Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 235 (2024) 2318–2331

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Machine Learning and Data Engineering (ICMLDE 2023)

Transformer Based Unsupervised Learning Approach for Imbalanced Text Sentiment Analysis of E-Commerce Reviews

Khushboo Taneja*, Jyoti Vashishtha, Saroj Ratnoo

Guru Jambheshwar University of Science and Technology, Hisar-125001, India

Abstract

In e-commerce, online reviews play a vital role in understanding customer requirements. Companies can perform analysis of these reviews to make right business decisions. However, manually analyzing the content generated by each user on social media is challenging due to its high volume and frequency. Sentiment analysis (SA) is a Natural Language Processing (NLP) technique which provides an automated solution to this problem. It is a challenging subtask of text classification. Over a long time, researchers have proposed different approaches to solve it. This paper proposes an unsupervised learning approach using transformer architecture to perform SA on women's clothing e-commerce dataset which is imbalanced in nature. In this study, we have fine-tuned DistilBERT, a cutting-edge transformer based pre-trained model and developed models for two subtasks of SA named as Sentiment Classification (SC) and Product Recommendation (PR). The proposed models have achieved the highest F1 scores (0.79 for SC and 0.85 for PR), AUC scores (0.98 for SC and 0.96 for PR) along with the highest accuracy of 0.96 for SC and 0.91 for PR. We found that the performance of our models is least affected by imbalanced dataset issues. The results show that the proposed models have significantly outperformed the traditional supervised approaches and various other existing state-of-the-art (SOTA) models. This study can contribute to a better understanding of consumer sentiment and consumer psychology in the e-commerce transaction business.

Keywords: Sentiment analysis; e-commerce; transformer; DistilBERT; unsupervised learning; pre-trained model.

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-mail address: khushbootaneja@gmail.com

10.1016/j.procs.2024.04.220

1. Introduction

Online reviews give deeper and valuable insight of customer's sentiments about various product or services offered by companies. Organizations use this data to analyze customer's behavior to understand their needs and potential customers seek this information to make purchase decisions. Positive reviews can increase company's sales and brand equity value while negative reviews can break brand trust, lower sales and degrade company's reputation. Therefore, it is important for organizations to analyze the online consumer's reviews in order to provide good customer support, develop better products and attract new customers to earn more profit. However, it is very costly and nearly impossible for organizations to manually analyze the huge amount of text generated by users on various social media platforms.

Sentiment analysis (SA) is a growing research area that has widely gained attention from researchers due to its application in different domains such as business, marketing, finance, medicine, etc. [1]. It can help companies to perform analysis of reviews in a cost effective manner. The area of recommender systems is another popular application domain of SA. Product recommendation based on customer reviews help companies to provide personalized view of products or services. Over a long time, the research community has faced many theoretical and technical issues related to NLP that makes this task more challenging. Researchers have proposed three main methods for solving the problem of SA: lexicon-based method, machine-learning method, and hybrid method [2].

Lexicon-based approach was the first to be used for SA. It can be either dictionary-based or corpus-based. In dictionary-based method, a dictionary of terms is used to perform SA, such as SentiWordNet [3] and WordNet-Affect [4]. However, in corpus-based method, various techniques based on conditional random field (CRF) [5], k-nearest neighbors (k-NN) [6] and hidden Markov models (HMM) [7] are used to perform statistical analysis on collection of documents.

Machine learning approach also known as supervised learning approach relies on traditional machine learning [ML] and deep learning [DL] models [8]. In this approach, a labeled dataset is fed to a ML or DL algorithm to train a SA model. This approach has been widely accepted for sentiment classification. The most popular traditional ML algorithms used for SA are Naïve Bayes classifier and Support Vector Machines [9]. Deep neural networks (DNN), convolutional neural networks (CNN), and recurrent neural networks (RNN) are the recently used DL models for SA [10].

Hybrid approach is a combination of lexicon approach and machine learning approach [11]. The first phase of this method uses a lexicon-based approach to analyze the text and provides an output that is given into a ML algorithm during second phase in the form of input training data. The output obtained in the first phase is used to expand the affective dictionary.

In the past, ML and hybrid approaches have dominated the field of SA. However, the recent studies in SA are mainly focused on using DL models. Dang et al. (2020) have performed a comparative study of DNN, CNN and RNN for SA and based on their experiments they found that RNN when combined with word embedding gives highest reliability as compared to DNN and CNN but it has high computational cost [10]. Several other research studies have focused on building powerful SA models based on DL techniques for different applications such as marketing [12], recommender systems [13] and movie reviews [14].

The necessity for a significant amount of labelled data to provide better accuracy is a major disadvantage of supervised learning approaches in NLP. Recently, NLP research has spawned the transformer [15] architecture, which is built on a self-attention mechanism that focuses on the context of each word in a sentence or document. Several transformer based models are now available in the literature to solve named entity recognition, part of speech tagging, text classification, sentiment analysis, and other NLP tasks. These models are known as pre-trained language models (PLMs) [16] which are trained on huge amount of unsupervised data and can be adapted to solve NLP task using a little amount of annotated data. This process is known as transfer learning [17].

Due to lack of human expertise, annotated data is not available for many NLP tasks. Moreover, we often find imbalanced data distribution in many real-world applications like fraud detection, spam detection, churn prediction, etc. which leads to biased results towards majority class due to more number of training examples. Supervised approaches cannot tackle the problem of data imbalance. Various techniques such as data under-sampling and over-sampling are available to fix this issue but they have their own limitations [18]. Over-sampling can cause model over-fitting and under-sampling might lead to poor model training. PLMs can solve both the above mentioned

problems. There are a number of transformer based PLMs available in the literature like BERT [19], DistilBERT [20], T5 [21], ELECTRA [22], etc. Among these, DistilBERT is a modified variant of BERT which is smaller in size and cost effective [23].

