Predicting Movie Review Sentiment

Group 4: Data Vis Biz
Cecelia Walker (Leader), Jeremy Grossman, Blake Richardson
DS 4002
16 February 2026

Outline

01 Question/Motivation

02 Data Acquisition

03 Analysis Plan

04 Analysis Decision

05 Bias & Uncertainty

06 Results & Conclusion/ Next Steps

Research Question

Can we determine which textual evidence cues are most strongly predictive of a movie review's sentiment (positive v. negative) and then use those determinations to accurately predict movie review sentiment?

Motivation

Hypothesis: Textual patterns and semantic meaning within movie reviews can accurately predict whether a review is positive or negative.

Why?

- Online movie reviews strongly influence a film's success and future projects
- The volume of reviews makes manual analysis impractical
- Automated sentiment analysis enables large-scale understanding of audience reception
- Could be applied in other contexts such as restaurant reviews

Modeling Approach:

(1) TF-IDF – identifies important words by their lexical meaning
(2) BERT language model → captures context and semantic meaning
(3) Classification models → predict sentiment

Data Acquisition

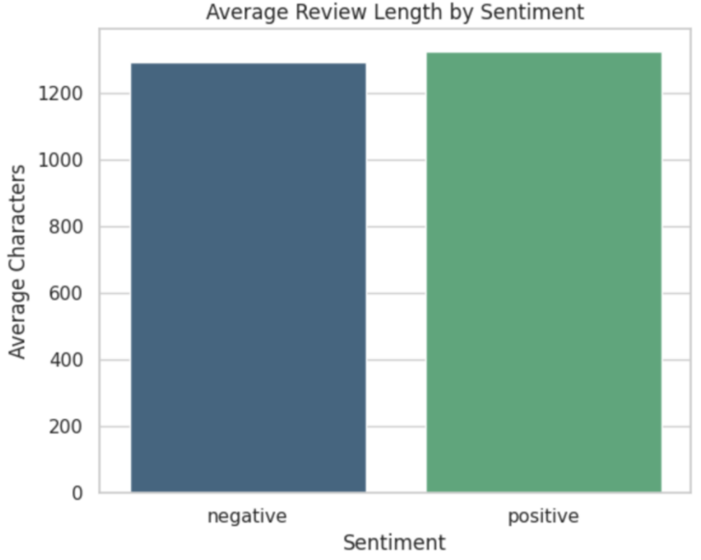
- Data from **Kaggle**: Text analytics drawn from 50,000 movie reviews on the International Movie Database (IMDB)
- 50, 000 movie reviews
- Balanced classes: 25,000 positive and 25,000 negative

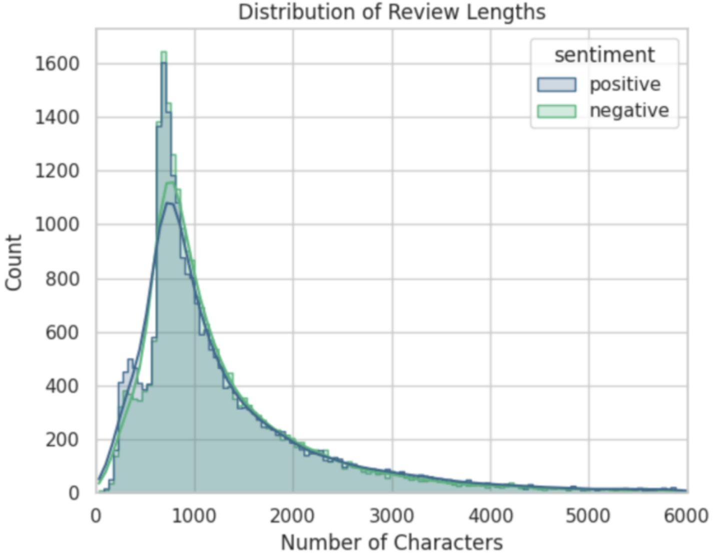
Review: raw text ⇨ Sentiment: positive/negative

	review
sentiment	
negative	25000
positive	25000

	review	sentiment
0	One of the other reviewers has mentioned that ...	positive
1	A wonderful little production. The...	positive
2	I thought this was a wonderful way to spend ti...	positive
3	Basically there's a family where a little boy ...	negative
4	Petter Mattei's "Love in the Time of Money" is...	positive
...	...	...
49995	I thought this movie did a down right good job...	positive
49996	Bad plot, bad dialogue, bad acting, idiotic di...	negative
49997	I am a Catholic taught in parochial elementary...	negative
49998	I'm going to have to disagree with the previou...	negative
49999	No one expects the Star Trek movies to be high...	negative

