

Multi-label Sentiment Analysis of Product Reviews of Online Shop

Animesh Chandra Roy, Ahasan Kabir, Zaima Sartaj Taheri and Md. Jahedul Alam Rifat

Chittagong University of Engineering & Technology
Chattogram-4349, Bangladesh
animeshroycse@gmail.com
ahasankabir146@gmail.com
u1704054@student.cuet.ac.bd
jahedulalamrifat@gmail.com

Abstract. Online Shopping has become very popular nowadays. It tends to share user experiences of buying products or dealing with the seller by posting reviews. So millions of reviews are being generated daily. It can be difficult for a new customer of that particular product to read all of those reviews and decide whether or not to purchase. In this situation, an overall sentiment(s) of the whole review might help them. Also, people being more creative sometimes post sarcastic or ironic statements. This may mislead other buyers. The Majority of the previous research work regarding product sentiment analysis was confined to two to three sentiments only. Also, a review may express several emotions at once. By doing binary classification we may miss other emotions present alongside the predicted one. So we have proposed a binary classifier model to separate the sarcastic reviews and then a multi-label classifier model to detect the emotions present in a particular review. We have applied several methods for multi-label classification naming Binary Relevance, Classifier Chain, Label Power-set on up to four different classifiers. Among them, the OnevsRest classifier along with the support vector classifier as an estimator performed better than the other methods. We have also trained and tested a few binary classifiers for sarcasm detection and got almost the same accuracy of 93.37% and 93.92% for Logistic Regression and Support Vector Classifier.

Keywords: Multi-label classification, Sentiment analysis, Product reviews, Sarcastic reviews

1 Introduction

We are now living in the era of the internet. The internet is used for everything from minor to major tasks. From buying groceries, doing shopping to work from home, everything could be possible by dint of the internet. The importance of online shopping is growing frequently in the modern society. This growth in e-commerce placed a strong emphasis on the requirements and client preferences,

giving rise to a critical component of internet purchasing known as "User Reviews". User reviews are opinions and judgements expressed by customers about a product that help other buyers make purchasing decisions. Product reviews can be beneficial to both consumers and manufacturers. This makes it more challenging to analyze and extract information from existing reviews.

People are becoming dependent on online purchasing. They are using different sites like Amazon, Flipkart, Daraz etc to purchase their products. Before making a purchase on online, people frequently read through the product's reviews and ratings. The buyer may be mislead as the overall sentiment that amazon gives is a collective one. Moreover, that much review can't be read or it will be very time consuming. But if he does not read any review then there might be a chance not getting the right product. That is why, to analyze the product reviews we propose a model that can evaluate a multi-label sentimental output of that review. Thus, a new buyer can easily get the message about the product that he is intended to buy. People being more creative, sometimes post comments or review in a positive manner although they mean almost the opposite of that. This situation is termed as irony or sarcasm. An example of that can be, "Your product is so good that I will never think of buying another piece". Here they might be mentioning good but actually meant that the product was very bad. We will be trying to detect such texts i.e sarcastic reviews. Along with that, A single review may not be expressed with just one emotion. It can express multiple emotion at once. An example of that can be, "I am so happy using your product. I will surely recommend it to my relatives." This can be extracted both happy and admiration in case of sentiment. Our model will try to extract all the sentiment available in a sentence after removing the sarcastic reviews.

2 Related Works

Sentiment analysis is like context mining which analyses subjective information in the original text. At first, by sentiment analysis, it was supposed to be classifying texts into binary classes i.e. Positive or negative. After this, researchers found out that, it's getting difficult to analyze text and express emotion in just two classes. Then it became a multi-class classification where a sentence can be any of the one class among several types of emotions. By now, it is well examined that, it's difficult to confine a text into just one class of emotion. A text can express multiple emotions at a time. This was termed a multi-label classification. Here, we will go through some research that is already done and closely related to ours.

Senticnet5 was used in the training phase to detect emotions by Cambria et al. [1]. Since a limited number of tweets are available in this dataset, if a numerous amount of words are encountered then it faced some difficulties. Additionally, Bouazizi et al. [2] introduced a scalable method for classifying sentiments in tweets. They applied 7 various sentiment classes, each of which consists of three pairings of various feelings and one impartial class. But a particular can express more than one emotion at a time.