Many supervised learning algorithms have been employed in the past to perform SA. The potential of unsupervised learning is still unexplored. To fill this gap, our primary focus in this research work is to apply transformer based unsupervised learning approach using DistilBERT to solve the problem of SA in the area of e-commerce on an imbalanced women's clothing dataset [24].

1.1 Main Contributions and Significance

- In this research work, we have proposed unsupervised learning approach employing transformer based DistilBERT and developed fine-tuned models for two related subtasks of sentiment analysis: Sentiment Classification (SC) and Product Recommendation (PR) on women's clothing e-commerce dataset [24].
- We have performed extensive experimentation and also developed several baseline models for both sub-tasks with traditional supervised ML algorithms such as SVM, Naïve Baiyes (NB), Logistic Regression (LR), Ridge Classifier (RC), Passive Aggressive Classifier (PAC), Voting Classifier (VC), and Random Forest (RF) along with deep learning based "bidirectional RNN with long-short term memory unit (BiLSTM)" [25] and presented a comparative analysis with the proposed models.
- The experimental results indicate that both the proposed models are outperforming baseline models and other SOTA models in the literature on the same dataset. We have also observed that performance of proposed models is least influenced by the imbalanced nature of data.
- This study can aid companies in better comprehending consumer behavior. We expect that it will close the gap between corporate viewpoints and consumer attitudes.

The remainder of the paper is structured as follows. In section 2, related work is discussed. The materials and methods utilized to conduct the experiments is explained in Section 3. Section 4 presents the implementation details. The experimental study's findings are discussed in Section 5. Section 6 concludes the work along with its future scope.

2. Related Work

NLP research has mostly focused on classic ML and DL methods for SA in the literature [26]. Noor and Islam (2019) have proposed different ML algorithms such as NB, JRip, J4 and Sequential Machine Optimization (SMO) each from four different categories including Bayes theorem, rules, trees and SVM to perform SA on Amazon's women's e-commerce dataset [27]. In their work, SMO achieved the highest accuracy of 80.875%. Agarap (2020) developed models for sentiment classification (SC) and product recommendation (PR) using BiLSTM which achieved the test accuracy scores of 0.92 and 0.88 respectively on women's e-commerce clothing dataset [28]. Due to imbalanced nature of dataset, both the models are giving skewed performance towards the majority class and obtained very low recall and F1 scores in case of minority class. Nawaz et al. (2021) performed the study of SA for PR on the same dataset using CNN and LSTM models. Based on their results, LSTM has outperformed CNN with a classification accuracy of 91.69% and achieved F1-score of 56.67%. In this work, the authors recommended the use of LSTM over CNN for SA [29]. Mehta and Panda (2022) have used CNN with TFIDF and Word2Vec for sentiment classification of e-commerce reviews [30]. Loukili et al. (2023) has also compared various supervised ML algorithms such as RF, LR, K-Nearest Neighbor (KNN) and Catboost for Product Recommendation [31]. Table 1 presents the summary of existing work in the literature for SA on women's e-commerce reviews.

Table 1. Summary of existing work in the literature for SA on women's e-commerce clothing reviews.

References	Sub-task	Models	Results (Highest performing model)	Merit	Demerit
Noor and Islam (2019) [27]	Sentiment Classification	NB, JRip, J48, SMO	Accuracy: 0.81 Precision: 0.809 TP rate: 0.809 FP rate: 0.191	Low computational cost	Low accuracy
Agarap (2020) [28]	Sentiment Classification	BiLSTM	Accuracy: 0.92 Precision: 0.58 Recall: 0.55 F1: 0.56	High accuracy	Skewed performance
	Product Recommendation	BiLSTM	Accuracy: 0.88 AUC: 0.79	Good accuracy	Skewed performance
Nawaz et al. (2021) [29]	Product Recommendation	LSTM, CNN	Accuracy: 91.69 Recall: 0.71 F1: 56.67 AUC: 0.71	Good use of SOTA technique	Low recall, F1 and AUC scores
Mehta and Panda (2022) [30]	Sentiment Classification	CNN	Accuracy: 0.94 Precision: 0.94 Recall: 0.94 F1: 0.94	Good use of SOTA technique	Model over-fitting
Loukili et al. (2023) [31]	Product Recommendation	RF, LR, KNN, Catboost	Accuracy: 0.90 Precision: 0.854 Recall: 0.786 F1: 0.814	Low computational cost because of simple techniques	Low recall and F1 score

It can be seen from table 1 that the existing approaches employing ML or DL algorithms are showing skewed performance towards majority class due to their inability to deal with imbalanced data. Also, these approaches are based on non-contextual feature extraction methods such as TFIDF and word2vec which do not consider the dynamic context of words in the text. None of the approaches have used data resampling to deal with imbalanced dataset problem. The results of previous studies show very high accuracy score but low recall, F1 and AUC scores. However, accuracy is considered as a misleading performance metric in case of imbalanced dataset problem. Generally, a model with high recall, F1 and AUC scores is considered as the best quality model in the domain of imbalanced text classification.

In this research work, we addressed the above mentioned limitations with the application of transformer based unsupervised learning approach. Transformer was originally proposed for the task of machine translation in NLP. Due to its success, it was soon adopted for other NLP tasks as well. BERT [19], a PLM pre-trained on two tasks: Masked Language Model and Next Sentence Prediction, was proposed by Devlin et al. (2019) and has proven to be a milestone due to its bidirectional nature and contextual understanding of text. Researchers have used BERT to solve various NLP tasks like document classification, sentiment analysis, text summarization etc. It has also been adapted for similarity search applications [32]. BERT's large model size is responsible for difficulty in pre-training it from scratch. Victor et al. (2020) proposed a modified variant of BERT known as DistilBERT [20] which is smaller in size and cost effective. In our research work, we have employed transformer based DistilBERT PLM for both subtasks of SA.

