SENTIMENT CLASSIFICATION

STUDENT ID: 2320757



Motivation and Objective

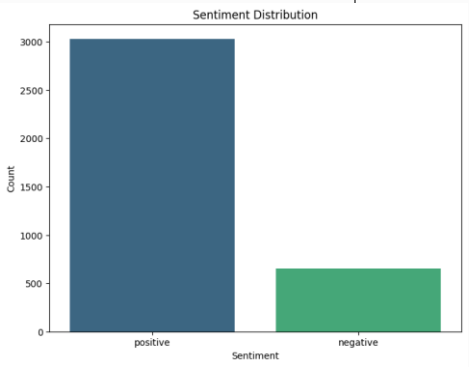
In the wake of technology, the ecommerce industry has taken over a major business area and hence there is a lot of competition among the service providers. They want the customer to be happy and be better than others. Sentiment analysis could play a vital role in this as if the customer reviews are negative, it will be bad for business. Also, customers too, can choose the best products available analyzing the sentiment of the reviews to make a fine choice.

Objective:

- To build and implement effective classifiers
- To understand how to choose the best model
- To efficiently train, validate and evaluate the models
- To effectively draw comparison between models

DATA

- Automotive reviews on ecommerce given by the customers and its sentiment
- Highly imbalanced with negative sentiment being minority class.
- 3 sets: Train, validation and Test
- Ratio of both the sentiments is equal in both validation and training dataset



Dataset	Total	% Positive	% Negative
Train	3681	82.31	17.69
Valid	454	82.38	17.62
Test	409	NA	NA

Two Main Methods

01

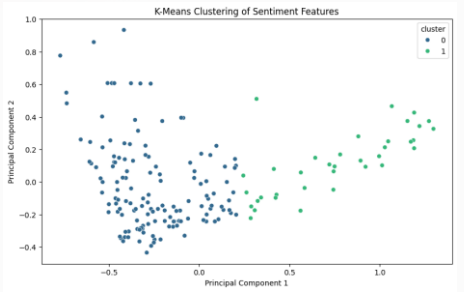
Unsupervised

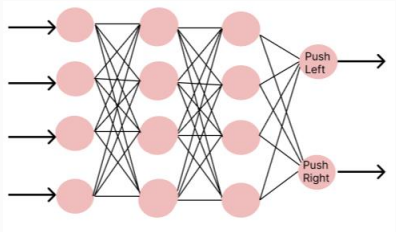
Training without labelled data

02

Discriminative

Training with labelled data





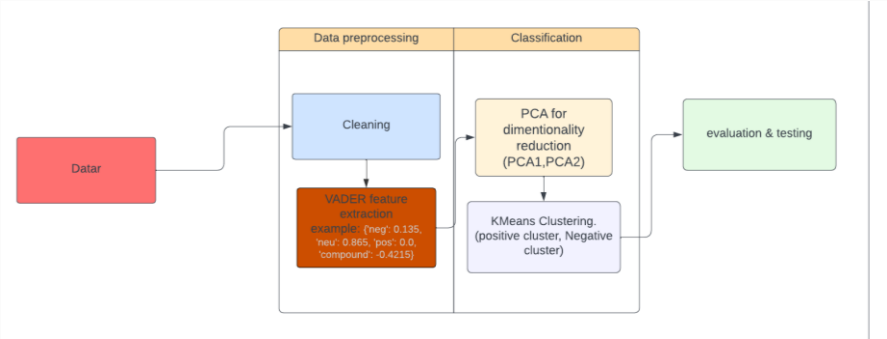
01

Unsupervised

Unsupervised Model

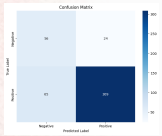
The model architecture is broadly divided into 3 parts:

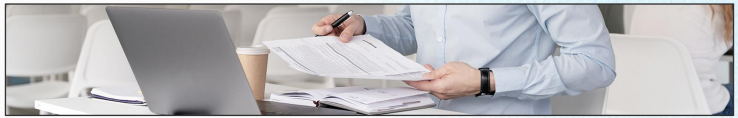
- Extracting Features VADER Score
- PCA to reduce dimensionality
- K-Means clustering to group alike sentiments together

