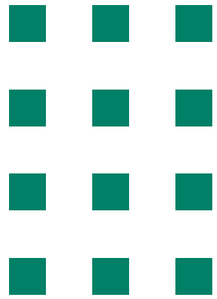
Sentiment Analysis on Woman's Clothing Reviews Using BERT



By: Team Job Application
Christian Arifin Gouw - F11315020

Task: sentiment analysis

Dataset: Women's E-Commerce Clothing Reviews by Nick Brooks

Objective: Experimenting with word embeddings to see if it can improve model performance.

Baseline approach: BoW + traditional classifier

Proposed approach: BERT + traditional classifier

Baseline result: 84% macro average on F1-Score

INTRODUCTION

DATASET

Rows: 23486 rows

Features: 10 feature variables

- Review Text
- Rating

Consideration:

- 4% Null value in ‘Review Text’

Women's E-Commerce Clothing Reviews

23,000 Customer Reviews and Ratings

kaggle.com

PUBLICATION

Class	Precision	Recall	F1-Score	Support
Negative	0.47	0.50	0.49	289
Neutral	0.31	0.18	0.23	22
Positive	0.96	0.96	0.96	4215
Average	0.93	0.93	**0.93**	
Total				4526

Statistical Analysis on E-Commerce Reviews, with Sentiment...

Understanding customer sentiments is of paramount importance in marketing strategies today. Not only will it give companies an insight as to how customers perceive their products and/or services,...

DATASET

Note:

The baseline came from his work, which got an 84% macro average F1-Score on Logistic Regression and Naive Bayes

Predicting Sentiment from Clothing Reviews

Explore and run machine learning code with Kaggle Notebooks | Using data from Women's E-Commerc...

