Applied
Machine
Learning

PROJECT PRESENTATION

TEAM : Longshot

MAY 2025

Content

Team and Roles

Gauranga

gauranga.mds2023@cmi.ac.in

Nikita

nikitakumari@study.iitm.ac.in

Dhruv

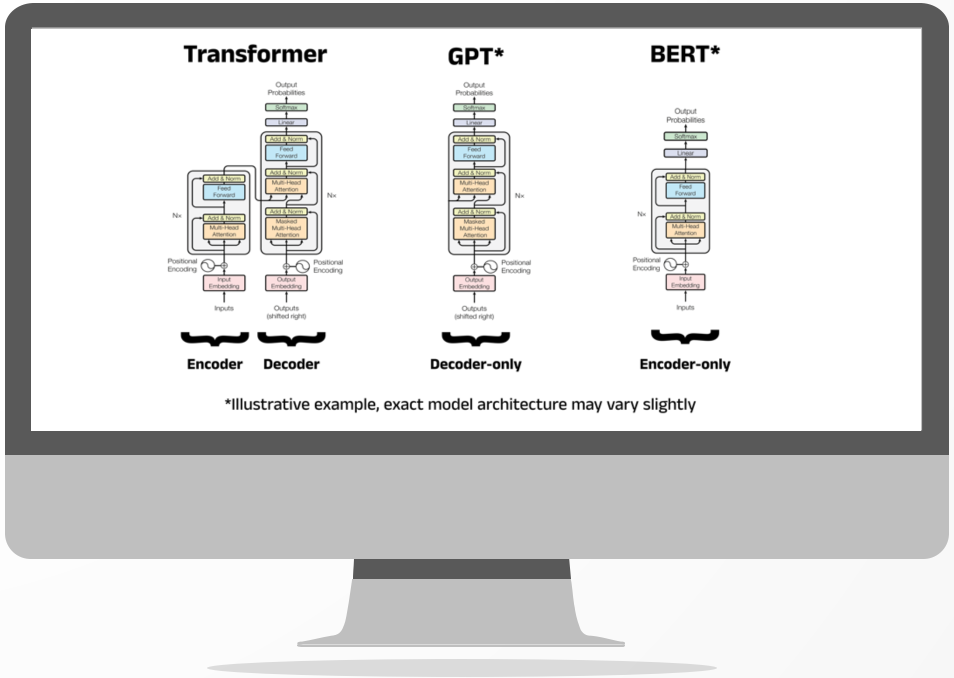
22f3001413@ds.study.iitm.ac.in

Vasu

21f3002975@ds.study.iitm.ac.in

Member	Responsibilities
Nikita	Background research, Data pre-processing and Design
Dhruv	Code base and ANN Modelling
Gauranga	ML Modelling and Experiments
Vasu	Interpretation and Reporting

Backdrop



Situation

There is a strong belief that Deep Learning based models can solve any problem way better than traditional ML models.

Question

Are the modern AI modeling architectures the silver bullet?
Are there any considerations to keep in mind?

Answer

Systematic comparison of Classical Models and Modern Deep Learning based Architectures.

Problem of choice

I. Good Reads is a very popular source of book reviews for users. Book reviews impact reader perceptions and can directly impact sales.

II. Currently likes and comments on reviews are being used to determine the popularity of a review which in turn determines the placement of the review. Higher the popularity, higher is placement.

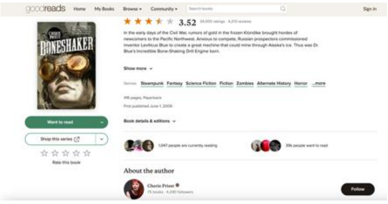
III. However, this is our best guess and a black box. To verify this, we have reconstructed the popularity metric and tested whether a model can be trained to determine which reviews will be popular vs. not so popular.

IV. While doing so, we are evaluating whether simpler, faster classical methods can match modern deep learning models at this classification task.

Study design

	Choices	Metrics	Experiments
Data	• 15 Mn reviews data set (2 Mn books, 465 k users) • Construct book review popularity measure based on likes and comments • Meta data of reviews (length of the review, % verbs etc.) • English reviews	• Class imbalance : ~15% reviews are popular	• Down sampling • Tfidf, CBoW, Skip-Gram and BERT embeddings. • Normalization of meta data
Models	• Model = f (Review + Review's Meta data) -> Classification	• Precision Recall • FN is costlier => Recall	• Logistic • XGBoost • ANN • Transformer
Action	• If prob(popular) > default threshold push the review up the order on the website	• Update threshold to improve recall	• Pick highest recall subject to acceptable precision score