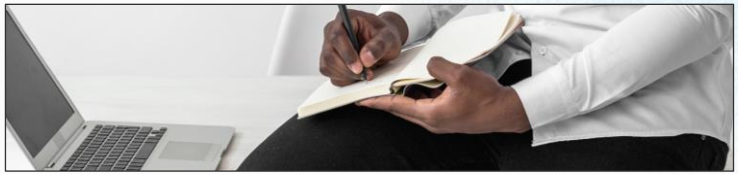


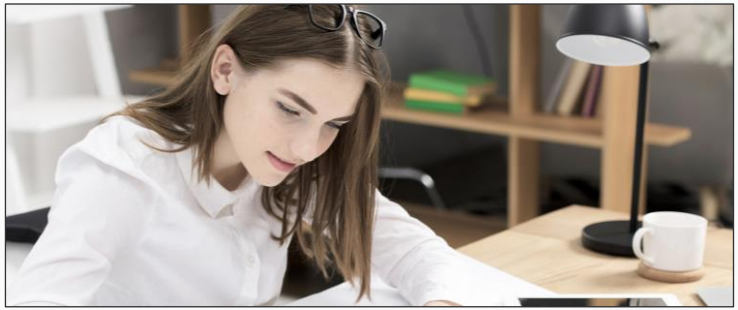
Parameter tuning was done using GridSearchCV for: PCA components and K-Means iterations.

Results


Confusion Matrix
Negative
Positive
True Label
58
24
472
979
Negative
Positive
Predicted Label
300
250
200
150
100
50



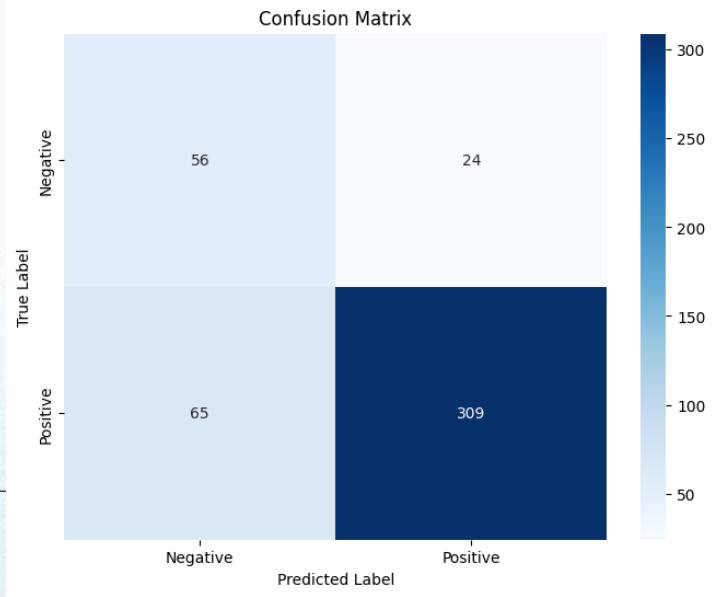




Confusion Matrix

•**True Positive (TP)**: 309
•The model correctly predicted 309 positive instances.
•**True Negative (TN)**: 56
•The model correctly predicted 56 negative instances.

•**False Positive (FP)**: 24
•The model incorrectly predicted 24 negative instances as positive.
•**False Negative (FN)**: 65
•The model incorrectly predicted 65 positive instances as negative.



Performance Evaluation

Silhouette Score

A score of 0.52.

moderate clustering quality

F1 Score

Macro: 0.71 otherwise : 0.87

the model performs well overall, there is a notable discrepancy in performance between classes

Precision & Recall

Recall (weighted): 0.80

model correctly identifies 80.4% of the actual instances of each class

Precision (weighted): 0.84

the model predicts a class, it is correct about 84.6% of the time

Strength & Weakness

Strength

- The model is highly confident in sentences which has more words.
- decent confidence in both classes.
- The model shows high confidence in reviews with clear, unambiguous sentiments. For example, "just as described. delivered very quickly! No complaints" is correctly identified as positive with high confidence (0.84)
- The model accurately identifies strongly negative reviews with high confidence. For instance, "Worked first 2 times. Now won't work at all. Junk!" is correctly predicted as negative with a confidence score of 0.89.

