香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

ACCT 6243: Text Analytics in Financial Market

Final Research Report

Identify Spam Reviews

Student ID	Name
222021054	白含岭
222025036	邱杏颖
222021047	宋天澍
222021090	左昕叶
222021060	罗清月

1. Introduction and Problem Identification

1.1 Industry Background

In February 2023, Amazon took legal action against six defendants in an effort to protect its customers and selling partners from fake review brokers. Each lawsuit targets defendants who sell fake review services to bad actors attempting to operate Amazon selling accounts. These fraudsters commit fraud by providing false reviews, intentionally deceiving customers, and attempting to gain unfair competitive advantages over honest sellers on Amazon. Because of the ravages of fake reviews, Amazon had to launch a cleanup of fake reviews to change the user environment.

People largely rely on user-generated reviews to make buying decisions, especially when selecting electronic products such as cameras online. The number of reviews is growing, yet there are many fake reviews posted by businesses to complain on competitors' products or to beautify their own products. Such reviews are regarded as spam reviews which make bias and lower information quality. Thus, identifying spam reviews has become an interest for both academics and the industry. Literature suggests approaches such as logistic regression, SVM and Bayesian, etc. Before applying those models, preprocessing steps including tokenization are applied. While some literature provides satisfactory outcomes, changing data horizons and sources as well as machine learning steps may lead to different results. Thus, identifying spam reviews for a specific product remains a challenge.

1.2 Problem Statement

Problem:

Spam reviews can harm the business platform and provide unreliable references for potential customers. Generating spam reviews has lower costs while identifying them becomes costlier. It is challenging to balance the in-performance and out-performance of classification models seeking to identify spam reviews from various contents. Besides, identifying machine-generated reviews may be difficult.

Objective:

Improve the quality and purity of commodity reviews by identifying the spam reviews effectively through textual analysis and machine learning. Evaluate the model performance using both human-generated and machine generated fake reviews. Compare the identifying ability (performance) of classification model(s) and that of AI.

Solution:

Most methods for identifying spam reviews are based on text binary classification. The specific process includes: first obtain raw data through web crawling, construct a standard dataset through data preprocessing, select part of the data in the dataset as training samples, then find out the features for the review to be classified, and manually label them as spam review and non-spam review, finally use Logistic regression, Bayesian, SVM and other algorithms to classify spam review and compare the performance of the models.

This project focuses on product reviews about electronic products on Amazon, one of the most popular E-commerce platforms throughout the world. After obtaining reviews data through web scrapping, we preprocess the text by implementing tokenization, removing stop words, etc. Experimenters (group members) will manually label the preprocessed spam reviews and non-spam reviews.

We plan to extract the features for identifying spam reviews from the textual data, i.e., the reviews. Possible features include the sentiment of reviews, text length and fog index, whether images are included, etc.

After feature engineering, we will randomly select some reviews as the training sample to train different classification models, and use Precision, Recall and F-measure to evaluate the performance of spam review classification model.

Expected Impacts：

We expect to establish a classification model (models) that can effectively distinguish spam reviews and non-spam reviews. The performance of the model(s) will be evaluated by several quantitative metrics, including accuracy, precision, recall, F1-score, ROC, and AUC. We anticipate that the model(s) will achieve high values for these metrics, indicating a high degree of accuracy and reliability in identifying spam reviews. The model(s) should also have good generalization performance, providing insights into the underlying patterns and features relevant for identifying spam reviews and informing future research in the field of text classification.

In addition, the establishment of a reliable classification model can significantly affect various stakeholders. For the Amazon platform, the model can help improve the quality and credibility of product reviews, increasing customer trust, satisfaction and loyalty. For electronic product companies, the model can provide valuable insights into the effectiveness of their marketing strategies, as well as help them identify and address issues with their products. For consumers, the model can provide a trustworthy source of information for making purchasing decisions.

