Abstract

This study suggests an information retrieval-based method for identifying false online platform reviews. To preserve the credibility of user-generated content in light of the rise in false reviews, it is crucial to create an efficient detection system. In this work, we gathered a dataset of reviews from several internet sources and utilised information retrieval techniques to find patterns and traits that distinguish real reviews from false reviews

This paper proposes a solution for detecting fake online reviews using a combination of machine learning algorithms and natural language processing techniques. The proposed method involves extracting additional features such as sentiment, verified purchases, ratings, product category, and overall score to gain insights into the authenticity of reviews. The classifiers used for classification include Naive Bayes, SVM, Random Forest, Decision Trees, and Logistic Regression, while advanced natural language processing techniques such as LSTMs, transformers, and the BERT model are utilized to improve accuracy. The proposed solution aims to provide a reliable and accurate approach to address the problem of opinion spamming in online platforms. The performance of the proposed solution is evaluated using various metrics such as accuracy, precision, recall, and F1 score. The results demonstrate the effectiveness of the proposed solution in detecting fake reviews and providing trustworthy user-generated content for consumers.

~~.Online platforms may identify phoney reviews with the help of the suggested fake review detection system, which employs an information retrieval strategy. This increases the legitimacy of user-generated material.~~

~~Lastly, there are occasions when individuals make statements that are in opposition to one another, making it hard to understand the nature of opinion. A negative review may conceal a positive impression. Additionally, there may be mixed reviews of the product at times. A person's words and actions can be significantly influenced by their emotions. combining a positive and negative comment with the same emoji. Finding reviews that are not genuine or that are used to influence consumers' opinions becomes even more difficult after these difficulties.~~

~~ProblemSet~~

~~In order to promote their products, people write unworthy positive reviews. In some instances, maliciously negative reviews of other (competitive) products are written to hurt their reputations. The main test is that a word can be both positive and negative in various settings. For instance, a positive opinion is expressed when the battery life of a product is described as "long," while a negative opinion is expressed when the start time is described as "long." Another challenge is that people don't always express opinions the same way.~~

Introduction

In the modern world, online reviews are crucial since they have a big impact on consumers' purchase decisions. But not all reviews are true, and the prevalence of phoney reviews has raised severe issues for both consumers and companies. Fake reviews have the potential to mislead customers, harm a company's reputation, and ultimately reduce sales. Determining fraudulent reviews and removing them from internet sites are therefore vital.

In recent years, the detection of bogus reviews has demonstrated encouraging results when using machine learning algorithms and natural language processing approaches. Due to the continually changing methods used by spammers to manufacture phoney reviews, it is still difficult to identify bogus reviews.

The BERT (Bidirectional Encoder Representations from Transformers) model and information retrieval techniques are used in this research study to provide a method for identifying bogus reviews. The suggested approach entails the extraction of supplementary attributes that reveal information about the veracity of the evaluations. Sentiment, confirmed purchases, ratings, product category, and total score are among the retrieved characteristics. Patterns that distinguish between the two types of reviews may be found by comparing these characteristics between real and false reviews.

Several classification methods, including Naive Bayes, SVM, Random Forest, Decision Trees, and Logistic Regression, are used to categorise reviews. Additionally, to increase the models' accuracy, cutting-edge NLP methods including LSTMs, transformers, and recurrent neural networks are used.

The suggested solution seeks to solve the issue of opinion spamming on online platforms by offering a dependable and accurate method to identify bogus evaluations. The remaining section of this research article outlines the suggested technique and provides experimental data and analysis to assess the effectiveness of the suggested remedy.

Methodology

Our proposed solution for detecting fake reviews combined machine learning algorithms and natural language processing techniques. The methodology involved the following steps:

Extracting additional features: We extracted features such as sentiment, verified purchases, ratings, product category, and overall score to gain insights into the authenticity of the reviews. By comparing these features between genuine and fake reviews, we identified patterns that distinguish between the two types of reviews.

Classification algorithms: We used Naive Bayes, SVM, Random Forest, Decision Trees, and Logistic Regression classifiers to perform the classification. These classifiers were trained on the preprocessed data and the additional features extracted earlier. We used techniques like cross-validation and hyperparameter tuning to improve the accuracy of the models.

Advanced natural language processing techniques: We also used advanced techniques like LSTMs, transformers, BERT (Bidirectional Encoder Representations from Transformers) model, and recurrent neural networks to capture the context and semantics of the text data and provide more accurate predictions.

Model evaluation: We evaluated the performance of the models using metrics such as accuracy, precision, recall, and F1 score. We also performed a comparative analysis of the different models to identify the best-performing model for detecting fake reviews.