We have also ran experiments with traditional ML models such as SVM, NB, LR, PAC, RC, VC, and RF along with BiLSTM which act as baseline models. While developing the baseline models, we have applied data under-sampling to address the problem of class imbalance. The performance of each model is measured in terms of precision, recall, accuracy, F1 and AUC scores. The study shows that DistilBERT has obtained significant results over traditional baselines surpassing the issue of imbalanced data. Our models have also achieved SOTA results in comparison to existing approaches in the literature on women's clothing e-commerce dataset.

3. Materials and Methods

This section gives details about DistilBERT that is used to develop fine-tuned models for SC and PR in subsection 3.1. Subsection 3.2 presents the overall proposed model architecture adopted to fine-tune DistilBERT for SA. Subsection 3.3 describes the dataset. Subsection 3.4 presents the list of tools used to conduct the experiments. Subsection 3.5 gives the details of various preprocessing steps on the dataset. Subsection 3.6 shows the exploratory data analysis.

3.1 DistilBERT

BERT [19] is a multilayer bidirectional transformer encoder which is based on the original architecture of transformer [15], a state-of-the-art (SOTA) network that can perform parallel processing of data and reduces the model training time significantly. It is based entirely on self-attention mechanism that helps the transformer to establish contextual relation between words. Self-attention can be calculated using equation (1) [15].

$$Attention(Q,K,V) = soft\max\left(QK \Big/ \sqrt{d_k}\right)V \tag{1}$$

Here Q, K, V stands for Query, Key and Value matrices respectively, d is the dimension of Key (K) vector.

However, BERT is very large in size and cannot be adopted for on-device real-time applications due to high computational budget and memory requirements. Victor et al. (2020) proposed a compressed version of BERT named as DistilBERT which is based on the same general architecture of BERT. In their work they have shown that pre-training of a smaller language model can be possible by using a compression technique called knowledge distillation. Using a triple loss combining language modelling, distillation and cosine-distance losses, DistilBERT can retain 97% of BERT performance on many downstream-tasks while being 40% smaller and 60% faster than BERT. It can also be deployed for on-the-edge applications for small devices like mobile phones.

3.2 Proposed Model Architecture

In this work, we have employed DistilBERT and developed two fine-tuned models corresponding to each subtask of SA i.e. Sentiment Classification and Product Recommendation. A model can be fine-tuned by adding a task specific layer on top of original DistilBERT architecture as shown in figure 1. The process of fine-tuning begins by first initializing DistilBERT model with pre-trained parameters and then updating all the parameters using labeled data for each sub-task. Figure 1 illustrates the unified overall architecture for proposed models. The input to the model consists of a sequence/sentence. A sequence is processed by first converting it into tokens for each word and adding a special [CLS] token in the beginning. These tokens are then passed to multiple transformer layers. The final transformer layer outputs the embeddings for each token. The output embedding of [CLS] token is passed through Feed Forward Neural Network (FFNN) to perform sentence classification. The sigmoid function is used to determine the output in terms of various categories.

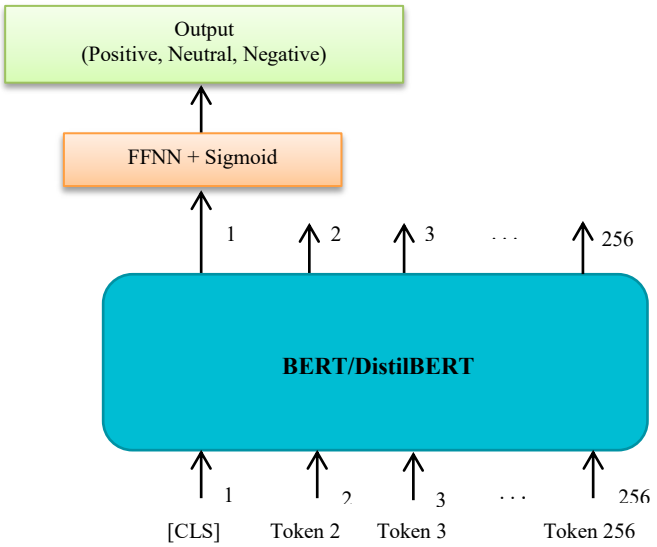


Fig. 1. Overall architecture of proposed model.

3.3 Dataset Description

We have used the women's clothing e-commerce reviews [24] dataset for this research study. It consists of 23,486 rows and 10 feature variables. Each row consists of a review written by real customers along with other information. The dataset is downloaded from kaggle website. Table 2 shows the statistical distribution of feature variables. Table 3 represents the frequency distribution of different features of original dataset.

Table 2. Statistical distribution of feature variables.

Index	Mean	Std.	Min.	25%	50%	75%	Max
Recommended IND	0.82	0.38	0	1	1	1	1
Rating	4.19	1.11	1	4	5	5	5
Positive Feedback Count	2.53	5.70	0	0	1	3	122
Clothing ID	918.11	203.29	0	861	936	1078	1205
Age	43.19	12.27	18	34	41	52	99

Table 3. Frequency distribution of feature variables.

Feature variable	Frequency count	Type of variable
Recommended IND	2	Binary
Rating	5	Positive ordinal integer
Positive Feedback Count	82	Positive integer
Clothing ID	1206	Integer categorical
Age	77	Positive integer
Title	13993	String
Review Text	22634	String
Class Name	20	Categorical
Department Name	6	Categorical
Division Name	3	Categorical

3.4 Tools

The research work conducted in this paper is carried out using Google Colab. The following python libraries are used to perform the experimental work.