Weakness

- The model misclassifies some positive reviews as negative with high confidence. For example, "High quality, but bought as spares" is predicted as negative with a confidence score of 0.890453. This indicates overconfidence in certain incorrect predictions.
- It is not confident enough in predicting the negative class.
- Its confidence is not much especially for scentences with fewer words like in ‘good value !’
- Sometimes influenced by negative details despite the overall positive sentiment.

02

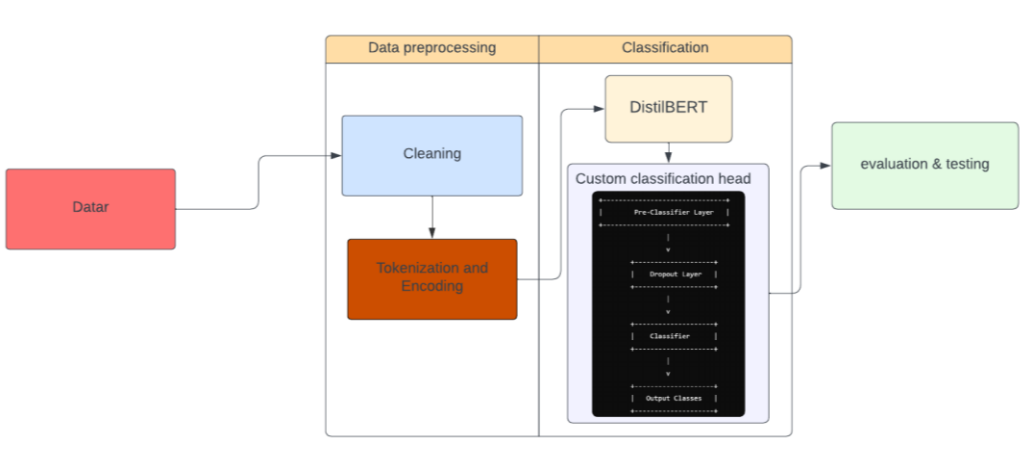
Discriminative

Discriminative Model

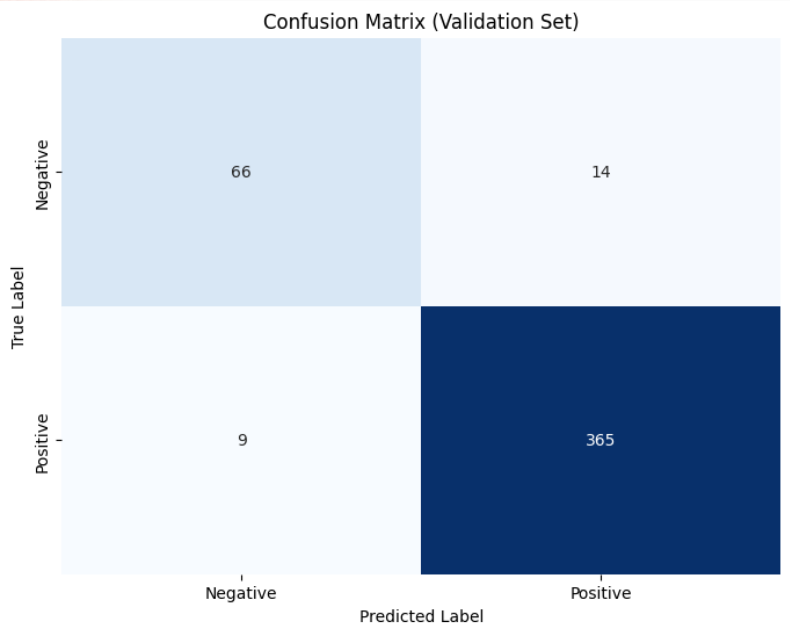
The model architecture is broadly divided into 3 parts:

- Tokenization and Encoding
- DistilBERT Model
- Custom Classification Head

Hyperparameter tuning was done using Kfold CV for: Learning rate.