2. Data

2.1 Data sources

This project would focus on the product review from Amazon. Because Amazon has a large customer base which would provide sufficient samples. High reviewer diversity in Amazon could help model do better generalization. Amazon usually has high-quality reviews. Detail description would give opportunity to conduct more kind of research. In Amazon products, the proportion of electronic products is relatively high, so we will choose electronic product field. This project has two data sources.

Kaggle

This dataset contains nearly a million reviews of electronic product range from 2000 to 2014. It has labelled whether the comment is spam. It will be used as our training data.

Scraping

This dataset contains 5,000 reviews of product Airpods pro range from 2019 to 2023. We choose it because is a popular product with a lot of data and views. High competitiveness in this field gives merchants more motivation to manipulate comments. The dataset would include: User ID, Review date, Review rating, Content, Anonymized, Helpful votes, verified purchase.

2.2 Data Inspections and Exploration

2.2.1 Preprocessing and statistics

We preprocess both datasets by implementing lemmatization, tokenization and the removal of punctuation and stop words. At first, we utilize NLTK to preprocess data, but the word clouds show that NLTK cannot effectively remove stop words. Thus, we utilize SPACY to process data and make the word clouds below.

The first word cloud is drawn from the scrapped dataset and the second one from the Kaggle dataset.

The first word cloud shows that people buy airpods mainly for listening to music. Customers care about functional features, e.g., battery life, sound quality, noise cancellation, etc. People use sentiment words such as amazing, great and nice to describe their feelings.

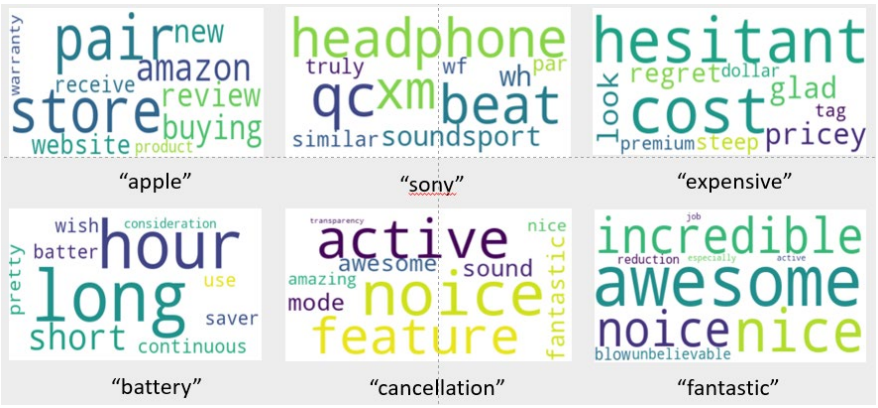
The second datasets show that 1) functional features are mentioned more frequently than sentiment words, 2) reviewers compare brands with each other, i.e., Samsung galaxy and iPhone, etc.

Descriptive statistics

We calculated net sentiment score, fog index and similarity score for the textual data of the airpods dataset. We also conducted word embedding and topic modelling (LDA) for the airpods dataset.

Word embedding

The word embedding results show that Sony Wf, Xm and Wh as well as Bose QC are competitors of Airpods Pro. Battery and noise cancellation mode of Airpods seemed to be appraised.



Topic Modelling (LDA)

The topic modelling results are as follows. These results generally show that noise cancellation, battery, sound quality are always the themes of user reviews. And these features of Airpods are generally praised. Besides, the earbuds may fall out of ear.