- numpy: It is used to perform various mathematical operations in the experiment.
- pandas: It is used to read the dataset and helps in performing various data cleaning and feature engineering operations.
- matplotlib.pyplot: It is a data visualization library and helps in visualizing data in various forms like bar chart, histogram, scatter plot, line plot etc. It is used to analyse the dataset.
- seaborn: It is also a data visualization library with more advance features and offers more axis control for graphs and figures.
- nltk: It is one of the popular NLP library in python. We have used SentimentIntensityAnalyzer from nltk library to automatically generate the sentiment polarity for each review.
- ktrain: It is a very efficient library for building text classifiers using transformer models [33].
- sklearn: It is a popular machine learning library to perform various operations required while building a model.
- tensorflow: It is a deep learning framework to build deep learning models.
- preprocess_kgptalkie: It is a python package that is used exclusively for data cleaning [34].
- word cloud: It is a python package to visualize the most frequently occurred words belonging to a particular category.
- imblearn: It is used to perform resampling on training data while running baseline models.
- scikit-plot: We have used this python package to plot ROC curves.

3.5 Data Pre-processing

3.5.1 Data Cleaning

We have performed the following data cleaning operations before analysing the original dataset.

- Drop the columns: Clothing ID and Title columns are dropped from the dataset as we found these insignificant for data analysis.
- Removal of null values: Null values are removed from Review Text, Division Name, Department Name, and Class Name columns.
- Review Text Cleaning: Review Text column is the heart of the analysis so we have performed text cleaning operation on each review by applying get_clean function of preprocess_kgptalkie package and removed emails, urls, html tags, special characters, etc.

After performing the above operations we obtained the cleaned dataset which is comprised of 22,628 rows and 8 columns.

3.5.2 Feature Engineering

To perform the sentiment analysis, instead of manually tagging the review text, we have used nltk's SentimentIntensityAnalyzer to generate the sentiment polarity scores for each review. In this step, we have introduced the Sentiment column in which reviews with Polarity Score greater than zero are labelled as Positive, reviews with Polarity Score equal to zero are labelled as Neutral, and reviews with Polarity Score less than zero are labelled as Negative.

3.6 Exploratory Data Analysis

3.6.1 Univariate Distribution

We have performed the univariate distribution of various features and made some interesting observations as described below.

- Age: The age distribution in Figure 2 shows that most of the reviews are given by customers under age group 35 to 44. This data reveals that the customers of this age group are highly satisfied and work can be done to find out various ways in order to satisfy the customers of other age group.

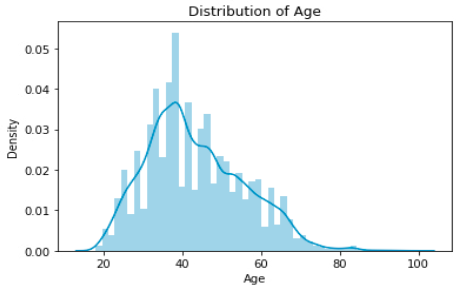
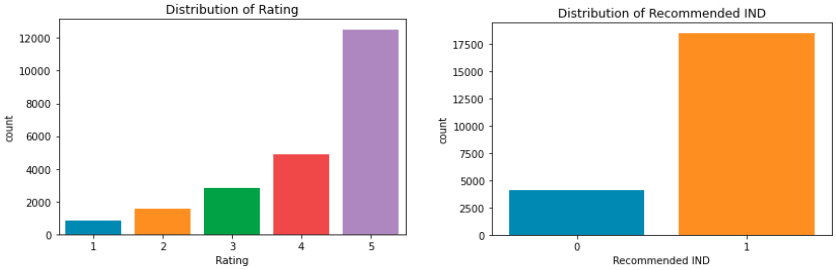


Fig. 2. Univariate distribution of age.

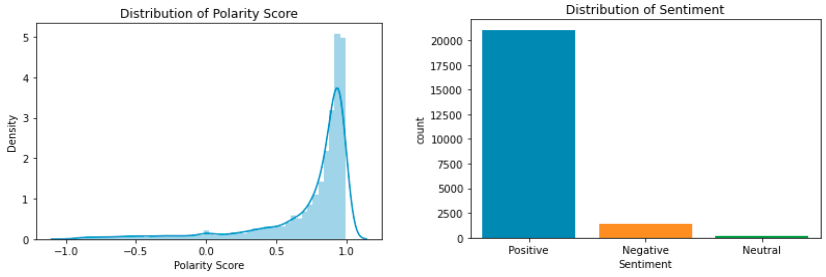
- Rating and Recommended IND: The univariate distribution of Rating and Recommended IND as shown in Figure 3 suggests that a large segment of customers are satisfied with e-commerce products as count of reviews having 5 star rating is highest and 1 star rating is lowest. Also, the count of recommended products is very high.



(a) (b)

Fig. 3. (a) Univariate distribution of Rating; (b) Univariate distribution of Recommended IND.

- Polarity Score and Sentiment: The distribution of Polarity Score and Sentiment in Figure 4 indicates high percentage of satisfied customers for e-commerce products as most of the reviews are positive.



(a) (b)

Fig. 4. (a) Univariate distribution of Polarity Score; (b) Univariate distribution of Sentiment.

- Department Name, Division Name and Class Name: The frequency distribution of Department Name in Figure 5(a) shows that the Department of Tops and Dresses have large number of reviews. It also suggests

e-commerce to find ways to improve Trends Department having least number of reviews. The bar chart of Division Name as shown in Figure 5(b) reveals that General Division Name has highest number of reviewers. Similarly, the frequency distribution of Class Name in Figure 5(c) indicates that Dresses, Knits and Blouses are top 3 apparel with highest count of reviews.