Sarcasm in the Amazon Alexa Sample Set was investigated by Avinash et al. [3]. Nevertheless, they omitted to acknowledge their dataset. They used SentiWordNet and TextBlob to eliminate noise and extract features. They used three classifier approaches to get an accuracy of 70.95%. Another research group Sahil Jain et al. [4] also proposed a methodology applying Naive Bayes, Support Vector Machine and Neural Network in Amazon product reviews. They have considered the ratings along with the review text in determining sarcasm. Joshi et al. [5] researched the importance of using Word Embedding-based features in terms of detecting sarcasm in text. They have shown their concern by taking GloVe, Word2Vec, and other word embeddings. They finally showed Classification reports on them and concluded that word embedding-based features can play a vital role in detecting sarastic sentences.

Singla et al. [6] applied the three well-known classifiers Support Vector Machine, Naive Bayes, and Decision Tree to roughly 4,000 000 reviews in order to categorize positive and negative ratings. Rajkumar et al. [7] Satuluri Vanaja and Meena Belwal [8] analyzed two Machine Language approaches for performing Sentiment Analysis on reviews of a specific product. Islam et al. [9] suggested a multi-label classification paradigm, categorizing strategies as problem transformation or algorithm adaptation. They also used Senticnet5 to improve the accuracy. Yang et al. [10] used sentiment analysis on customer reviews where they have over 100,000 orders of magnitude, which are applied to Chinese sentiment analysis. A deep learning method was introduced to analyse emotions by Xiao et al. [11]. In their work, LSTM had better performance than Naive Bayes and the logistic regression method. With single-attention and multiple-attention networks, Ameer et al. [12] studied the usage of LSTMs with fine-tuned Transformer Networks, reaching an accuracy of 62.4% for English and 45.6% for Chinese in multi-label emotion categorization.. Combining transformers with handcrafted features has demonstrated enhanced performance in various tasks. For example, Uto et al. [13] achieved an accuracy increase from 71% to 80% in essay scoring, and Kumar et al. [14] showcased accuracies of 78% versus 80% in sarcasm detection.

We have been to some recent research regarding sentiment analysis. Sentiment Analysis has been a very popular topic among the researchers worldwide but a few of them that were related to our work. Our objective is to perform multi-label sentiment analysis of product reviews after filtering the sarcastic reviews. The contributions of our work are summarized below:

- Developing a properly Labelled Product Review Dataset.
- Multi-Label Classification Model to extract sentiment of product reviews.
- Building a model to classify sarcastic text in product reviews.

3 Methodology

We have implemented the classification task maintaining some specific steps. In this particular section, we will elaborately discuss each of the steps we have done in terms of completing our work.

3.1 Overview of Framework

The framework for our Proposed work is depicted in Figure 1. This shows the two separate modules of our thesis. One is Binary Classification and the other one is Multi-Label Sentiment Analysis. In the Multi-Label Sentiment Analysis module we have applied and analyzed various methods to accurately label our emotions naming sarcastic, happy, sad, neutral, admiration, and angry.