Topic 0: airpod pro **noise** like sound earbud feature mode apple control

Topic 1: ear airpod fit fall sound stay great love **noise** pro

Topic 2: text read message **love** airpod good los phone **fall** run

Topic 3: case earbud apple ear bud work charge hour **good** sound

Topic 4: **sound quality** great noise apple product **good buy love battery**

Topic 5: airpod pro apple ear sound tip come use fit work

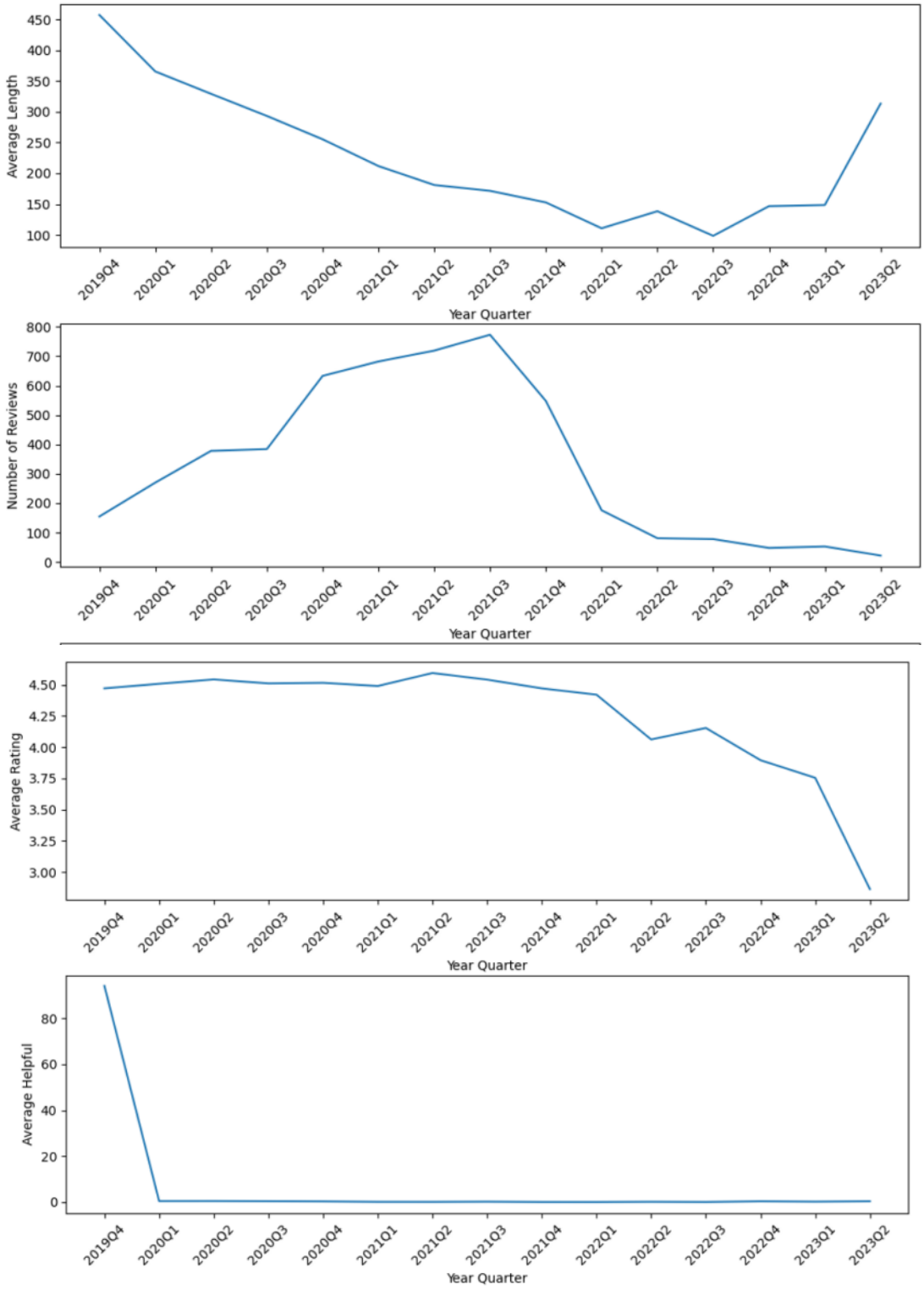
Topic 6: noise hear sound **cancellation music** cancel work **great good** listen