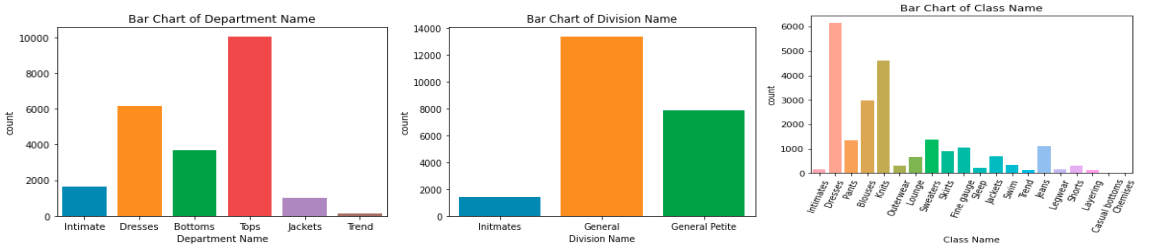


Fig. 5. (a) Univariate distribution of Department Name; (b) Univariate distribution of Division Name; (c) Univariate distribution of Class Name.

3.6.2 Bivariate Distribution

- Rating by Recommended IND: Figure 6(a) shows the correlation of Rating and Recommended IND. It indicates that many recommended products have rating of 3 ,4 or 5 and the products which are not recommended have rating of 1 or 2.

- Sentiment by Recommended IND: Figure 6(b) shows a positive correlation between Sentiment and Recommended IND. It reveals that Recommended IND is a strong indicator of Sentiment.

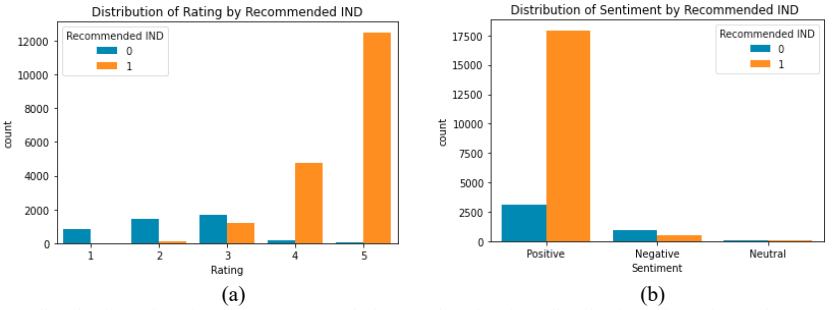


Fig. 6. (a) Bivariate distribution of Rating by Recommended IND; (b) Bivariate distribution of Sentiment by Recommended IND.

4. Implementation

In this research work, we have employed two models related to SA, one is PR and other is SC. For each model, we have used python's ktrain [33] library and fine-tuned Hugging face's *distilbert-base-uncased* with the hyper-parameter settings given in Table 4. Subsection 4.1 and 4.2 presents the implementation details for each model.

Table 4.Hyper-parameter settings.

Hyper-parameter	Value
Batch size	6
Maxlen	500
Learning rate	2e-5
Epochs	2

4.1 Sentiment Classification

In our dataset, there is total number of 21,053 reviews with Positive Sentiment, 1435 reviews with Negative Sentiment and 140 reviews have Neutral Sentiment. For building the model for SC, we set "Sentiment" as target variable and split the dataset into training set (80%) and validation set (20%). We then fine-tuned *distilbert-base-uncased* model with the settings mentioned in Table 4. The model get trained with validation accuracy 0.96 and 0.79 macro F1 score. The model is trained with a total of 66,955,779 parameters.

4.2 Product Recommendation

For building PR model, we set "Recommended IND" as target variable. It consists of a count of 18,527 reviews with Recommended IND value equal to 1 and 4,101 reviews having Recommended IND value equal to 0. We then split the dataset into training set (80%) and validation set (20%) and fine-tuned *distilbert-base-uncased* model with the settings mentioned in Table 4. Our model gives validation accuracy of 0.91 and 0.85 macro F1 score. The model is trained with a total of 66,955,010 parameters.

5. **Experimental Results**

5.1 Performance of proposed models against baseline models

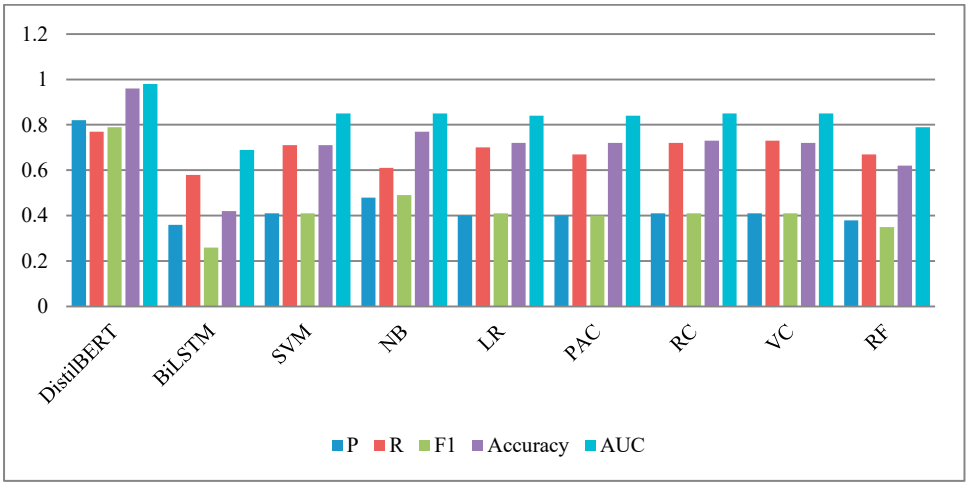
Table 5 and 6 represent the performance of proposed models using DistilBERT and various other supervised traditional algorithms trained in this study. The performance of each model is evaluated in terms of precision (P), recall (R), F1 score, accuracy and AUC score. It can be observed from both tables that transformer-based DistilBERT is providing exceptional performance with a very high precision, recall, macro F1 and AUC scores in comparison to conventional models for both sub-tasks, which declares the proposed models as highest quality models trained on imbalanced data. Along with this, we have also achieved very high accuracy in both cases. BiLSTM is the least performing model and showing highly skewed performance, this is because deep learning models are data hungry and here we have applied data under-sampling (most common technique to deal with imbalanced data) which leads to poor model training in this case. We have also tried oversampling for BiLSTM but there is no significant performance gain with it also. DistilBERT is giving remarkable results without application of any data resampling technique. Figure 7 graphically shows the performance of proposed models in comparison to baseline models.