The proposed methodology aimed to achieve high accuracy in detecting fake reviews by using a combination of features and classifiers. By using machine learning and natural language processing techniques, we provided a reliable solution for addressing the problem of opinion spamming in online websites. Specifically, we aimed to train a precise cumulative model by using both CNN and ANN, which could provide more accurate predictions than traditional machine learning models. Overall, the proposed method offered an effective approach to detecting fake reviews and providing trustworthy user-generated content for consumers.

Literature Review

Since 2007, the study of fake review detection has been conducted through review spamming analysis . The case of Amazon was looked at in this study, and the authors came to the conclusion that manually labeling fake reviews can be difficult because fake reviewers may carefully craft their reviews to make them more trustworthy for other users. As a result, they suggested using duplicates or nearly duplicates as spam to create a model that could identify fake reviews . Additionally, research on distributional footprints has demonstrated a link between distribution anomalies and deceptive hotel and Amazon product reviews. Some of the links we have referred from are :

https://ieeexplore.ieee.org/document/8335018
→This paper proposes a CNN-based approach for fake review detection and compares its performance with traditional machine learning methods.

Fake Review Detection: Classification and Analysis of Real and Pseudo Reviews →This paper describes features used in the model including sentiment analysis, part - of - speech tagging. And user behaviour analysis. The author concludes with the approach that it can automatically detect and remove fake reviews, improving the overall quality.

Detection of fake reviews using NLP &Sentiment Analysis | IEEE Conference Publication→This paper uses CNN and DT. This system extracts features from the text of reviews using CNN and DT to classify whether they are genuine or fake.

Detecting Fake Reviews Utilizing Semantic and Emotion Model | IEEE Conference Publication → This paper proposes a heuristic-based approach for detecting opinion spam. The approach uses the sentiment of the review and the frequency of specific words to identify spam.

A Study on Identification of Important Features for Efficient Detection of Fake Reviews | IEEE Conference Publication → This paper discusses the impact of the fake review on business, and the different techniques. Used to detect them, including ML, NLP, DL and data mining.

A Deep Learning Approach for Fake Review Detection →This paper proposes a deep learning framework that combines CNN and ANN for fake review detection and achieves high accuracy on a large dataset.

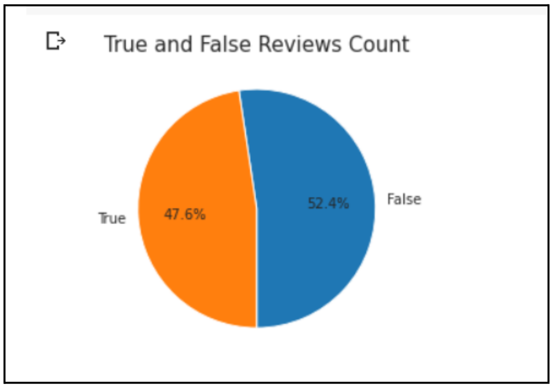
Detecting Fake Reviews Using Convolutional Neural Networks" → by H. F. El-Sofany et al. This paper presents a CNN-based approach for fake review detection and compares its performance with traditional machine learning methods on a dataset of Amazon product reviews.

Dataset Description

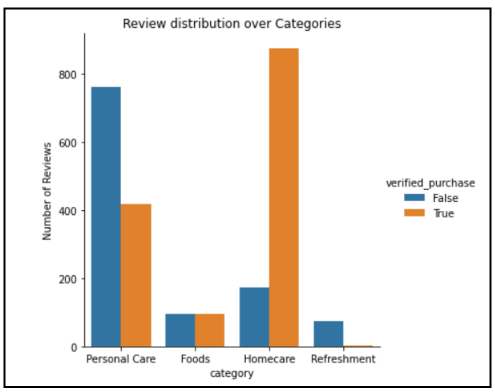
```
Data columns (total 32 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   report_date              2501 non-null   object
 1   online_store             2501 non-null   object
 2   upc                      2501 non-null   float64
 3   retailer_product_code    2501 non-null   object
 4   brand                    2501 non-null   object
 5   category                 2501 non-null   object
 6   sub_category             2501 non-null   object
 7   product_description      2501 non-null   object
 8   review_date              2501 non-null   object
 9   review_rating            2501 non-null   int64
 10  review_title             2403 non-null   object
 11  review_text              2501 non-null   object
 12  is_competitor            2501 non-null   int64
 13  manufacturer             2501 non-null   object
 14  market                   2501 non-null   object
 15  matched_keywords         0 non-null      float64
 16  time_of_publication      0 non-null      float64
 17  url                      1654 non-null   object
 18  review_type              2501 non-null   object
 19  parent_review            2501 non-null   object
 20  manufacturers_response   0 non-null      float64
 21  dimension1               2501 non-null   object
 22  dimension2               2501 non-null   object
 23  dimension3               2310 non-null   object
 24  dimension4               0 non-null      float64
 25  dimension5               0 non-null      float64
 26  dimension6               0 non-null      float64
 27  dimension7               2499 non-null   object
 28  dimension8               2501 non-null   object
 29  verified_purchase        2501 non-null   bool
 30  helpful_review_count     2501 non-null   int64
 31  review_hash_id           2501 non-null   object
```