50000 rows × 2 columns





Analysis Plan

1) TF-IDF Vectorization

- Vocabulary of 20,000 words
- Identifies words most associated with sentiment

2) BERT Embeddings

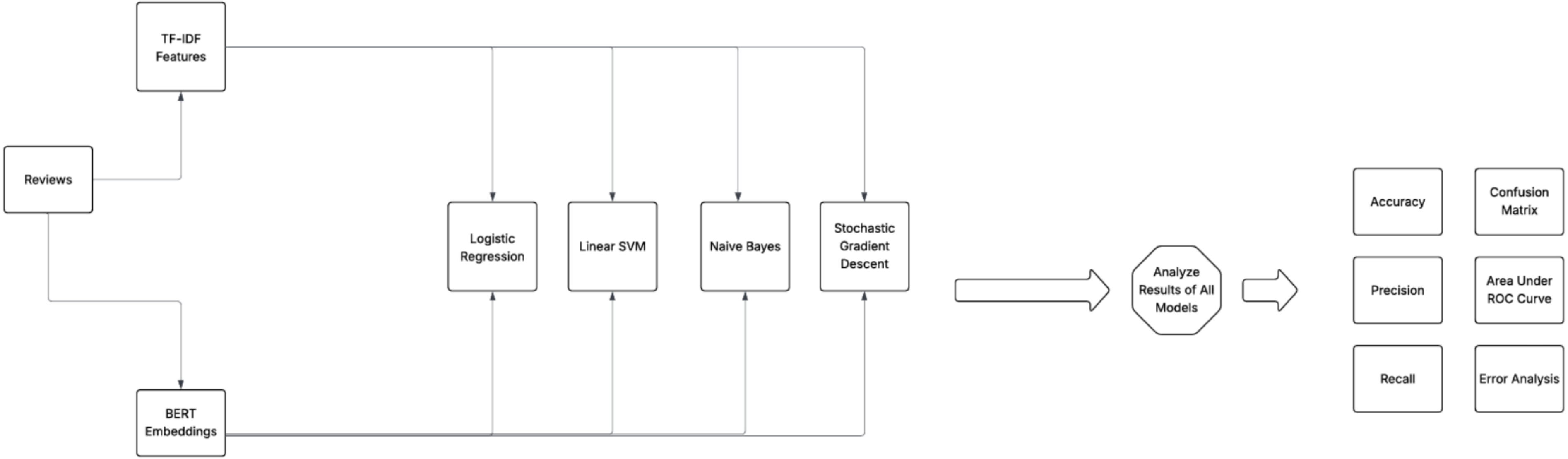
- Captures semantic meaning and word order

3) Classification Models

Trained classifiers: logistic regression, Linear SVM, Naive Bayes, & SGD

4) Evaluate Models

Measured and evaluated accuracy, precision, recall, & AUC


Reviews
TF-IDF Features
BERT Embeddings
Logistic Regression
Linear SVM
Naive Bayes
Stochastic Gradient Descent
Analyze Results of All Models
Accuracy
Confusion Matrix
Precision
Area Under ROC Curve
Recall
Error Analysis

Important Analysis Decision

BERT → **Fine-Tuning**

Slow, impractical to run

BERT → **Embedding**

More efficient while maintaining semantic meaning

Bias & Uncertainty Validation

Sources of Bias

- Self-selected by data collector → more extreme reviews that are easier to analyze
- English-only → not representative of global population

Solutions

- Balanced data set: 25,000 positive and 25,000 negative reviews
- Tested numerous models and feature representations ensuring reliable results

Determining Uncertainty

- Evaluated model performance using multiple metrics (accuracy, precision, recall, AUC)
- Compared results across multiple classifiers to verify consistency