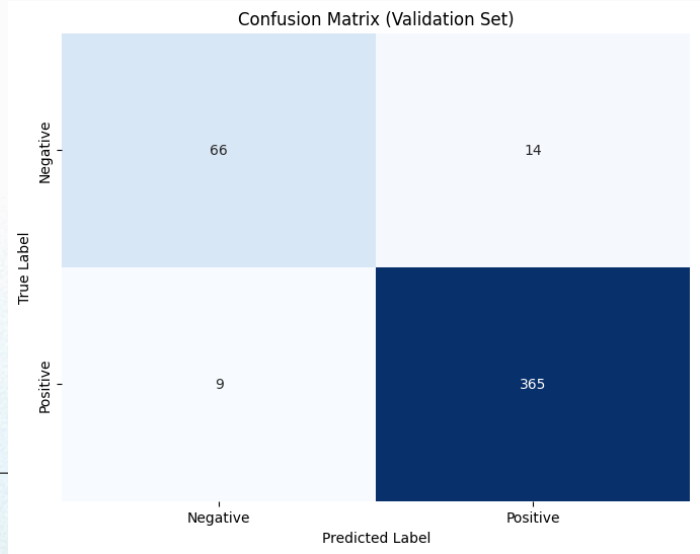
Results - Discriminative



Confusion Matrix

•**True Positive (TP): 309**
•The model correctly predicted 309 positive instances.
•**True Negative (TN): 56**
•The model correctly predicted 56 negative instances.

•**False Positives (FP)**: 14
•These are the instances where the model incorrectly predicted the positive class when the actual class was negative (Type I error).
•**False Negatives (FN)**: 9
•These are the instances where the model incorrectly predicted the negative class when the actual class was positive (Type II error).



Performance Evaluation

Class-wise Accuracy

The accuracy for the negative class is 82.5%, indicating the model is less accurate in predicting negative sentiments.

The accuracy for the positive class is 97.59%, indicating the model is highly accurate in predicting positive sentiments.

This discrepancy suggests an imbalance in the model's performance, where it performs significantly better on positive samples than negative ones.

F1 Score

Macro: 0.91 otherwise : 0.9695

good performance across both classes but reflects that performance on the negative class is bringing down the average.

Precision & Recall

Recall (weighted): 0.949

the model correctly identifies positive cases accurately, balanced by the actual distribution of instances in each class

Precision (weighted): 0.948

the model predicts a class, it is correct about most of the time

Strength & Weakness

Strength

- - High Confidence in Predictions
- - Effective Handling of Clear Sentiments
- - Correctly classifies the misclassified examples too.
- - Effectively distinguishing between positive and negative sentiments in most cases

Weakness

- - There are cases where the model is highly confident but incorrect in its prediction.
- - The model sometimes fails to grasp the nuanced context of a review, especially in cases with mixed sentiments
- - Over-reliance on Keywords

Combined Model

Comnined Model

Rule based model in which the decision rule for choosing which model's prediction to take is based on a weighted comparison of confidence scores.

By combining models, this approach leverages the strengths of both supervised and unsupervised methods. For example, the unsupervised model might excel in capturing underlying data patterns not apparent during training, while the supervised model provides precision due to its training on labeled data.

Results

exactly same performance as that of discriminative model is seen which implies that:

* The discriminative model alone is sufficiently powerful, and the unsupervised model’s contribution becomes negligible.
* The unsupervised model does not capture additional meaningful patterns or insights that the discriminative model misses.
* The combined model essentially mirrors the performance of the discriminative model, indicating that the discriminative model's predictions are dominant.

Comparison

Comparison between the 2 models

Class-Wise precision & Recall

discriminative model has a better balance between precision and recall across all classes compared to the unsupervised model.