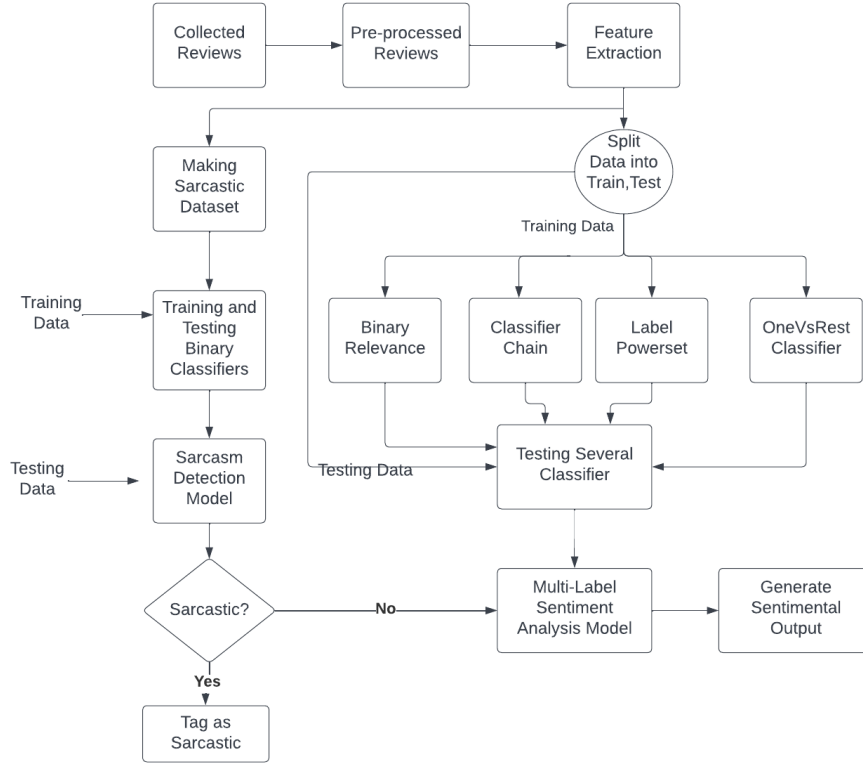


Fig. 1. System Architecture

3.2 Data Pre-processing

As machines don't understand our national languages, we have to pre-process those reviews and extract features from them. After that, we have to split the whole dataset into parts for training and testing.

1. **Pre-processing the reviews:** After collecting the dataset and labeling, we have to pre-process the reviews for further use. For this, we have removed ev-

ery non-alphabetic character from the review texts. Punctuations and URLs also being removed, as they won't enhance the learning.

After pre-processing the text reviews, we have implemented Word Tokeniza-

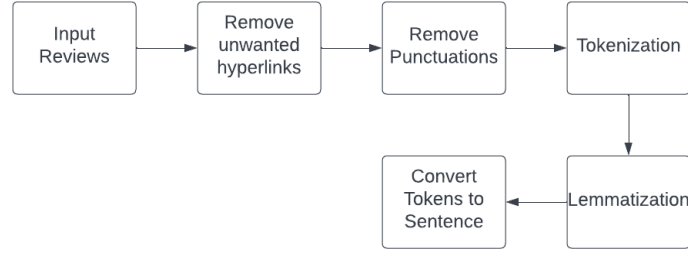


Fig. 2. Pre-processing Steps

tion. Then we performed Lemmatization and stopwords were also removed from the tokens.

After the completion of the pre-processing task, we made a separate dataset containing only the sarcastic and an equal amount of non-sarcastic review-texts. This process is depicted in Figure 2.

2. **Tf-idf vectorization, Glove, CountVect** We have analyzed performing Tf-idf, CountVect, and Glove simultaneously. Among all of them, Glove outperformed the rest of the two in case of the sarcasm detection process. Also, We have monitored the performance of Tf-idf and Word2Vec in terms of implementing multi-label sentiment analysis.
3. **Applying Problem Transformation Method:** As multi-label classification has no specific classifier to directly detect emotion, we have to convert the task to Binary Classification or Multi-class classification problem. For this task, we used some problem transformation method to apply. This methods do this tough task on behalf of us and bring us our desired labels as output.

After pre-processing and removing the unwanted parts from the text, we have applied tokenization as the following sample shown at Table 1.

3.3 Implementation of Multi-Label Classifier


There are two popular ways to implement a multi-label classifier. These are Problem Transformation and Adaptation. In the case of the problem transformation method, we have applied three techniques shown in Figures 3, 4, and 5. They are Binary Relevance, Classifier Chain and Label Powerset. In terms of Binary Relevance, the dataset is subdivided into six different binary classification problems. After the application, the results are merged to get the multi-label

Table 1. Pre-processing steps

Steps	Description
Input Text	Amazon batteries-can't be used in many appliances
Remove Unwanted Symbols	amazon batteries can not be used in many appliances
Tokenization	['amazon', 'batteries', 'can', 'not', 'be', 'used', 'in',...]
Lemmatization	['amazon', 'battery', 'can', 'not', 'be', 'use', 'in',...]
Pre-processed Text	amazon battery can not be use in many appliance

output. The Classifier Chain technique applies by shifting the classes one by one. and continues to each six classes. Label powerset changes the problem into a multi-class problem and predicts the result. After that, all of the results are merged to get the final output. We have also applied OneVsRestClassifier technique. In this case, one particular classification model is compared to the rest of the classifiers. We have used the Jaccard_score function for the calculation of accuracy in terms of the OnevsRest Classifier and the default accuracy_score function from sklearn for the rest of the classifiers.