Topic 7: pod air work airpod **great connect** phone **sound ear buy**

Topic 8: sound noise good charge battery ear life **great** quality cancellation

Topic 9: apple airpod work **buy** issue pair product time charge pro

Time series



The above results show that 1) average length of comments decreases with time but has begun to increase since 2023, 2) the number of reviews is negatively correlated with average length of

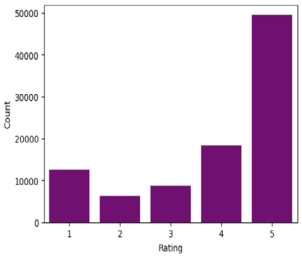
reviews, 3) For the data crawled, which are sorted from “to extent to which other users feel helpful”, the average rating given by reviewers has been decreasing since 2022. 4) The earlier the review date, the more possible that the review would be regarded as helpful.

The above results show that generally the anonymized reviews (which sum up to over 200) are dispersed among the 5000 reviews. The earlier the review date, the more possible the reviews are long and complex. But somehow the reviews have become longer and more complex again since 2023. The similarity score seems to walk randomly. Although the length and complexity of

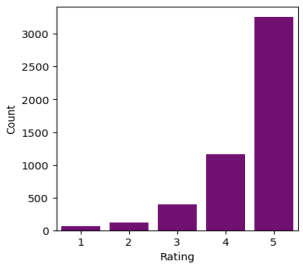
comments has a “high-low-high” pattern, the net sentiment score decreases slowly before 2023 but decreased sharply during 2023. The sharp decrease in net sentiment is in contrast with the increase in fog index and review length in 2023.

2.2.2 Distribution plot

We have created distribution plot for both databases. The following figure shows the distribution of user ratings. People tends to give high rating. The proportion of the two databases is similar. Rating 5 accounts for over 50%. Rating 4 accounts for approximately 20% of all comments.

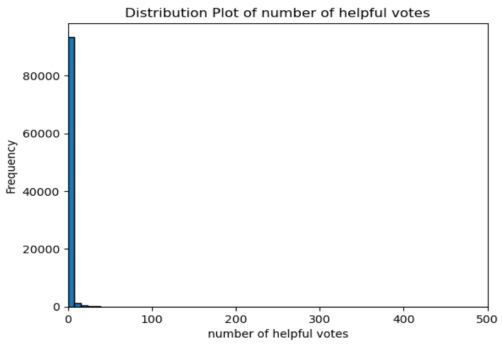


Kaggle

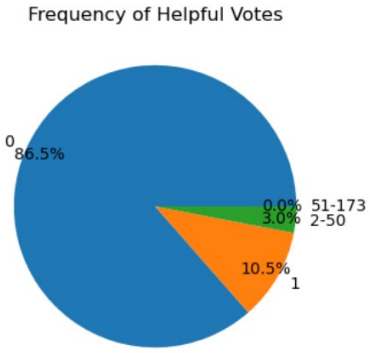


Scraping

For helpful votes, most reviews have not received any votes from other user. The helpful votes of both datasets are concentrated between 0 and 50.