k Kaggle / Apr 20, 2020

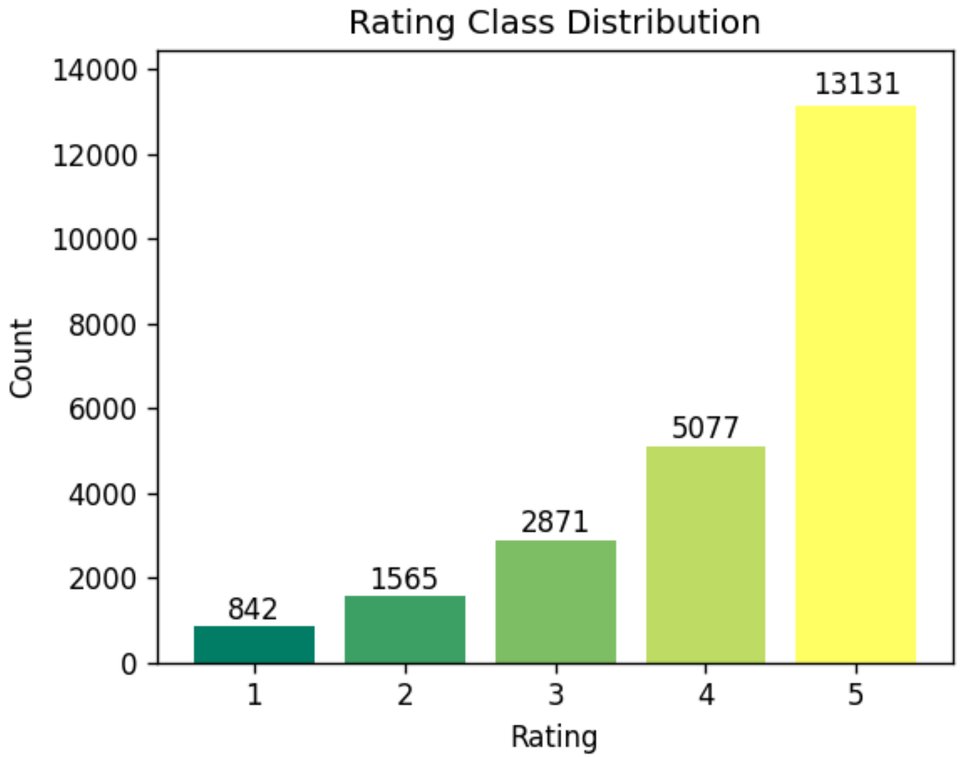
All Words in the Reviews

Word Cloud

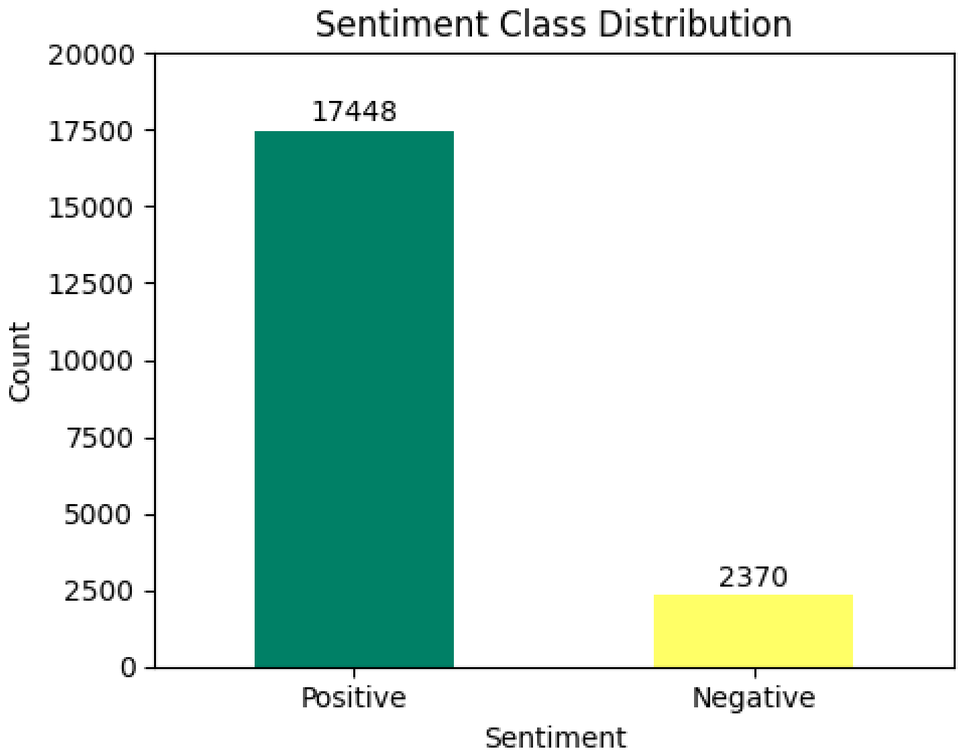
DATASET

Bar Chart

Before preprocessing



After preprocessing

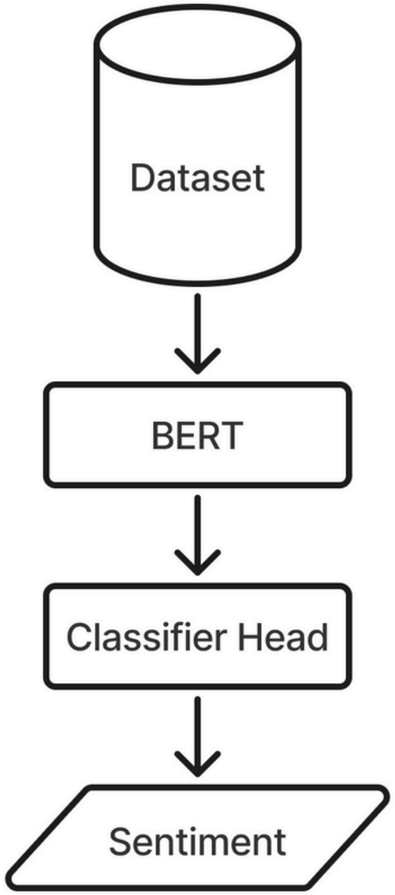


METHODOLOGY

PROPOSED METHOD

Classifier Head:

- Logistic regression
- Naive bayes
- SVM



Why BERT?

Because it captures deep contextual relationships between words, unlike traditional methods like Bag-of-Words or TF-IDF, which treat words independently.

What is the rationale behind using traditional classifier models?

- Isolate and evaluate the effectiveness of BERT embeddings as feature extractors.
- Traditional models are easier to implement.

DATA PREPROCESSING

Removing neutral polarity reviews.

Removing null value

Label mapping.

Splitting the dataset into an 80% training dataset and a 20% testing dataset.

Tokenize using BertTokenizerFast tokenizer.

FEATURE EXTRACTION

BERT base model

Approximately 110M parameters and 12 transformer layers.

BERT for embedding extraction (without fine-tuning), the computational burden lies mostly in inference.

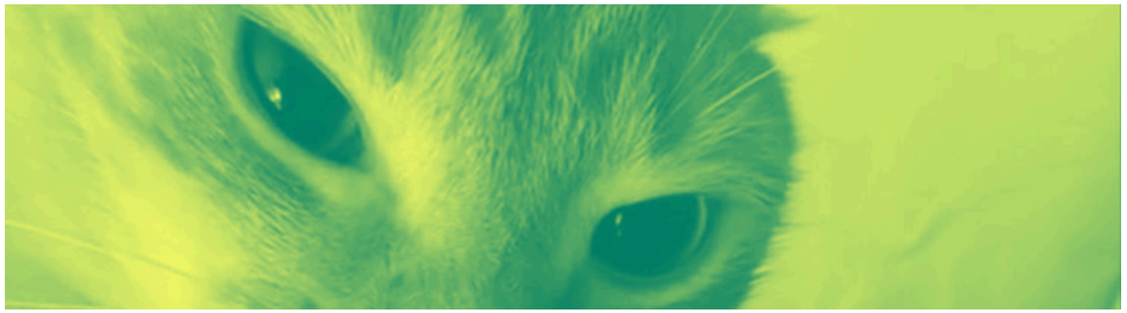
The downstream classifiers are lightweight, with logistic regression having a negligible parameter count in comparison.