The dataset taken for our model is the Amazon Reviews dataset. It is focused on the products sold on Amazon website and the reviews that are added by the customers about the products. The dataset contains 2500 rows and 32 columns.

Exploratory Data Analysis



Percentage of True and False Reviews in the dataset



No of True and False Reviews per category

WordCloud showing common words in the reviews

Data Preprocessing

> Few columns contain many null values, so replacing these values with the mean values of the columns.
> Duplicate values were also removed.
> Feature Transformation: The data present is categorical, so the string values have been scaled to Integer for model prediction.
> Feature selection: On the basis of Information Gain, some of the attributes like dimensions, time-of-publication, category and few more columns which are product descriptive and which provide low information gain for review (dependent variable) are dropped.

After these steps, text processing was done like spelling correction: using TextBlob() removing punctuations: using regex, removing stopwords: from NLTK SW library , lemmatisation and tokenisation.

Baseline Models

After preprocessing the data, we have used several Machine Learning models to classify reviews based on the features in the sample. We have used the following classifiers:

1. Logistic Regression: It is a type of regression used in case of classification problems. It learns a linear relationship from the given dataset and then introduces a non-linearity in the form of the Sigmoid function.

2. Gaussian Naive Bayes: It is a type of classification model which uses Bayes

algorithm. It is easy and fast in multiclass classification as it needs less training data. It is used to determine the benchmark performance of the models.

3. Random Forests Classifier: It is ensemble learning of Decision Trees(which provides interpretability and is non-parametric in nature) where some weak classifiers are combined and the prediction is done by majority voting for classification problems.

4. Decision Tree Classifier: A decision tree is a non-parametric supervised learning algorithm which provides interpretability while doing classification. At each level, a feature is chosen as per its information gain or entropy for classifying data and final classification is obtained at the leaf level.

5. SVM : A support vector machine (SVM) is a supervised learning algorithm to classify or predict data groups. The goal of the SVM is to determine the unique decision boundary known as Optimum Separating Hyperplane (OSH) that can segregate n-dimensional space into the required number of regions for classification.

6. Additional Models used -
CNN - Convolutional Neural Networks are deep learning models used to reduce multidimensional data to scalar data so that it can be used in a neural network for classification.
ANN- Artificial Neural Networks are based on the working of a human brain neuron. Each node in the network is assigned a weight to it which is updated during backpropagation of the error between predicted and actual labels. Finally we get the set of optimal weights.

Results and Analysis

We have implemented two additional models CNN and ANN . We have used word Vectorization method of converting words or phrases from a vocabulary into a corresponding vector of real numbers, which can be used to analyze word predictions and semantics. This vectorizer matrix is passed into the models as training data . Some of the results of our models are shown below.

Accuracies of baseline models

	MNB	SVM	LR	DT	RN
Count Vectorizer	80	80	82	76	78
Tfidf Vectorizer	79	82	82	76	76

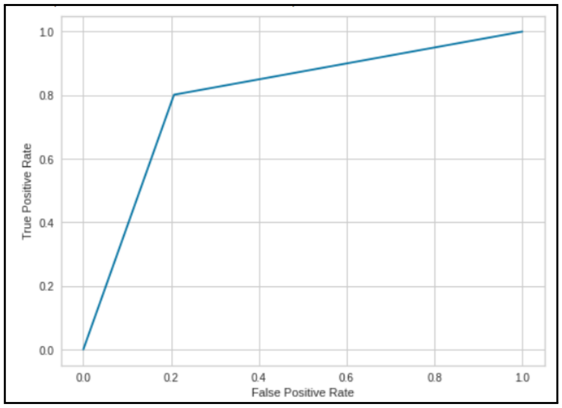
Accuracy of CNN

```
### Test Accuracy
model_cnn.evaluate(test_c.toarray(),y_test)

22/22 [==============================] - 2s 105ms/step - loss: 0.5829 - accuracy: 0.7460
```

Accuracy of ANN

```
### Test Accuracy
model_ann.evaluate(test_c.toarray(),y_test)

22/22 [==============================] - 0s 2ms/step - loss: 0.5320 - accuracy: 0.8423
[0.5319725871086121, 0.8423357605934143]
```