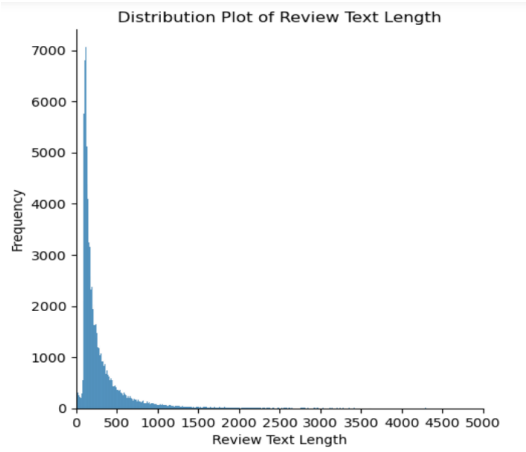


Kaggle

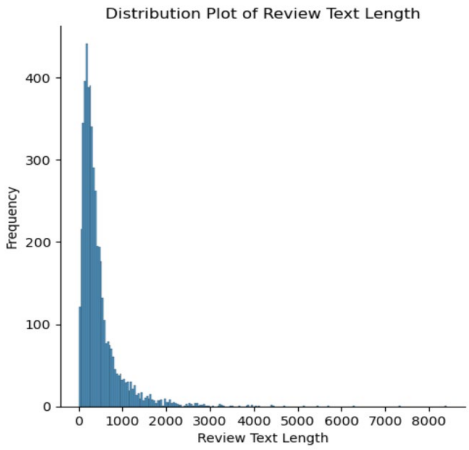


Scraping

The review length of the Kaggle dataset is around 0-1000 words. Airpods often have comments with a slightly more words. Buyer of airpods usually not only commenting on their feelings, but also comparing them with Apple's previous products and other brands.



Kaggle



Scraping

3. **Feature Engineering**

Class	Whether the review is fake or not
rating	Consumer rating of the product
fog_index	A numerical score assigned to an input text where larger values indicate greater difficulty of reading the text
len_num	Percentage of the length of number and character in the whole sentence
len_letter	Percentage of the length of letter in the whole sentence
helpful	The number of people who found the review useful
Per_pos	Percentage of the length of positive word in the whole sentence
Per_neg	Percentage of the length of negative word in the whole sentence
Net_sen	The difference between length of positive word and length of negative word

3.1 Introduction

In this module, we will use various machine learning models to explore which words in preprocessed comment text have the greatest weight in predicting spam comments.

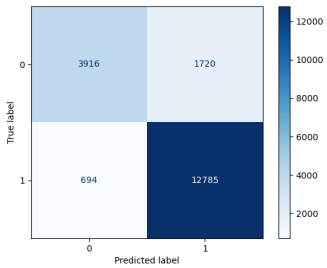
3.2 Research process

1. Store the corresponding comment text in CSR sparse matrix. 2. Divide the samples into test set and training set. 3. Use logistic regression, Bayesian classifier, and SVC classifier to model the training set, and find the best hyperparameters through RandomizedSearchCV. 4. Apply the fitted model to the test set, calculate the accuracy score through confusion matrix, and find the words with the highest weight. 5. Select the model with the highest accuracy score.

3.3 Results and illustration

3.3.1 logistic regression

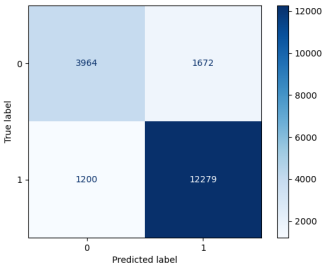
```
lr_model: C: 0.1 penalty: l2
True Positive(TP)  =  12785
False Positive(FP) =  1720
True Negative(TN)  =  3916
False Negative(FN) =  694
['love' 'perfectly' 'perfect' 'exactly' 'great']
['waste' 'disappointed' 'broke' 'stopped' 'useless']
Accuracy of the binary classification = 0.874
```



Results illustration:For logistic regression, the best hyperparameters we found were C: 0.1 and penalty: L2. With these hyperparameters, when we applied the model to the test set, the results were TP: 12785, TN: 3916, and accuracy: 0.874. The top five words with the highest weight and the bottom five words with the lowest weight are shown in the figure.

3.3.2 Bayesian classifier

```
nb_model: alpha: 0.1
True Positive(TP)  =  12279
False Positive(FP) =  1672
True Negative(TN)  =  3964
False Negative(FN) =  1200
['phone' 'case' 'great' 'one' 'like']
['blahh' 'unhappy' 'worthless' 'disappointment' 'trash']
Accuracy of the binary classification = 0.850
```



Results illustration: For Bayesian classifier, the best hyperparameters we found were Alpha: 0.1 . With the hyperparameter, when we applied the model to the test set, the results were TP: 12279, TN: 3964, and accuracy: 0.850. The top five words with the highest weight and the bottom five

words with the lowest weight are shown in the figure.