X	Y_1	Y_2	Y_3	Y_4
$X^{(1)}$	0	0	0	1
$X^{(2)}$	1	0	0	0
$X^{(3)}$	1	0	0	1
$X^{(4)}$	0	1	0	0
$X^{(5)}$	0	1	1	0



X	Y_1	X	Y_2	X	Y_3
$X^{(1)}$	0	$X^{(1)}$	0	$X^{(1)}$	0
$X^{(2)}$	1	$X^{(2)}$	0	$X^{(2)}$	0
$X^{(3)}$	1	$X^{(3)}$	0	$X^{(3)}$	0
$X^{(4)}$	0	$X^{(4)}$	1	$X^{(4)}$	0
$X^{(5)}$	0	$X^{(5)}$	1	$X^{(5)}$	1

Fig. 3. Binary Relevance

X	Class1
X1	0
X2	0
X3	1

X	Class2
X1	0
X2	0
X3	0

X	Class3
X1	1
X2	0
X3	1

Fig. 4. Classifier Chain

X	Class1	Class2	Class3
X1	0	0	1
X2	0	0	0
X3	1	0	1

X	Class
X1	1
X2	2
X3	3

Fig. 5. Label Powerset

3.4 Implementation of Sarcasm Detection Classifier

We have tried different types of available vectorizers for feature extraction. Namingly, Senticnet5, CountVectorizer, Tf-Idf Vectorizer, TextBlob, Glove. Among them, we have found better performance by using Glove, CountVectorizer and Tf-idf vectorizer. Implementation of CountVectorizer, TF-IDF Vectorizer and Glove Training Binary Classifiers. We have used multiple Binary Classifier models to train and detect sarcastic reviews.

4 Result and Discussion

4.1 Dataset Description

Unfortunately, there is no standard and well-labeled dataset for multi-label emotion categorization of product reviews. So we collected a dataset of product reviews from the well renowned website for the data scientists, Kaggle. This dataset consists of a portion of user reviews from the Amazon Website with the ratings and other related information. So for making it sentiment analysis worthy we had to manually input the sentiments for each of the reviews. As there were not any pre-established dataset for multi-label sentiment analysis. This is the dataset used for multi-label sentiment analysis. It consists of 4863 rows and nine columns for our output labels. we have also used a portion of this dataset for binary classification in detecting sarcastic reviews. That dataset consists of 903 rows and two columns.

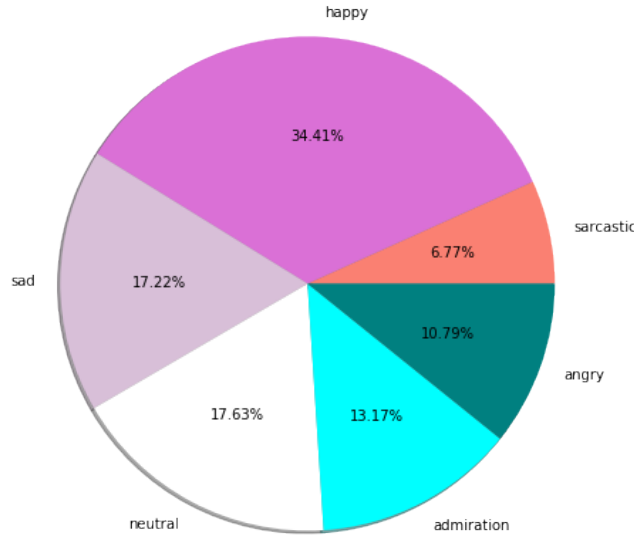
Table 2 shows some example of our dataset and Figure 6 shows the percentages of data of each of the six levels. Its a bit imbalanced dataset in terms of the label count.