Table 5. Performance of sentiment classification model as compared to traditional supervised models.

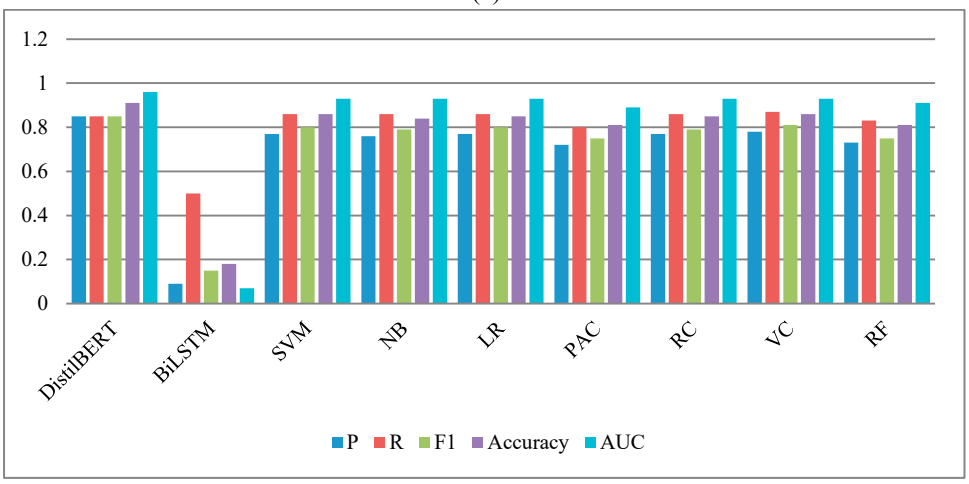
Model	P	R	F1	Accuracy	AUC
DistilBERT	**0.82**	**0.77**	**0.79**	**0.96**	**0.98**
BiLSTM	0.36	0.58	0.26	0.42	0.69
SVM	0.41	0.71	0.41	0.71	0.85
NB	0.48	0.61	0.49	0.77	0.85
LR	0.40	0.70	0.41	0.72	0.84
PAC	0.40	0.67	0.40	0.72	0.84
RC	0.41	0.72	0.41	0.73	0.85
VC	0.41	0.73	0.41	0.72	0.85
RF	0.38	0.67	0.35	0.62	0.79

Table 6. Performance of product recommendation model as compared to traditional supervised models.

Model	P	R	F1	Accuracy	AUC
DistilBERT	**0.85**	**0.85**	**0.85**	**0.91**	**0.96**
BiLSTM	0.09	0.50	0.15	0.18	0.07
SVM	0.77	0.86	0.80	0.86	0.93
NB	0.76	0.86	0.79	0.84	0.93
LR	0.77	0.86	0.80	0.85	0.93
PAC	0.72	0.80	0.75	0.81	0.89
RC	0.77	0.86	0.79	0.85	0.93
VC	0.78	0.87	0.81	0.86	0.93
RF	0.73	0.83	0.75	0.81	0.91



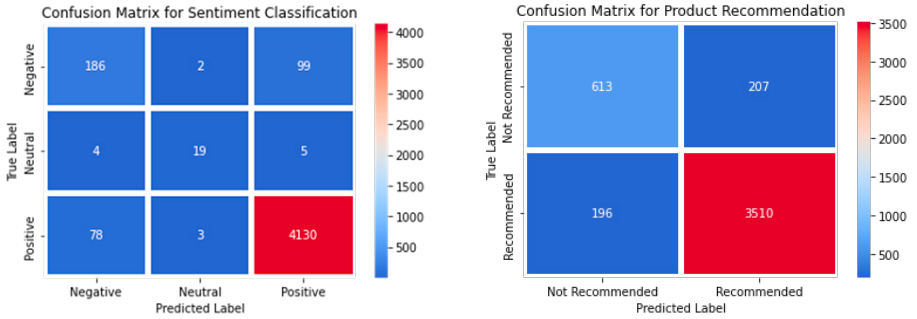
(a)



(b)

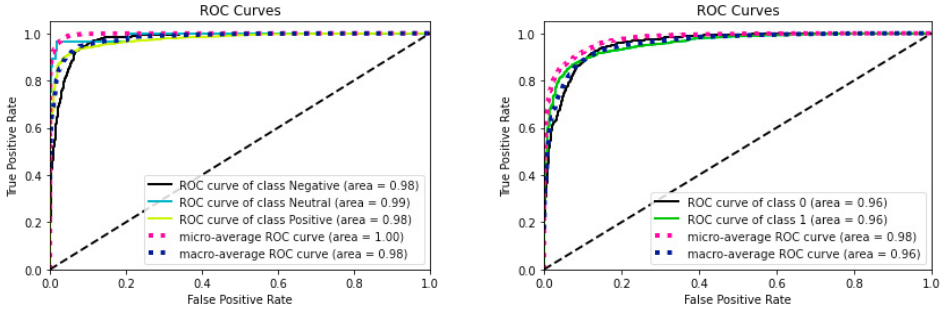
Fig. 7. (a) Performance comparison of proposed model over baseline models for SC; (b) Performance comparison of proposed model over baseline models for PR.