3.3.3 LinearSVC

Results illustration:For LinearSVC, the best hyperparameters we found were C: 0.01 and penalty: L2. With these hyperparameters, when we applied the model to the test set, the results were TP: 12826, TN: 3878, and accuracy: 0.874. The top five words with the highest weight and the bottom five words with the lowest weight are shown in the figure.

3.4 Summary for model generation

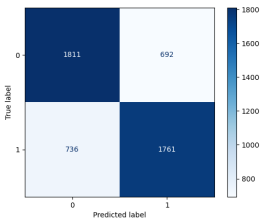
In summary, both the logistic regression model and LinearSVC model have high accuracy scores, so we have selected these two models as our final choices. Additionally, we found that the top weighted words obtained by both models were the same and were mostly positive. We speculate that this is because the use of positive words may have a two-fold origin; one may come from sincere customer evaluations, while the other may come from spam comments generated either by AI or humans.

3.5 Model applying

In this section, we will apply the two models generated above to the crawled data to study their generalization ability.

3.5.1 logistic regression

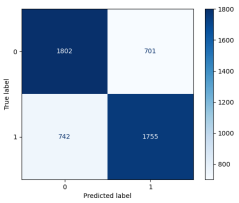
```
True Positive(TP)  =  1761
False Positive(FP) =  692
True Negative(TN)  =  1811
False Negative(FN) =  736
['matter' 'connecting' 'favorite' 'include' 'conference']
['number' 'along' 'these' 'amazed' 'near']
Accuracy of the binary classification = 0.714
```



Results illustration:For logistic regression, with the hyperparameters C: 0.1 and penalty: L2, when we applied the model to the test set, the results were TP: 1761, TN: 1811, and accuracy: 0.714. The top five words with the highest weight and the bottom five words with the lowest weight are shown in the figure.

3.5.2 LinearSVC

```
True Positive(TP)  =  1755
False Positive(FP) =  701
True Negative(TN)  =  1802
False Negative(FN) =  742
['matter' 'favorite' 'connecting' 'include' 'conference']
['number' 'along' 'these' 'amazed' 'near']
Accuracy of the binary classification = 0.711
```



Results illustration:For LinearSVC, with hyperparameters C:0.1 and penalty: L2, when we applied the model to the test set, the results were TP: 1755, TN:1802 , and accuracy: 0.711. The top five words with the highest weight and the bottom five words with the lowest weight are shown in the figure.

3.5.3 Summary for model applying

Based on the results of the models, the logistic regression model achieves the highest accuracy. Using this as a basis, we found that the highest weighted words in the crawled data were 'matter', 'favourite', 'connecting', 'include', and 'conference'. Therefore, we infer that the positive emotion words 'matter' and 'favourite' play an important role in judging spam comments, while the application-oriented words 'connecting' and 'conference' can assist us in better discrimination. Meanwhile, the verb 'include', which often serves as a prompt for specific

items listed, still plays an important role. We believe that the high-weighted words designed in the training set and test set are important for our research discrimination.

4. **Classification Model**

In this study, we compared two binary classification models: SGD classifier, Support Vector Machines(SVM), random forest to distinguish spam reviews using Kaggle data source, which has been labeled spam or non-spam.

SGD Classifier stands for Stochastic Gradient Descent Classifier. It is a type of linear classifier that uses stochastic gradient descent as the optimization algorithm to train the model. It is particularly useful for large-scale machine learning tasks, where the dataset is extensive or continuously updated.