4.2 Problem Transformation Methods

We have analyzed our Binary Relevance model performance with four different Classifiers. Then we used the Classifier Chain method which shows better performance in case of maintaining label correlation. Also, it's more accurate than the binary relevance method. We have also implemented the Label Powerset method. Given a classification problem with N possible solutions, a OnevsRest solution

Table 2. Dataset Overview

Reviews	Sarcastic	Happy	Sad	Neutral	Admiration	Angry
We loved the Echo..	0	1	0	0	0	0
My first tablet was..	0	0	1	0	0	1
Worst batteries ever..	0	0	1	0	0	1
Best buy associate..	0	1	0	0	1	0
So fast delivery took..	1	0	1	0	0	1
I am enjoying my kir..	0	1	0	0	1	0
The kindle over all..	0	0	0	1	0	0
Bought this for my..	0	1	0	0	0	0
Great electronic, ve..	0	1	0	0	0	0
The features that..	0	0	0	1	0	0

**Fig. 6.** Dataset overview for Multi-Label Sentiment Analysis

consists of N distinct binary classifiers—one binary classifier for each potential outcome. Each binary classifier is trained to respond to a different classification problem when the model goes through a sequence of them during training.

Table 3 depicts the accuracy we could get using the above mentioned methods. From the table we can see that OnevsRest gives more accurate results than the other three models.

4.3 Sarcasm Detection

we have tested several Binary Classifiers to detect sarcasm in product reviews. We have applied both CountVect and Tf-idf Vectorizer and analyzed there respective outputs.

Table 3. Accuracy and Loss For Problem Transformation Methods

Method	Classifier Name	Accuracy	Hamming Loss
Binary Relevance	Naive Bayes	41%	14.08%
	Random Forest	47%	13.17%
	Support Vector	48%	12.30%
	Logistic Regression	46%	13.78%
Classifier Chain	Naive Bayes	45%	14.12%
	Random Forest	54%	13.27%
	Support Vector	57%	12.31%
	Logistic Regression	55%	12.50%
Label Powerset	Naive Bayes	49%	15.10%
	Random Forest	55%	13.44%
	Support Vector	58%	12.70%
	Logistic Regression	56%	13.07%
OnevsRest	Naive Bayes	54%	15.10%
	Random Forest	59%	12.34%
	Support Vector	65%	11.40%
	Logistic Regression	57%	13.57%

Table 4 depicts the resulting accuracy we have managed to get in terms of detecting sarcasm in product review text. Here, Support Vector classifier gives better accuracy than Random Forest and Logistic Regression.

Table 4. Accuracy using Tf-idf Vect

Classifier	Vectorizer	Accuracy	Log Loss
Random Forest	Count Vect	84.73%	37.59%
	Tf-idf Vect	87.08%	31.42%
Logistic Regression	Count Vect	8.17%	33.08%
	Tf-idf Vect	93.37%	29.12%
Support Vector	Count Vect	85.73%	33.49%
	Tf-idf Vect	93.92%	27.04%

4.4 Testing Real Data

Among all of the classifiers, SVC and Logistic Regression classifier acquired highest accuracy. So we can use any of the model for predicting real reviews. Following Tables 5 and 6 shows some of examples that our model predicted.

Table 5. Predicting Sarcasm

Input Text	Prediction	Expected	Correct?
A women needs a man just like..	sarc.	sarc.	yes
Your service is so good that..	sarc.	sarc.	yes
My mother liked your product..	sarc.	non-sarc.	no
Your so called test service..	sarc.	sarc.	yes

Table 6. Predicting Sentiment

Input Text	Prediction
The product was excellent, love it	happy, admiration
The product is not so good	neutral
None of the batteries that i opened worked fine	sad, angry
I will consider buying frequently from them	happy, admiration

5 Evaluation of Performance

5.1 Multi-Label Classifier

We have applied several methods and several classifiers. Among all of them, we had the most accuracy using OneVsRest Classifier with SVC. Table 7 shows the classification accuracy of SVC.

