DALHOUSIE UNIVERSITY

Fake Review Classifier

Hirva Dhandhukia (B00987633)

Motivation

Online product reviews strongly influence customer purchasing decisions. However, fake or AI-generated reviews can mislead buyers and weaken trust in a brand.

In my project, I built a deep learning-based model to detect fake reviews, enabling organizations to monitor and maintain the credibility of their product feedback. This helps businesses protect their reputation and ensure the quality standards they promise to customers.

Problem of Study

Detecting fake reviews based only on the review text

- Fake reviews can be well-written to mimic genuine ones.
- Labeling isn't inherently reliable - my dataset has "genuine" and "fake" labels made using rules, but some reviews might still be unclear.

Related Work

Dataset: Kaggle Fake reviews dataset

Size: 40k reviews

Labels: Genuine (CG) / Fake (OR) - generated heuristically

```
Columns: ['category', 'rating', 'label', 'text_']

Sample rows:
```

	category	rating	label	text_
33150	Toys_and_Games_5	5.0	OR	special Christmas gift never seen these lego b...
11072	Electronics_5	4.0	CG	Got to be happy with this lens, as it is a len...
34140	Toys_and_Games_5	5.0	CG	Happy with purchase, currently using this set ...
23474	Pet_Supplies_5	5.0	CG	I chose this food because it has the quality i...
3875	Home_and_Kitchen_5	5.0	OR	My sons room is done in an alligator theme, an...

```
Dataset shape: (40432, 4)
```

Mexwell, “fake-reviews-dataset,” *Kaggle*, Dataset, Aug. 2025. [Online]. Available: kaggle.com/datasets/mexwell/fake-reviews-dataset.

Approach

Data Cleaning and Encoding Labels: 0:1

Tokenization: Used *bert-base-uncased* tokenizer

Splitting (Train-Val-Test): 80-10-10

Model: *Bert* with Sequence Classification of 2 labels

Training loop args with batch-size, learning rate, early stopping, and best model loading.

[500/8088 59:38 < 15:08:44, 0.14 it/s, Epoch 0/4]

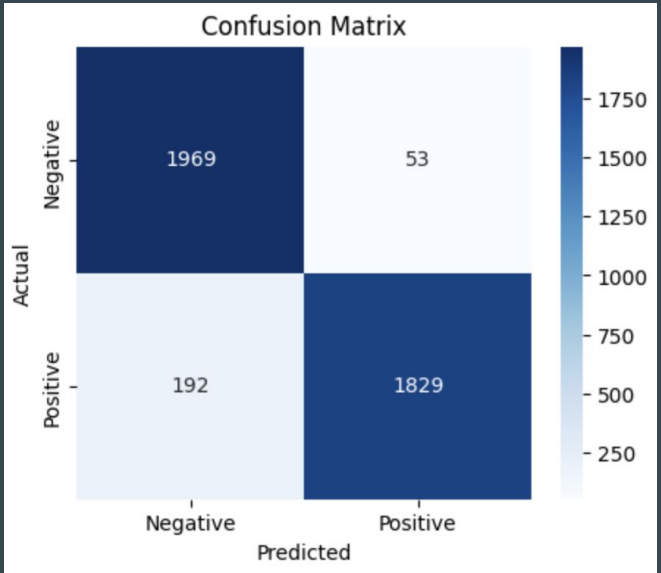
Step	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
100	0.443300	0.235849	0.910710	0.900579	0.923305	0.911801
200	0.215500	0.192481	0.919119	0.954399	0.880257	0.915830
300	0.209900	0.153802	0.939401	0.971838	0.904998	0.937228
400	0.178100	0.307237	0.917883	0.983400	0.850074	0.911890
500	0.210400	0.243006	0.920356	0.990757	0.848590	0.914179

Evaluation Results

Accuracy: 94%

Class-wise Performance:
Original: Precision 0.91, Recall 0.97, F1 = 0.94
Fake: Precision 0.97, Recall 0.90, F1 = 0.94

Understanding: more likely to miss a fake review (false negative) than to incorrectly flag an original review as fake (false positive).



Limitations & Error Analysis

LIME to explain predictions

"which words influenced the model's decision the most"

Misclassifications happened with

Very short reviews
Neutral, generic wording
Mixed sentiment reviews

```
[Sample 24] True: Genuine Review, Predicted: Fake Review
super high quality
just what i expected to receive from ck
the gas lens works great
[('the', -0.4757581559434148), ('great', -0.15876293637867295), ('quality', -0.09505542417808734), ('to', -0.087402
```

Turn this into a Product ?

1. Browser extension for customers to classify online reviews.

2. Companies integrating it in their workflow to analyze manipulated reviews of products.

Q/A