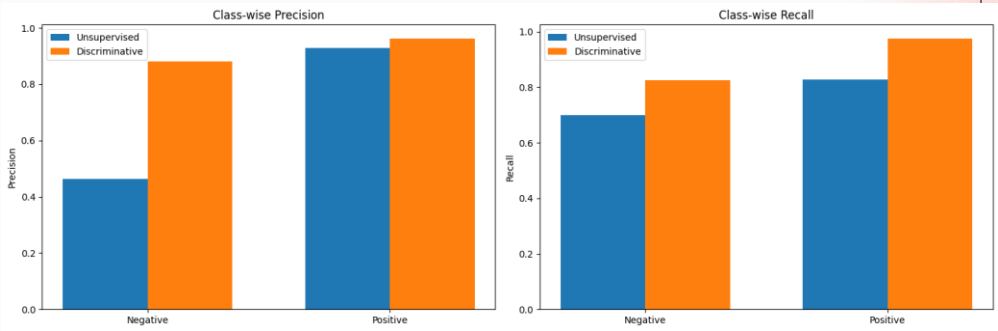
Confidence

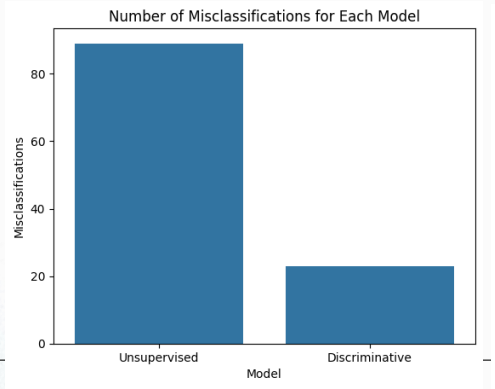
It can be seen that the discriminative model is consistently more confident in its predictions

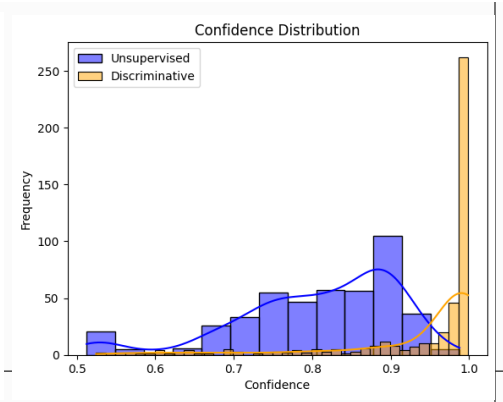
The unsupervised model, while also confident, shows more variability in its confidence levels.

Misclassifications

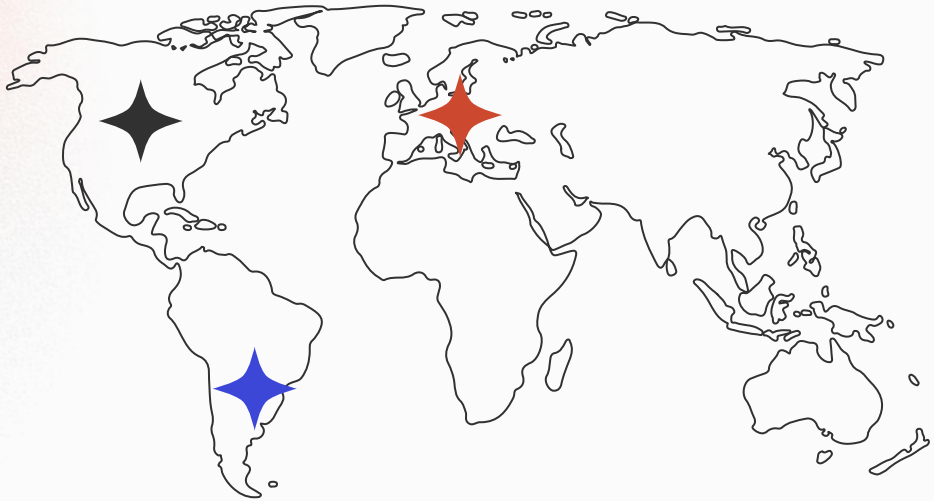
The discriminative model has fairly less missclassification.







State-of-The-Art Model



LLMS

LLMs like GPT are the top SoTA models as of this year for most of the NLP task including sentiment analysis.

GPT

In this section GPT-3.5-Turbo model would be used via API for sentiment classification.

BERT

The Discriminative model that we used (Fine-tuned DistilBERT) is already a SoTA model however more advanced models like GPT are prevalent nowadays

Comparison with SoTA Model

Our models were compared with GPT model using their classification over 7 different examples:

Serial No.	Category	text	original_label	sentiment_should_be	Unsup model	Dis_model	GPT
1	easy	love the product easy to use	positive	positive	☑	☑	☑
2	easy	good value !	positive	positive	☑	☑	☑
3	hard	the dilution ratio on this allows you to make ...	negative	negative	☑	☐	☑
4	hard	i purchased this for my semi worn armrest to p...	positive	negative	☐	☑	☑
5	easy	i bought these to replace my worn out headligh...	positive	positive	☐	☑	☑
6	ambiguous	customer service and shipping was first rate a...	negative	positive or negative	☑	☑	☑
7	hard	this gel is not sticky even when warmed up! i ...	negative	negative	☐	☑	☑