Table 7. classification accuracy of SVC

	Precision	Recall	f1-score	Support
Sarcastic	0.91	0.62	0.74	97
Happy	0.81	0.85	0.83	506
Sad	0.90	0.78	0.83	234
Neutral	0.57	0.58	0.58	233
Admiration	0.55	0.47	0.51	203
Angry	0.79	0.80	0.79	147
Micro avg	0.75	0.72	0.73	1420
Macro avg	0.75	0.68	0.71	1420
Weighted avg	0.75	0.72	0.73	1420
Samples avg	0.74	0.72	0.72	1420

5.2 Sarcasm Detector

In the case of binary classification for detecting sarcasm in product review texts, we have gone through several vectorizers for feature extraction. Among all of them, Support Vector Classifier and Logistic Regression got quite similar accuracy of 93.92% and 93.37% accuracy respectively. Figures 7, 8 and 9 show the confusion matrix for sarcasm detection. For Logistic Regression and SVC, we got

higher True positive and True negative values of 77, 92, and 79, 91 respectively. But in the case of Random Forest, we got a higher false negative value of 26.

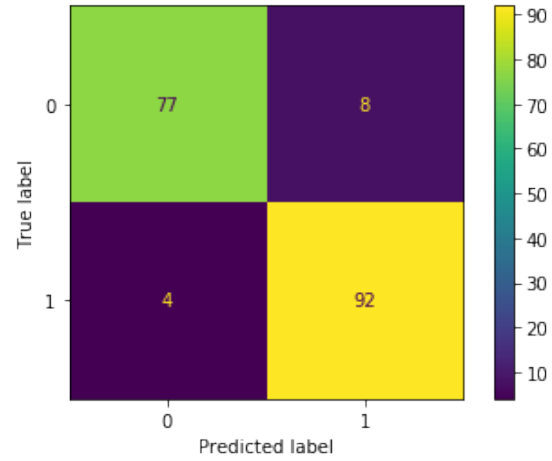


Fig. 7. Logistic Regression

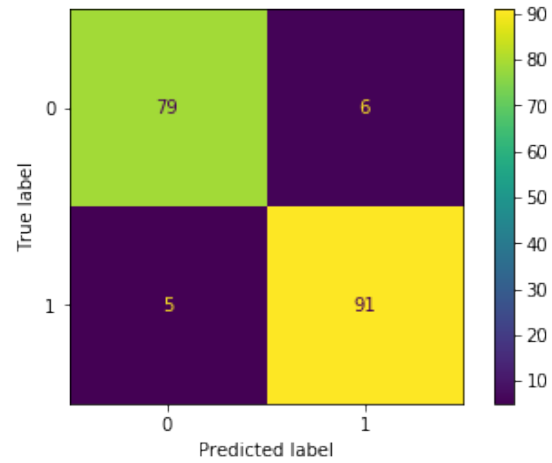
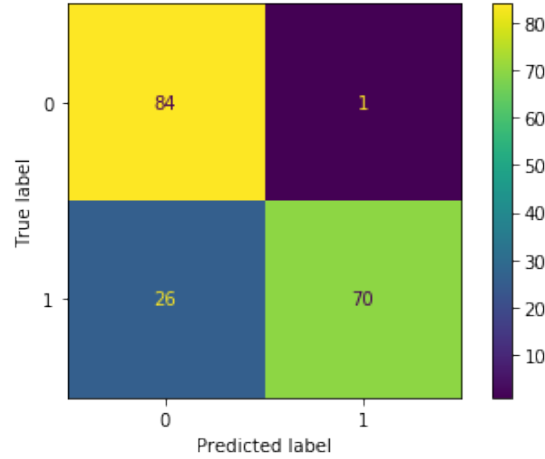


Fig. 8. SVC

**Fig. 9.** Random Forest

6 Conclusion

One of the most discussed aspects of LP is sentiment analysis. Online marketing press media have evolved to include tools that allow consumers to provide input on the products being marketed. Extracting information, especially sentiment, from product reviews is necessary for both customers and producers to understand market response and take suitable actions on the product based on the retrieved information. A multi-label emotion classification approach to classify product reviews given by the users is proposed in this work. We have applied multiple classifiers and extracted emotions like sarcastic, happy, sad, neutral, admiration, and angry. Along with that, we have separately considered a binary classification model to train and predict sarcasm in product reviews.