ROC- Curve of CNN



Precisions of various models

	MNB	SVM	LR	DT	RN
Count Vectorizer	79	74	75	69	71
Tfidf Vectorizer	81	77	81	69	69

Recalls of various models

	MNB	SVM	LR	DT	RN
Count Vectorizer	77	89	91	85	87
Tfidf Vectorizer	72	86	81	85	85

F1-scores of various models

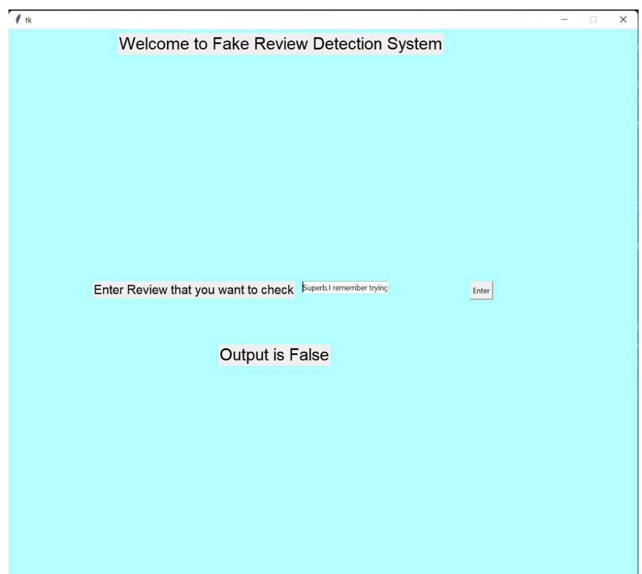
	MNB	SVM	LR	DT	RN
Count Vectorizer	78	81	83	85	87
Tfidf Vectorizer	76	81	81	85	85

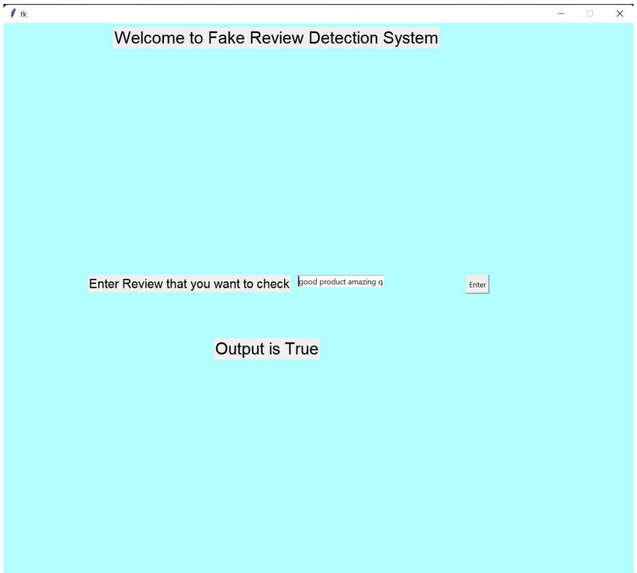
From the above results , the overall accuracy were close to 80% for all the models. Also the precision is higher for some models while recall
The CNN model was trained on a dataset using binary cross entropy loss. Adam optimizer. The model achieved 74.6% accuracy on the test data. The confusion matrix plotted shows the true positives , true negatives , false positives and false negatives. The area under ROC came out as 0.797,which is an indication of the model's ability to distinguish between positive and negative classes.
In ANN the model has achieved the accuracy of 84.23% on the test set, and the area under the CURVE (ROC) is 0.82. The ROC curve is a graphical representation of the performance of a binary classification model at different

thresholds. The Higher the AUC, the better the model's performance.

Front End Design

Our model takes user reviews and the classify the input into true or false review using ANN classifier. Some screenshots of our model are:





Conclusion

The purpose of the fake review detection is to filter out fake reviews. Random Forests outperformed other classifiers when it came to classifying our dataset in this study. It shows that absolute classifiers are better for this errand. Exploring the dataset was made easier by the data visualization, and the features that were found improved the classification's accuracy. The accuracy of the various algorithms used demonstrates how well they have performed in relation to their accuracy factors.
In addition, the method gives the user the ability to recommend the most honest reviews so that the customer can make decisions about the product.

Future Work

> To make use of real-time or time-based datasets, which will enable us to compare the timestamps of the user's reviews in order to determine whether a particular user is posting an excessive number of reviews in a short amount of time.

> To apply and compare NLP-based models like Lstms and Transformers in order to broaden the scope of the study to deep learning methods.

> To create a method that is comparable for unsupervised learning of unlabeled data in order to identify fake reviews