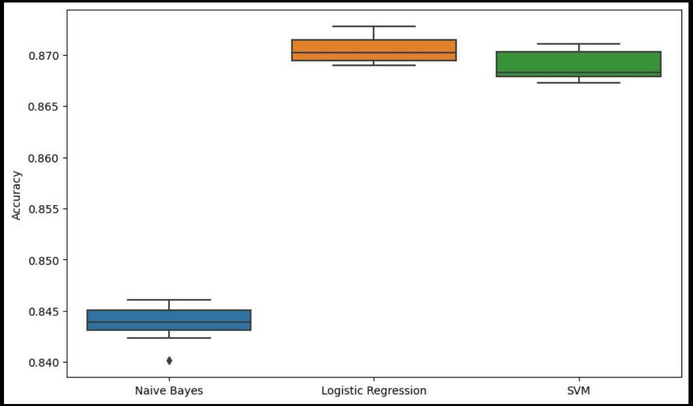


	Accuracy	Precision	Recall	F1-score
Naive Bayes	0.85	0.7	0.6	0.61
Logistic Regression	0.87	0.76	0.67	0.68
SVM	0.87	0.77	0.65	0.66

	Accuracy	Precision	Recall	F1-score
biLSTM + GloVe	0.92	0.85	0.81	0.82
biLSTM	0.92	0.85	0.82	0.83
DistilBERT	0.93	0.86	0.84	0.85

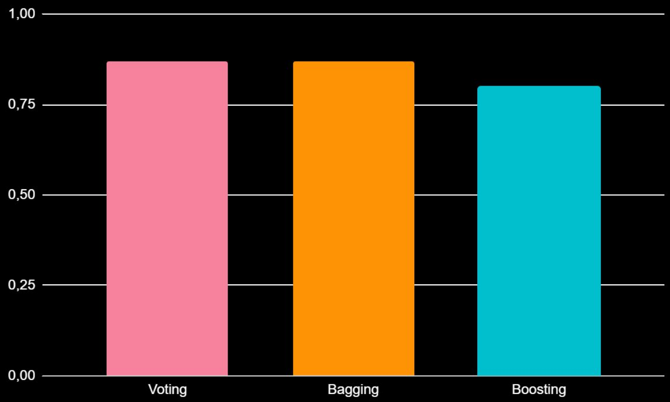
Experimental Results

Friedman Test



F statistic	20
p-value	4.54e-05

Ensemble Methods



	Accuracy
Voting Classifier	0.87
Bagging Classifier	0.87
Boosting Classifier	0.80

Key Insights

01

State-of-the-art models (biLSTMs and DistilBERT) demonstrated the highest performance, but simpler models also performed well.

02

Performance of biLSTM models with and without GloVe embeddings was unexpectedly similar, indicating a potential mismatch between GloVe and dataset language.

03

Employing techniques like feature engineering, dimensionality reduction, and SMOTE worsened performance, possibly due to dataset characteristics.

04

Friedman test confirmed statistically significant performance differences among Naive Bayes, Logistic Regression, and SVM models.

05

DistilBERT's performance was comparable to biLSTM, possibly due to task nature or suboptimal model exploitation.

Future work

Hyperparameter tuning for biLSTM and DistilBERT, different pre-trained word embeddings

Different preprocessing, data augmentation, dimensionality reduction and feature engineering techniques

Exploring other transformer-based models, ensemble methods, interpretation techniques

Conclusion

- The project conducted a comprehensive exploration of models and techniques for sentiment analysis on Amazon product reviews.
- Simple and SOTA models showed robust performance, with biLSTM and DistilBERT models performing the best.
- Pre-trained word embeddings may not always enhance performance, and advanced architectures may not always outperform simpler ones in sentiment analysis tasks.
- Ensemble methods showed potential for improving performance through model combination, although they did not exceed the performance of the best individual models.