Serial No.	Category	original_label	sentiment_should_be	prediction_unsup	confidence_unsup	prediction_dis	confidence_dis	prediction_GPT	confidence_GPT
Text 1	easy	positive	positive	positive	0.773061	positive	0.996934	positive	0.95
Text 2	easy	positive	positive	positive	0.630032	positive	0.99685	positive	0.92
Text 3	hard	negative	negative	negative	0.829751	positive	0.636293	negative	0.85
Text 4	hard	positive	negative	positive	0.781125	negative	0.608185	negative	0.85
Text 5	easy	positive	positive	negative	0.895176	positive	0.992711	positive	0.92
Text 6	ambiguous	negative	both	negative	0.890453	positive	0.886653	positive	0.85
Text 7	hard	negative	negative	positive	0.811304	negative	0.992177	negative	0.95

Comparison analysis

F1 Score

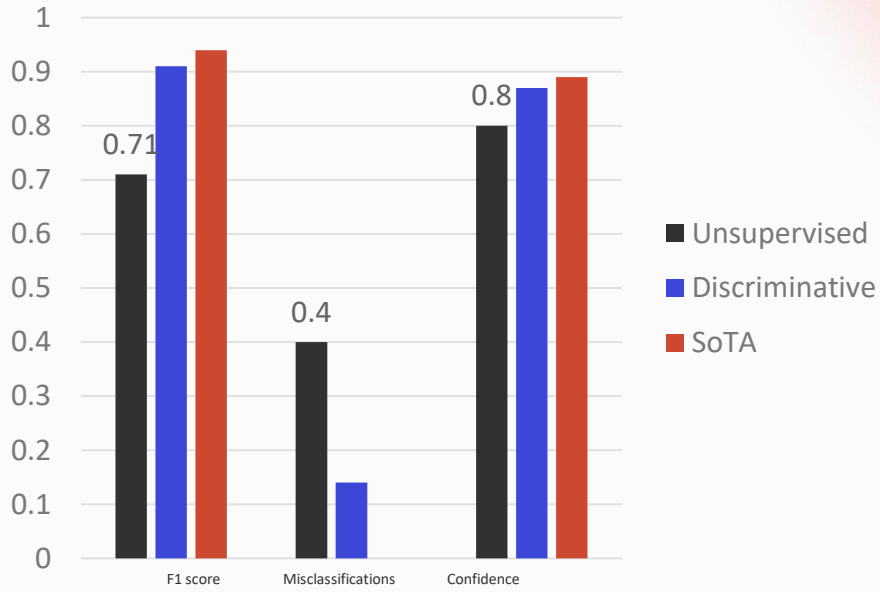
GPT has the highest F1 score, followed by Discriminative model

Misclassifications

Unsupervised model had the most misclassifications

Confidence

All show good confidence overall



Example Analysis

Category	Unsupervised Model	Discriminative Model	GPT Model
Easy	Correct predictions but less confident in short texts. Occasionally misclassifies with high confidence.	Generally correct predictions with high confidence.	Consistently correct predictions with high confidence.
Hard	Struggles with context, leading to misclassifications with high confidence.	More context-aware but occasional low-confidence misclassifications.	Handles context well with high confidence and accuracy.
Ambiguous	Correct predictions with high confidence when multiple interpretations are possible.	Correct predictions with high confidence when multiple interpretations are possible.	Correct predictions with high confidence when multiple interpretations are possible.

Final Take & Findings

In terms of dataset and training:
Unsupervised Models: Performance may vary based on the quality and nature of the data used for training. Simpler methods might miss nuanced context.
Discriminative Models: Performance generally improves with more labeled training data, allowing the model to learn from diverse examples.
GPT Models: Extensive pre-training on large datasets enhances their ability to understand and generate text, making them highly effective in sentiment analysis.

In terms of confidence Scores:
Discriminative and GPT Models: Higher confidence scores in correct predictions indicate the model's robustness in understanding context and sentiment.
Unsupervised Models: Lower confidence or incorrect predictions in complex cases reflect limitations in capturing nuanced sentiment.

Overall: SoTA model was able to beet both models including the Discriminative one but by close margin. Discriminative model should promising results but could improve in predicting negative sentiment andsame goes for unsupervised model.

THANKS!

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics & images by **Freepik**