We have gone through the analysis portion of our proposed work. After analyzing several methods and classifiers, we can say that, in determining both module, Tf-idf vectorizer showed outstanding performance in feature extraction. In particular for multi-label sentiment analysis, OnevsRest Classifier along with SVC model have the prediction rate. In terms of Sarcasm Detection, All of the classifiers we have tested were close in the accuracy result. But Logistic Regression and Support Vector Classifier have accuracy above 93%, which is an improved result then the available models.

Although we have managed to get better accuracy but that also can be improved in near future. This work may be expanded upon to train on a bigger, more evenly distributed dataset. Punctuations, Emoticons can also play a significant role in case of detecting emotion more accurately. Moreover, this dataset was labeled by only one person. So, Multiple person can be engaged for labeling and cross examine in order to get more precise and valid dataset.

References

1. Cambria, E., Hussain, A.: Senticnet. In: Sentic Computing, pp. 23–71. Springer (2015)
2. Bouazizi, M., Ohtsuki, T.: A pattern-based approach for multi-class sentiment analysis in twitter. *IEEE Access* 5, 20617–20639 (2017)
3. Pandey, A.C., Seth, S.R., Varshney, M.: Sarcasm detection of amazon alexa sample set. In: *Advances in Signal Processing and Communication*, pp. 559–564. Springer (2019)
4. Jain, S., Ranjan, A., Baviskar, D.: Sarcasm detection in amazon product reviews. *International Journal of Computer Science and Information Technologies* 9(3) (2018)
5. Joshi, A., Tripathi, V., Patel, K., Bhattacharyya, P., Carman, M.: Are word embedding-based features useful for sarcasm detection? *arXiv preprint arXiv:1610.00883* (2016)
6. Singla, Z., Randhawa, S., Jain, S.: Sentiment analysis of customer product reviews using machine learning. In: *2017 international conference on intelligent computing and control (I2C2)*. pp. 1–5. IEEE (2017)
7. Jagdale, R.S., Shirsat, V.S., Deshmukh, S.N.: Sentiment analysis on product reviews using machine learning techniques. In: *Cognitive Informatics and Soft Computing*, pp. 639–647. Springer (2019)
8. Rintyarna, B.S., Sarno, R., Fatichah, C.: Semantic features for optimizing supervised approach of sentiment analysis on product reviews. *Computers* 8 (2019)
9. Islam, S., Roy, A.C., Arefin, M.S., Afroz, S.: Multi-label emotion classification of tweets using machine learning. In: *Proceedings of the International Conference on Big Data, IoT, and Machine Learning: BIM 2021*. pp. 705–722. Springer (2022)
10. Yang, L., Li, Y., Wang, J., Sherratt, R.S.: Sentiment analysis for e-commerce product reviews in chinese based on sentiment lexicon and deep learning. *IEEE Access* 8, 23522–23530 (2020)
11. Xiao, S., Wang, H., Ling, Z., Wang, L., Tang, Z.: Sentiment analysis for product reviews based on deep learning. In: *Journal of Physics: Conference Series*. vol. 1651, p. 012103. IOP Publishing (2020)
12. Ameer, I., Bölücü, N., Siddiqui, M.H.F., Can, B., Sidorov, G., Gelbukh, A.: Multi-label emotion classification in texts using transfer learning. *Expert Systems with Applications* 213, 118534 (2023)
13. Uto, M., Xie, Y., Ueno, M.: Neural automated essay scoring incorporating hand-crafted features. In: *Proceedings of the 28th International Conference on Computational Linguistics*. pp. 6077–6088 (2020)
14. Kumar, A., Narapareddy, V.T., Gupta, P., Srikanth, V.A., Neti, L.B.M., Malapati, A.: Adversarial and auxiliary features-aware bert for sarcasm detection. In: *Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD)*. pp. 163–170 (2021)