Data and feature engineering

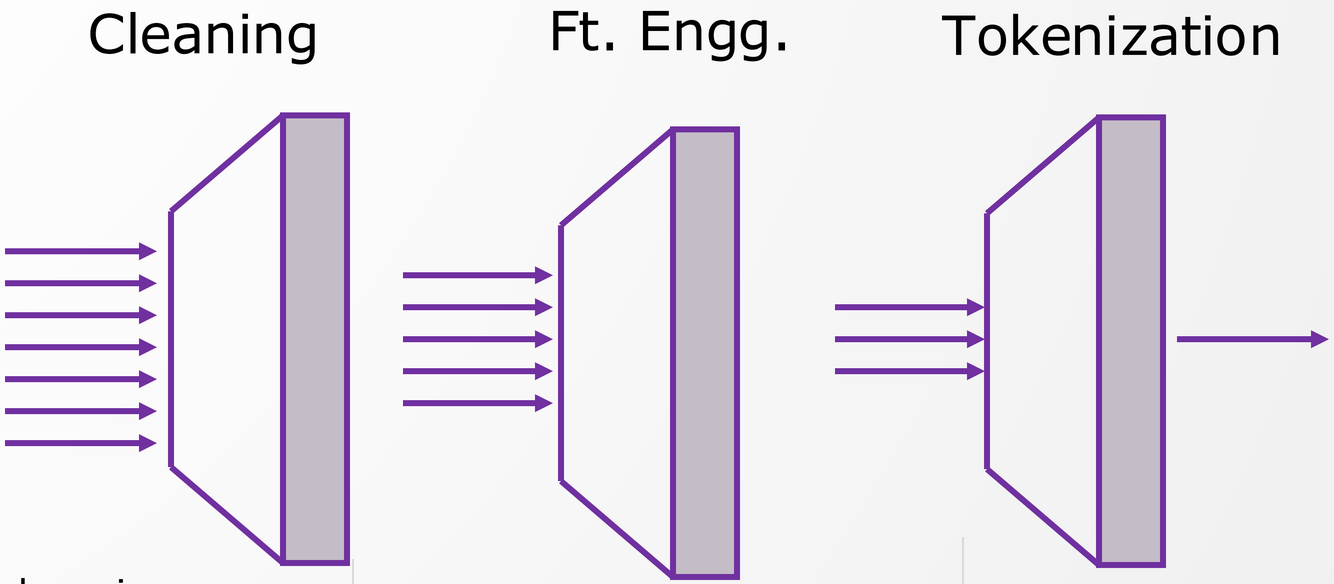


In the early days of the Civil War, rumors of gold in the frozen Klondike brought hordes of newcomers to the Pacific Northwest. Anxious to compete, Russian prospectors commissioned inventor Leviticus Blue to create a great machine that could mine through Alaska's ice. Thus was Dr. Blue's Incredible Bone-Shaking Drill Engine born.

15 Mn
multilingual reviews

2 Mn books

465k users



Cleaning

- English reviews
- Review reactions = Likes + Comments
- At least 100 reactions

Ft. Engg.

- No of reviews by reviewer
- # of sentences, length # of words, # and % verbs, %, # and % of nouns # and % of adjectives nouns,
- VADER sentiment score
- User rating, Diff from average

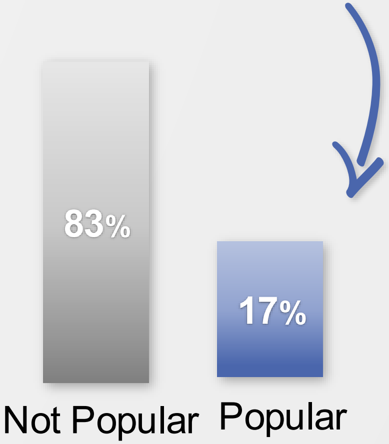
Tokenization

- NLTK
- Stop words removal
- Tokenization

POPULAR REVIEW IS DEFINED AS:

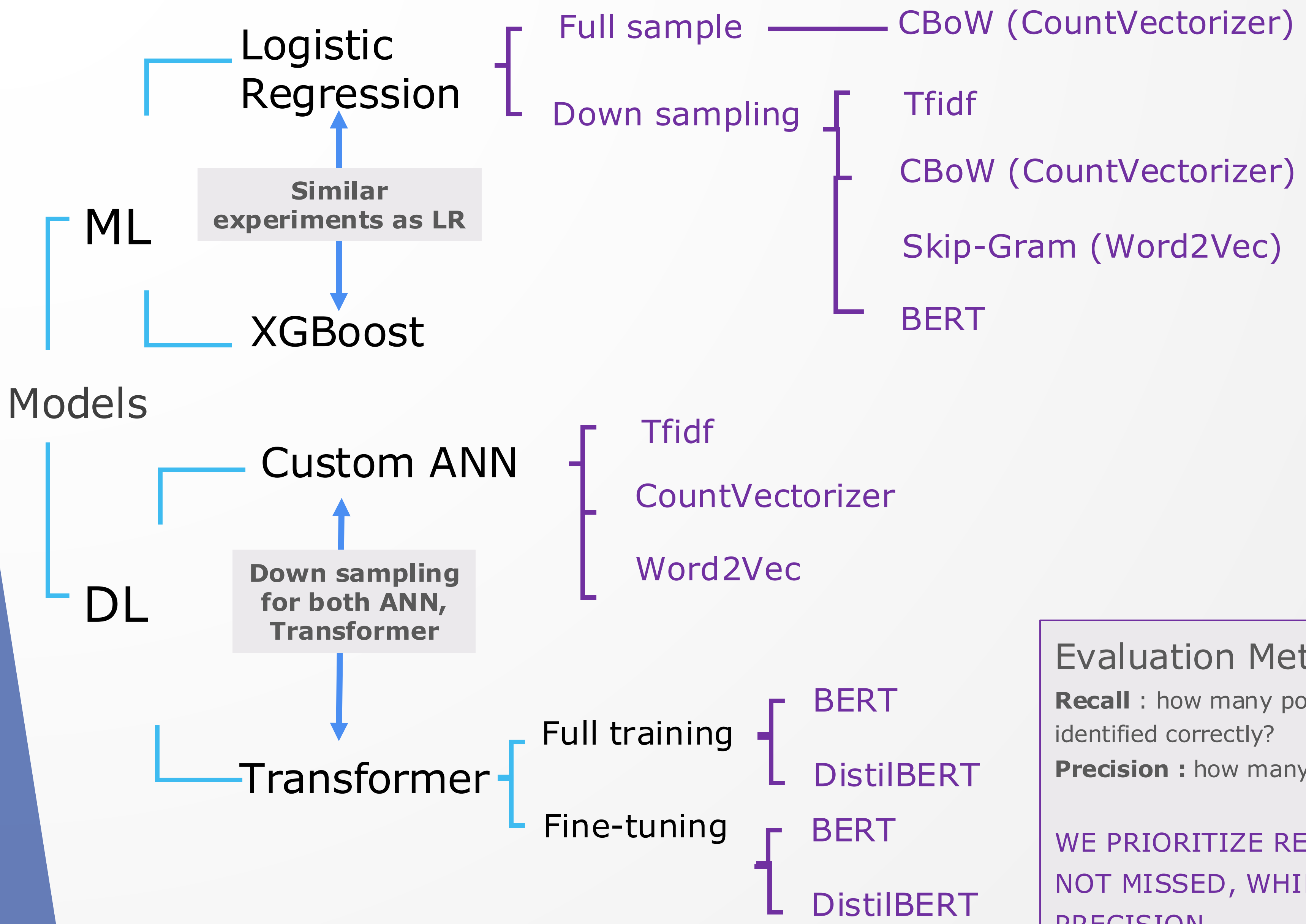
Share(Comment + 👍)

> 2 %



1.7 Mn Reviews

Key Experiments#



- Models
  - ML
    - Logistic Regression
      - Full sample
        - CBoW (CountVectorizer)
      - Down sampling
        - Tfidf
        - CBoW (CountVectorizer)
        - Skip-Gram (Word2Vec)
        - BERT
    - XGBoost
  - DL
    - Custom ANN
      - Tfidf
      - CountVectorizer
      - Word2Vec
    - Transformer
      - Full training
        - BERT
        - DistilBERT
      - Fine-tuning
        - BERT
        - DistilBERT

Similar experiments as LR

Down sampling for both ANN, Transformer

 # Not all experiments are mentioned here

Hyper-Parameters

LR

Regularization strength
(C = 10, 1, 0.01, 0.001)

XGBoost

Estimators : 100,1000
Depth: 4,6
Lr : 0.1, 0.3

BERT/ DistilBERT

No of tokens : 64, 128
Epochs: 1,3

Evaluation Metrics

Recall : how many popular reviews have been identified correctly?

Precision : how many identified as popular are correct?