DistilBERT is showing significant improvement for minority class without compromising the result for majority class. It can also be verified from the confusion matrix obtained for both the proposed models as shown in Figure 8. Figure 9 represents the ROC curves for transformer based SC and PR respectively.



(a) (b)

Fig. 8. (a) Confusion matrix for SC; (b) Confusion matrix for PR.



(a) (b)

Fig. 9. (a) ROC curve for SC; (b) ROC curve for PR.

5.2 Comparison with existing studies

Table 7 presents a comparison between the proposed and the existing models for the selected dataset. We find that several SOTA studies in the literature are based on ML and DL algorithms. It can also be seen from the table that most of the previous work could not deal with the problem of data imbalance and achieved low recall, F1 and AUC scores. In this study, we have overcome this limitation with the proposed models and obtained significant improvement over these scores. Our models have also achieved the highest accuracy among other SOTA models.

Table 7. Comparison with SOTA studies on women's clothing e-commerce dataset.

Reference	Sub-task	Highest performing model	P	R	F1	Accuracy	AUC
Noor and Islam (2019) [27]	SC	SMO	0.809	-	-	0.81	-
Agarap (2020) [28]	SC	BiLSTM	0.58	0.55	0.56	0.92	-
Ours (2023)	SC	**DistilBERT**	**0.82**	**0.77**	**0.79**	**0.96**	**0.98**
Agarap (2020) [28]	PR	BiLSTM	0.81	0.79	0.80	0.88	0.79
Nawaz et al.(2021) [29]	PR	LSTM	-	0.71	0.57	0.92	0.71
Loukili et al. (2023) [30]	PR	LR	0.854	0.786	0.814	0.90	-
Ours (2023)	PR	**DistilBERT**	**0.85**	**0.85**	**0.85**	**0.91**	**0.96**

6. Conclusion and Future Scope

Sentiment analysis of online reviews is challenging for companies because of high volume, frequency and complex structure of data. Previous studies have utilized various ML and DL techniques to solve this problem. In this paper, we have proposed transformer based unsupervised learning approach employing DistilBERT and developed fine-tuned models for Sentiment Classification and Product Recommendation on women's clothing reviews dataset. We have achieved significant amount of performance gain over SOTA studies in terms of F1 scores which are 0.79 for SC and 0.85 for PR, AUC with values 0.98 for SC and 0.96 for PR and accuracy of 0.96 for SC and 0.91 for PR. We have also performed a comparative analysis of proposed models with several supervised baseline models which are developed in this study with application of data resampling to tackle the problem of data imbalance. We found that unlike traditional supervised ML and DL approaches, DistiBERT can understand the contextual information in the sentence and it has the natural ability to deal with imbalanced data. From this research, we can conclude that the application of transformer based PLMs can provide an efficient solution in case of imbalanced text sentiment analysis. This study can help companies to gain customer insight about their products and services. However, this research work is limited to the field of e-commerce and we would like to extend it in future for different domains and real-life applications like spam detection, fraud detection, disease detection etc., which generally consist of imbalanced data.

References

[1] Tan, K. L., Lee, C. P., and Lim, K. M. (2023) "A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research." *Applied Sciences* **13 (7)**, doi: 10.3390/app13074550.

[2] Cui, J., Wang, Z., Ho, S. B., and Cambria, E. (2023) " Survey on sentiment analysis: evolution of research methods and topics" *Artificial Intelligence Review* **56 (8)**: 8469-8510, doi: 10.1007/s10462-022-10386-z.

[3] Baccianella, S., Esuli, A., and Sebastiani, F. (2010) "SENTIWORDNET 3.0: An enhanced lexical resource for sentiment analysis and opinion mining." In *Proeedings of 7th Inernational. Conference on Language Resources,. Eval. Lr*.: 2200–2204.

[4] Strapparava, C. and Valitutti, A. (2004) "WordNet-Affect: An affective extension of WordNet." In *Proceedings of 4th International Conference on Language. Resources, Eval. Lr*: 1083–1086.

[5] Pinto, D., McCallum, A.,Wei, X. and Crof, W. B. (2003) "Table extraction using conditional random fields." In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (SIGIR '03)* :235–242, doi: 10.1145/860435.860479.

[6] Rezwanul, M., Ali, A., and Rahman, A. (2017) "Sentiment Analysis on Twitter Data using KNN and SVM." *International Journal of Adanced. Computer Science and Applications* **8 (6)**:19–25, doi: 10.14569/ijacsa.2017.080603.

[7] Soni, S. and Sharaff, A. (2015) "Sentiment Analysis of Customer Reviews based on Hidden Markov Model." In *Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015)*: 1–5, doi: 10.1145/2743065.2743077.

[8] Rodrigues, A. P., Fernandes, R., Aakash, A., Abhishek, B., Shetty, A., Atul, K., Lakshmanna, K., and Shafi, R. M. (2022) "Real-Time Twitter Spam Detection and Sentiment Analysis using Machine Learning and Deep Learning Techniques." *Computational Intelligence and Neuroscience* vol. 2022, pp. 1-14, doi: 10.1155/2022/5211949.

[9] Rahat, A. M., Kahir, A., and Masum, A. K. M. (2019) "Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset." In *Proceedings of 8th International Conference System Modeling and Advancement in Research Trends (SMART)*: 266–270, doi: 10.1109/SMART46866.2019.9117512.

[10] Dang, N. C., Moreno-García, M. N., and Prieta, F. D. l. (2020) "Sentiment analysis based on deep learning: A comparative study." *Electronics* **9 (3)**: 1–29, doi: 10.3390/electronics9030483.