The main idea behind SVM is to find the optimal hyperplane that separates the data points of different classes with the maximum margin. In a binary classification scenario, the hyperplane is a decision boundary that divides the feature space into two regions corresponding to the two classes. The margin is the distance between the decision boundary and the nearest data points of each class, called support vectors.

Random Forest is a ensemble learning method used for classification and regression tasks in machine learning. It belongs to the family of decision tree-based algorithms and is known for its effectiveness in handling high-dimensional datasets and capturing complex relationships between variables.

In validation of Model, we used a simple train-test split of the dataset, with 80% of the observations belonging to the training dataset. Optimal hyper-parameters were chosen using three-fold cross-validation on the training dataset. The search space is defined as a grid of all possible combinations of the hyper-parameters.

In model evaluation, we selected f1_score, precision_score, recall_score indicators to compare three binary classification models in this study. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative. The best value is 1 and the worst value is 0.

The recall is intuitively the ability of the classifier to find all the positive samples.The best value is 1 and the worst value is 0.

The AUC value ranges from 0.5 (random classification) to 1 (perfect classification). Higher AUC value, better performance.

Evaluation	Conclusion	Recall	Precision	AUC
SGD Classifier	√	0.99	0.99	1
SVM	√	1	1	1
Random Forest	√	1	1	1

To sum up, according to these metrics, we can see that three models show great classification power.

5. **Conclusion**

At present, we have obtained a model for judging fake reviews with high accuracy. We can roughly judge the reviews of Amazon's AirPods products, which can help Amazon and customers filter fake reviews to a certain extent and get real and effective product feedback. However, we only focus on electronic products, not on a wider range of commodities, so there are certain

limitations in judging fake reviews.

In the long run, there are currently many language models to correct and judge false comments. We have not used such data for model training, and we can introduce more languages for training in the future.

Another process we can improve results from the tides of ChatGPT and other AI models. Tapping in simple instructions can make AI models generate thousands of seemingly credible reviews in a very short time. Do these reviews sound more credible than human-generated contents? The suggested models like logistic regression may achieve a satisfactory accuracy for identifying human generated reviews. But do they identify machine-generated reviews effectively as well? Do ChatGPT identify spam reviews more effectively than human-generate algorithms? From the perspective of identifying spam reviews, we can further discuss the boundaries of humans, algorithms and AI—possibly shedding a light on how humans can develop their advantages over AI.

Reference

[1] Crawford, M., Khoshgoftaar, T. M., Prusa, J. D., Richter, A. N., & Al Najada, H. (2015). Survey of review spam detection using machine learning techniques. *Journal of Big Data*, *2*(1), 1-24.

[2] Jindal, N., & Liu, B. (2008, February). Opinion spam and analysis. In *Proceedings of the 2008 international conference on web search and data mining* (pp. 219-230).

[3] Rodrigues, A. P., Fernandes, R., Shetty, A., Lakshmanna, K., & Shafi, R. M. (2022). Real-time twitter spam detection and sentiment analysis using machine learning and deep learning techniques. *Computational Intelligence and Neuroscience*, *2022*.

[4] Salminen, J., Kandpal, C., Kamel, A. M., Jung, S. G., & Jansen, B. J. (2022). Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services*, *64*, 102771.

[5] Fayaz, M., Khan, A., Rahman, J. U., Alharbi, A., Uddin, M. I., & Alouffi, B. (2020). Ensemble machine learning model for classification of spam product reviews. *Complexity*, *2020*, 1-10.

[6] 孙升芸, & 田萱. (2011). 产品垃圾评论检测研究综述. *计算机科学*, *38*(B10), 198-201.

[7] 彭庆喜, & 钱铁云. (2013). 基于量化情感的网店垃圾评论检测. *山东大学学报 (理学版)*, *48*(11), 66-72.

[8] 李霄, & 丁晟春. (2013). 垃圾商品评论信息的识别研究. *现代图书情报技术*, (1), 63-68.