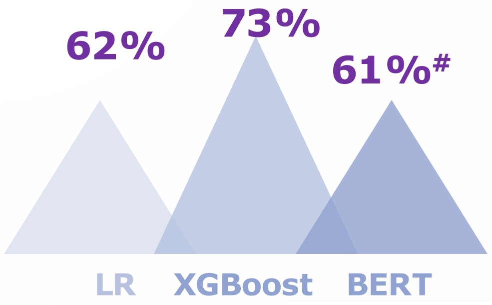
WE PRIORITIZE RECALL TO ENSURE POPULAR REVIEWS ARE NOT MISSED, WHILE MAINTAINING AN ACCEPTABLE LEVEL OF PRECISION

Results

Model	Down Sampling	Embeddings	Accuracy	Precision	Recall/ Sensitivity	f1	Specificity	AUC (ROC)		REMARKS
LR	Yes	Tfidf	0.74	0.32	0.66	0.43	0.75	0.70		
LR	Yes #	CBoW	0.77	0.34	0.62	0.44	0.79	0.71		Acceptable
LR	No	CBoW	0.86	0.60	0.18	0.28	0.98	0.58		Very poor recall
LR	Yes	Word2Vec	0.74	0.30	0.64	0.41	0.74	0.69		
LR	Yes	BERT	0.71	0.29	0.66	0.41	0.72	0.69		
XGBoost	Yes	Tfidf	0.72	0.31	0.73	0.44	0.72	0.72		
XGBoost	Yes	CBoW	0.72	0.31	0.73	0.44	0.72	0.72		XGBoost does better !
XGBoost	No	CBoW	0.68	0.68	0.22	0.34	0.98	0.60		
XGBoost	Yes	Word2Vec	0.71	0.30	0.72	0.42	0.71	0.72		
ANN	Yes	Tfidf	0.72	0.31	0.70	0.43	0.73	0.71		
ANN	Yes	CBoW	0.71	0.30	0.74	0.43	0.70	0.72		No upside vs. ML models
BERT	Yes	BERT	0.74	0.30	0.58	0.39	0.77	0.73		Not the best as expected
BERT– Ft*	Yes	BERT	0.64	0.22	0.57	0.32	0.65	0.65		

** Ft – Fine tune, # 15% of the full sample are positive reviews, train sample down sized to reflect 50-50% split*

Why is RECALL not improving with DL models?



I. Are we measuring the right thing?
II. Is embeddings an issue?
III. Is dataset imbalance making learning hard?

Popularity

- Popularity (dependent/outcome) metric was created via feature engineering and not independently or directly measured from users
- Correlation between popularity and features is low (0.3) to begin with

Embeddings

Diagnostics on TP and FN reviews revealed the following:

1. When emotional subtlety is missing, the model is failing to learn low key appreciation and classifying it as unpopular
2. Deeply thoughtful and philosophical reviews with no overt positivity are being deemed unpopular
3. Positivity and sentiment are being confused. Could be a popular review but highly critical and negative in sentiment

Imbalanced

- Yes, imbalance is an issue evidenced by poor model performance on full sample
- Down sampling is showing better modeling outcomes (70% vs. ~ 20% recall scores)

 # Average across experiments

A closer look at TP and FN reviews

ACTUAL

PREDICTED

	Not Popular	Popular
Not Popular	**TRUE NEGATIVE** • I was devastated when the book ended because I could have read it forever, as it took me to a literary high, making me feel like I can’t read another book right away for fear of ruining its aura, where everything felt like a metaphor down to the smallest thought. Jhumpa Lahiri…	**FALSE POSITIVE** • Ooooooo mmmmmmm gggggggI finished the book, and it was amazing; I loved every minute of the series.
Popular	**FALSE NEGATIVE** • I think the story grabbed me with its pure human aspect, maybe on a level I could relate to, as Wynne created a living, breathing character in Henry, a shy, bit-of-a-loner guy who you get to glimpse as a man, like many of us, going through the motions of life. Henry works as a security guard at a museum where a painting of a woman first arrives and captures his imagination; he grows obsessed with this sad, beautiful figure…	**TRUE POSITIVE** • I found this set of comics better than the first, and as a huge fan, the new Riverdale arc redeemed itself; while I still love the original, this one is slowly improving.

Conclusions and recommendations

- Define popularity clearly, what is it supposed to capture?
- Re-evaluate how to measure popularity metric
- Ideally an independent measure of popularity is required and not feature engineered
    - E.g. : review and recommendation
- And if required, capture from users (maintain a stratified sample)

- Domain Adaptive Pre-Training : Train BERT further using unsupervised masked language modelling (MLM) on reviews corpus
- Test for Sentiment ≠ Popularity
- Stop-words and tokenization should not remove evocative differentiators

- Deal with data set imbalance; down sampling has shown promise in mitigating this risk
- Deep Learning models do not provide an automatic advantage, this is a cognitive bias
- Do not overlook traditional ML models (particularly Boosting algorithms), more interpretable and computationally friendly
- Follow Occam's Razor : adding meta data and additional features may not automatically improve model performance

For suggestion and questions

21f3002975@ds.study.iitm.ac.in,
gauranga.mds2023@cmi.ac.in

TEAM : Longshot

MAY 2025