[11] Pandey, A. C., Rajpoot, D. S., and Saraswat, M. (2017) "Twitter sentiment analysis using hybrid cuckoo search method." *Information Processing Management,* **53 (4)**: 764–779, doi: https://doi.org/10.1016/j.ipm.2017.02.004.

[12] Keenan, M. J. S. (2020) "Advanced Positioning, Flow, and Sentiment Analysis in Commodity Markets: Bridging Fundamental and Technical Analysis", *2nd Edition. Wiley*.

[13] Yoon, J. H. and Jang, B. (2023) "Evolution of Deep Learning-Based Sequential Recommender Systems: From Current Trends to New Perspectives." *IEEE Access* vol 11, 54265-54279, doi: 10.1109/ACCESS.2023.3281981.

[14] Aishwarya, Wadhwa, P., and Singh, P. (2020) "A New Sentiment Analysis based Application for Analyzing Reviews of Web Series and Movies of Different Genres." In *Proceedings of 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India,* 390-396, doi: 10.1109/Confluence47617.2020.9058137.

[15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez A. N., Kaiser, L., and Polosukhin, I. (2017) "Attention is all you need." *Advances in Neural Information Processing Systems*: 6000–6010, doi: 10.5555/3295222.3295349.

[16] Min, B. et al. (2023) "Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey." *ACM Computing Surveys* **56 (2)**: 1-40, doi: 10.1145/3605943.

[17] Malte, A. and Ratadiya, P. (2019) "Evolution of transfer learning in natural language processing." *[Online]. Available: http://arxiv.org/abs/1910.07370.*

[18] López, V., Fernández, A., García, S., Palade, V., and Herrera, F. (2013) "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics." *Information Sciences (Ny)* **250**: 113–141, doi: 10.1016/j.ins.2013.07.007.

[19] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019) "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 4171–4186*, doi: 10.18653/V1/N19-1423.

[20] Victor, S., Debut, L., Chaumond, J., and Wolf, T. (2020) "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." *[Online]. Available: http://arxiv.org/abs/1910.01108.*

[21] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, A., Zhou, Y., and Liu, P. J. (2019) "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." *Journal of Machine. Learning Research* **21**: 1–67, doi: 10.48550/arxiv.1910.10683.

[22] Clark, K., Luong, M. T., Le, Q. V., and Manning, C. D. (2020) "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators." In *Proceedings of International Conference of Learning Representations.*. doi: 10.48550/arxiv.2003.10555.

[23] Taneja, K. and Vashishtha, J. (2022) "Comparison of Transfer Learning and Traditional Machine Learning Approach for Text Classification." *Proceedings of the 16th INDIACom; 2022 9th International Conference on Computing for Sustainable Global Development*: 397–402, doi: 10.23919/INDIACom54597.2022.9763279.

[24] Women's E-Commerce Clothing Reviews | Kaggle (2018). *https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews* (accessed Mar. 09, 2022).

[25] Hochreiter, S. and Schmidhuber, J. (1997) "Long Short-Term Memory." *Neural Computing.* **9 (8)**: 1735–1780, doi: 10.1162/neco.1997.9.8.1735.

[26] Sharma, S., Diwakar, M., Joshi, K., Singh, P., Akram, S. V., and Gehlot, A. (2022) "A Critical Review on Sentiment Analysis Techniques." 3rd International Conference on Intelligent Engineering and Management (ICIEM), London, United Kingdom, pp. 741-746, doi: 10.1109/ICIEM54221.2022.9853140.

[27] Noor, A. and Islam, M. (2019) "Sentiment Analysis for Women's E-commerce Reviews using Machine Learning Algorithms." *10th International Conference on Computing, Communication and Networking Technologies (ICCCNT),* Kanpur, India, pp. 1-6, doi: 10.1109/ICCCNT45670.2019.8944436.

[28] Agarap, A. F. (2020) "Statistical Analysis on E-Commerce Reviews, with Sentiment Classification using Bidirectional Recurrent Neural Network (RNN)." *[Online]. Available: http://arxiv.org/abs/1805.03687.*

[29] Nawaz, Z., Zhao, C., Nawaz, F., Safeer, A. A., and Irshad, W. (2021) "Role of artificial neural networks techniques in development of market intelligence: a study of sentiment analysis of ewom of a women's clothing company." Journal of Theoretical and Applied Electronic Commerce Research **16(5)**: 1862–1876, 2021, doi: 10.3390/jtaer16050104.

[30] Mehta, K. and Panda, S. P. (2022) "Sentiment Analysis on E-Commerce Apparels using Convolutional Neural Network." International Journal of Computing **21 (2)**: 234–241, doi: 10.47839/ijc.21.2.2592.

[31] Loukili, M., Messaoudi, F., and Ghazi, M. El. (2023) "Sentiment Analysis of Product Reviews for E-Commerce Recommendation based on Machine Learning," International Journal of Advances in Soft Computing and its Applications **15 (1)**: 1–13, doi: 10.15849/IJASCA.230320.01.

[32] Taneja, K., Vashishtha, J., Ratnoo, S. (2023) "Efficient Deep Pre-trained Sentence Embedding Model for Similarity Search." International Journal of Computer Information Systems and Industrial Management Applications **15**: 605-615.

[33] Maiya, A. S. (2020) "ktrain: A Low-Code Library for Augmented Machine Learning." Accessed: Jul. 14, 2021. *[Online]. Available: https://arxiv.org/abs/2004.10703v4.*

[34] GitHub - laxmimerit/preprocess_kgptalkie (2020). https://github.com/laxmimerit/preprocess_kgptalkie (accessed Mar. 10, 2022).