Classifier	Elapsed Time
Logistic Regression	0:00:01.056909
Naive Bayes	0:00:48.484381
Support Vector Machine	0:01:41.359712

EVALUATION METRIC

F1-Score

The harmonic mean of precision and recall. A high F1-score means the model is performing well in both precision and recall.



Precision

The ratio of true positives to all predicted positives. High precision means fewer false positives.

Recall

The ratio of true positives to all actual positives. High recall means fewer false negatives.

RESULT

Initial experimentation without handling imbalance dataset.

Model	Macro Average		
	Precision	Recall	F1-Score
Logistic Regression	0.86	0.81	0.83
Naive Bayes	0.68	0.82	0.71
SVM	0.87	0.75	0.79

Logistic Regression				
Class	Precision	Recall	F1-Score	Support
0 (Negative)	0.76	0.65	0.70	494
1 (Positive)	0.95	0.97	0.96	3470

Naive Bayes				
Class	Precision	Recall	F1-Score	Support
0 (Negative)	0.40	0.83	0.54	494
1 (Positive)	0.97	0.82	0.89	3470

Support Vector Machine				
Class	Precision	Recall	F1-Score	Support
0 (Negative)	0.81	0.51	0.62	494
1 (Positive)	0.93	0.98	0.96	3470

RESULT

Experimentation result after undersampling.

Model	Macro Average		
	Precision	Recall	F1-Score
Logistic Regression	0.88	0.88	0.88
Naive Bayes	0.84	0.84	0.83
SVM	0.88	0.88	0.88

Logistic Regression

Class	Precision	Recall	F1-Score	Support
0 (Negative)	0.86	0.89	0.88	465
1 (Positive)	0.89	0.87	0.88	483

Naive Bayes

Class	Precision	Recall	F1-Score	Support
0 (Negative)	0.81	0.87	0.84	465
1 (Positive)	0.87	0.80	0.83	483

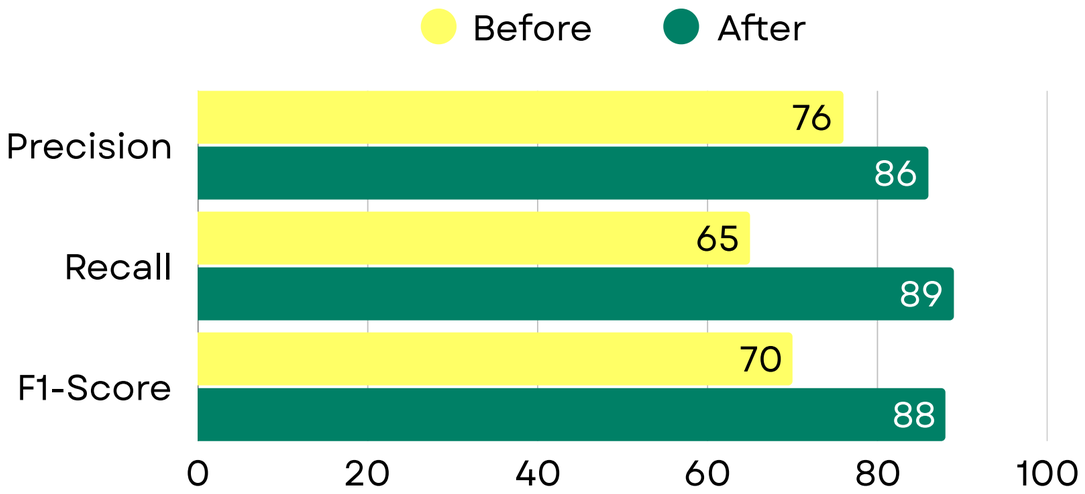
Support Vector Machine

Class	Precision	Recall	F1-Score	Support
0 (Negative)	0.85	0.92	0.88	465
1 (Positive)	0.92	0.84	0.88	483