Results & Conclusions

- Identified words most strongly predictive of positive or negative sentiment
- TF-IDF classifiers generally outperformed BERT, but all models were successful with > 85% accuracy
- Our hypothesis is confirmed: *Textual patterns and semantic meaning within movie reviews can accurately predict whether a review is positive or negative.*

```
Top 30 Positive Words:
great          0.014406
love           0.008619
best           0.008143
excellent      0.007262
wonderful      0.006149
loved          0.005117
life           0.005015
perfect        0.004664
amazing        0.004588
beautiful      0.004523
favorite       0.004053
brilliant      0.003865
family         0.003812
world          0.003707
enjoyed        0.003705
years          0.003662
war            0.003522
series         0.003519
story          0.003502
performance    0.003481
highly         0.003449
today          0.003430
fun            0.003406
superb         0.003328
young          0.003322
fantastic      0.003208
definitely     0.003128
performances   0.003061
true           0.002998
music          0.002947
dtype: float64
```

```
Top 30 Negative Words:
bad        -0.021252
movie      -0.013158
worst      -0.012846
br         -0.011799
just       -0.009229
waste      -0.008646
awful      -0.008553
terrible   -0.007995
plot       -0.007511
boring     -0.007168
acting     -0.007124
stupid     -0.006928
minutes    -0.006808
horrible   -0.006321
don        -0.006316
worse      -0.006230
poor       -0.006189
like       -0.005999
money      -0.005774
script     -0.005752
make       -0.005440
thing      -0.005343
crap       -0.004904
better     -0.004738
didn       -0.004633
supposed   -0.004415
avoid      -0.003963
actually   -0.003929
instead    -0.003892
poorly     -0.003868
dtype: float64
```

	model	accuracy	precision	recall	auc
0	TFIDF_LogReg	0.8938	0.881294	0.9102	0.960655
1	TFIDF_LinearSVM	0.8899	0.880094	0.9028	0.957939
3	TFIDF_SGD_LogLoss	0.8843	0.866349	0.9088	0.953393
5	BERT_LinearSVM	0.8689	0.872250	0.8644	0.944599
4	BERT_LogReg	0.8672	0.870909	0.8622	0.943477
2	TFIDF_MultinomialNB	0.8637	0.867448	0.8586	0.938572
7	BERT_SGD_LogLoss	0.8626	0.843241	0.8908	0.940497
6	BERT_GaussianNB	0.7761	0.792913	0.7474	0.851948

Next Steps

Improvements

- Collect and analyze more non-English reviews from a diverse array of platforms (not just IMBD)
- Fine-tune BERT to gain even better semantic understanding of reviews
    - Improve model performance on sarcastic or ambiguous reviews

Further Exploration & Future Questions

- How well does the model generalize to different types of text (e.g., product reviews, social media posts)?
- Which combinations of words or phrases are most indicative of review sentiment?
- Will these trends be consistent with the top individual words identified in this study?

References

[1] M. M. Danyal, S. S. Khan, M. Khan, S. Ullah, M. B. Ghaffar, and W. Khan, "Sentiment analysis of movie reviews based on NB approaches using TF–IDF and count vectorizer," *Social Network Analysis and Mining*, vol. 14, no. 1, Apr. 2024, doi: https://doi.org/10.1007/s13278-024-01250-9.

[2] N. M. Gardazi, A. Daud, M. K. Malik, A. Bukhari, Tariq Alsahfi, and Bader Alshemaimri, "BERT applications in natural language processing: a review," *Artificial Intelligence Review*, vol. 58, no. 6, Mar. 2025, doi: https://doi.org/10.1007/s10462-025-11162-5.

[3] Scikit learn, "TfidfVectorizer," *Scikit-learn.org*, 2018. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.h tml

[4] B. Muller, "BERT 101 - State Of The Art NLP Model Explained," *huggingface.co*, Mar. 02, 2022. https://huggingface.co/blog/bert-101

[5] "DistilBERT," *huggingface.co*. https://huggingface.co/docs/transformers/en/model_doc/distilbert

[6] GeeksforGeeks, "Handling Imbalanced Data for Classification," *GeeksforGeeks*, Jul. 08, 2020. https://www.geeksforgeeks.org/machine-learning/handling-imbalanced-data-for-classification/


iStock
by Getty Images