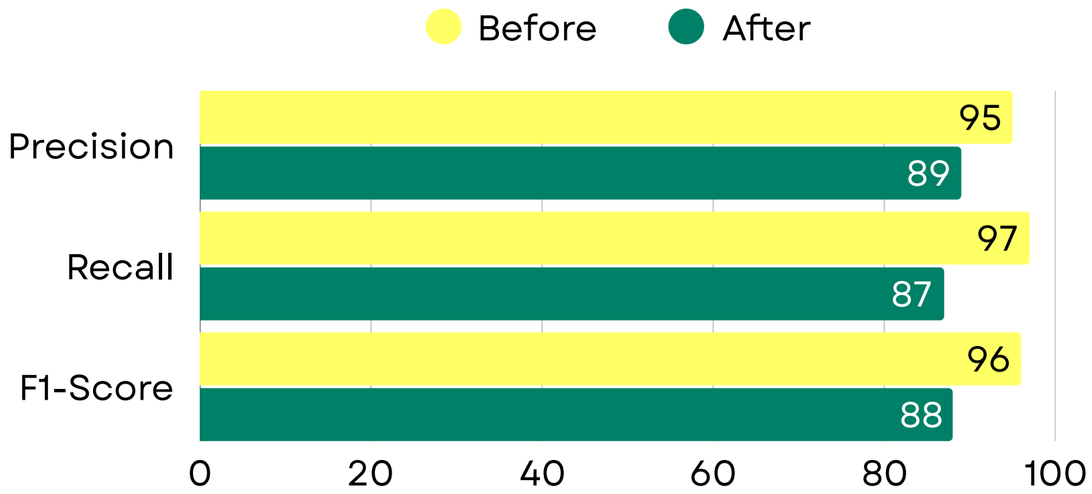
DISCUSSION

Logistic Regression

Negative Sentiment

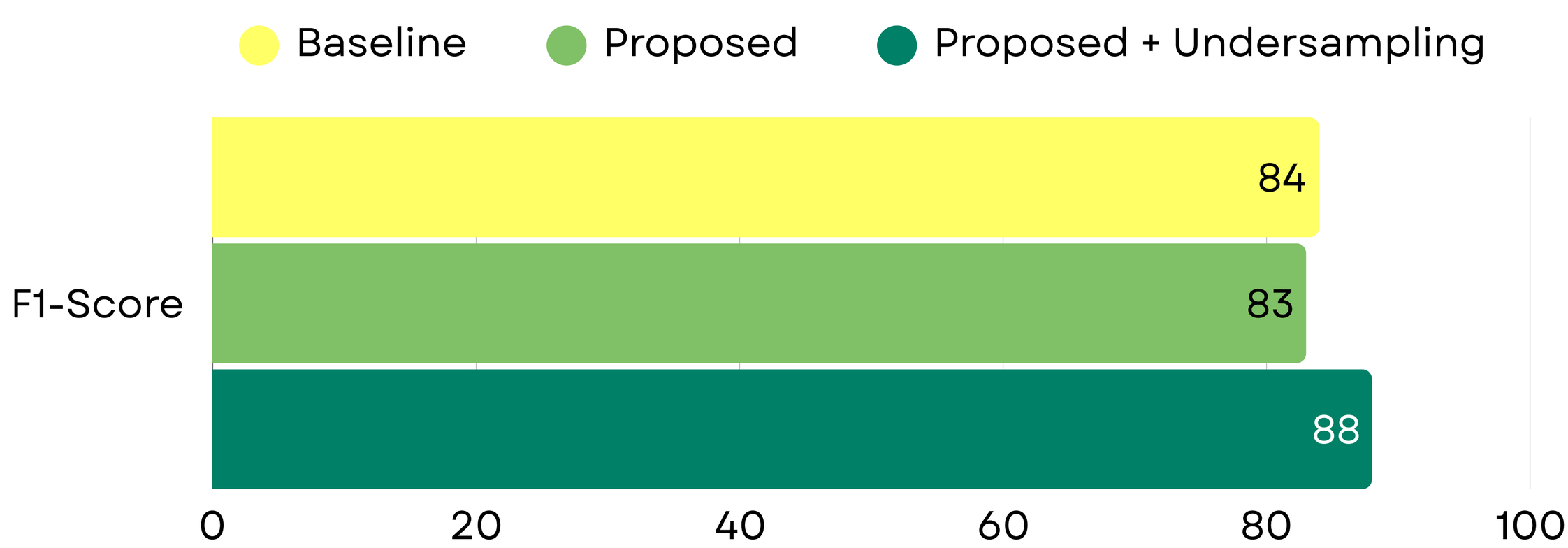


Positive Sentiment



DISCUSSION

Logistic Regression



CONCLUSION

This experiment shows that using word embeddings as the feature extractor in sentiment analysis helps even traditional model such as logistic regression and SVM to achieve great performance. Both models performs better than the BoW approach, with 88% macro average on F1-Score.

Thank you

Any questions?