Fraudulent Product Detection Using Customer Reviews and Ratings on Amazon Product Data

Submitted by

PALANIVELRAJAN P – 2019202039

MCA REGULAR

under the guidance of

Ms' T. SINDHU

TABLE OF CONTENTS

ABSTRACT

- In recent years, Customers buy a lot of products on online which leads to increased rate in online fraud activities. Many fraud and defective products are sold online even on very known eCommerce platforms[7].
- Several machine learning algorithms are developed over the years with an increasing trend in anomaly detection techniques. These techniques can be utilized to intimate the upcoming buyer that there is a possibility of buying a low-quality or fraudulent product from the seller.
- Clustering and classification of product information empower the end customer to identify counterfeits accurately and efficiently by comparing them with trained models.

INTRODUCTION

- Shopclues is a ecommerce platform valued over $1.1B dollars in 2016 but due to increased sales of fake products and lack of tech commitment, the company failed and just sold for $70M[6].
- Several machine learning algorithms are developed over the years with an increasing trend in anomaly detection techniques. These techniques can be utilized to intimate the upcoming buyer that there is a possibility of buying a low-quality or fraudulent product from the seller.

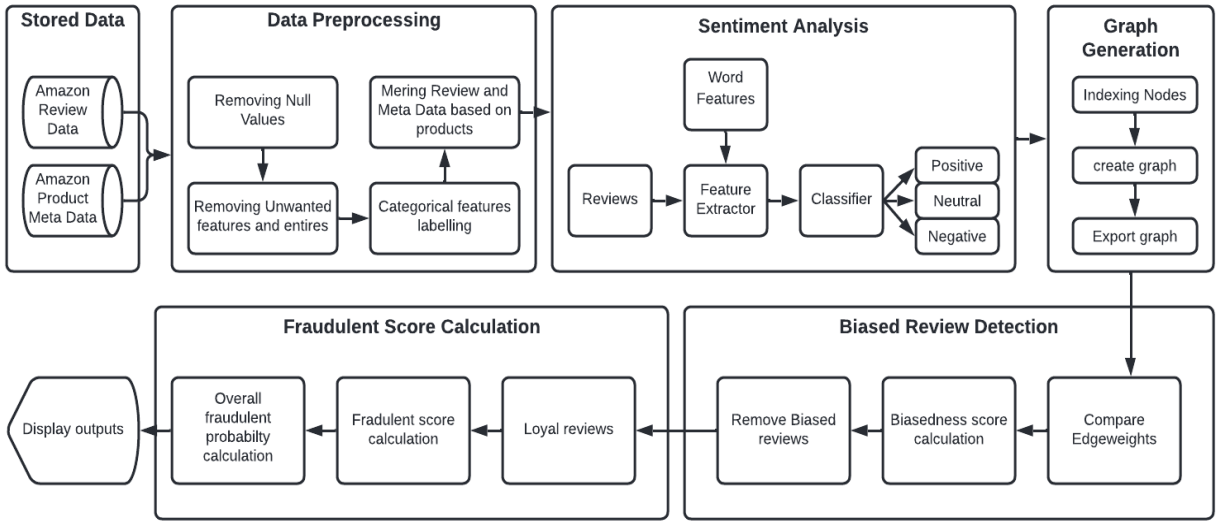
OBJECTIVE

1. To recognize the sentiments from the review texts.
2. To identify outlier nodes/ biased reviews.
3. To calculate the fraudulent score of the product from its loyal product reviews.

PROBLEM STATEMENT

1. One of the main problems is selling low-quality products, Customers can't able to find good products in the market.
2. Customers write a lot of reviews these days about the product qualities. One important factor is to identify the helpful reviews and with Only those reviews and ratings from previous users of the product can give a hint about the quality of the product.

ARCHITECTURE DIAGRAM



LIST OF MODULES

1. Data preprocessing
2. Sentiment Analysis on reviews
3. Graph Generation
4. Biased Reviews Detection
5. Fraudulent Score Calculation

BRIEF DESCRIPTION OF MODULE

1) DATA PREPROCESSING:

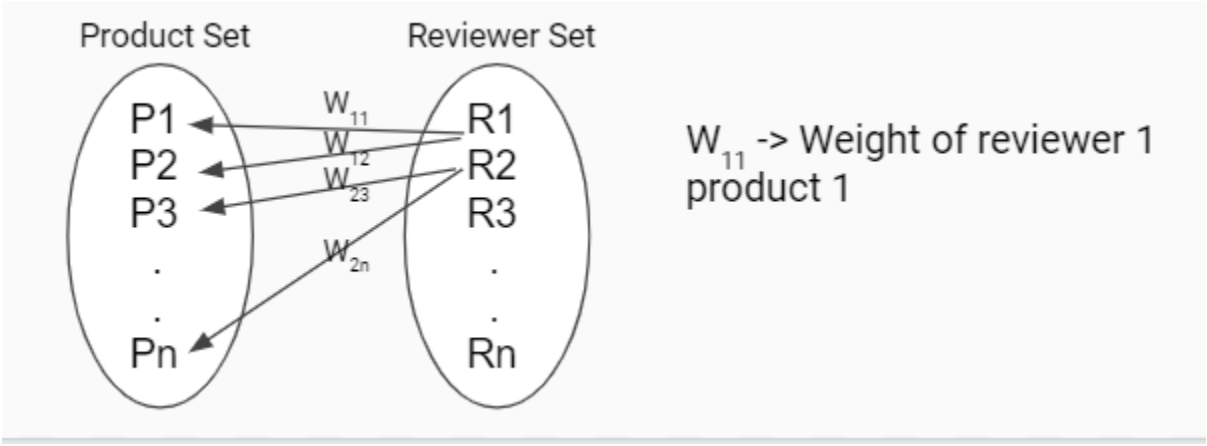
- This Module is a Data Cleaning where Null values are removed, Unwanted features are removed and categorical features will be labeled.
- There is a feature named helpfulness score for the review given in the dataset. Unwanted reviews are those reviews which are not helpful are removed.
- Finally the Review Data with the product ID will be mapped with the product ID in meta data and all the features are combined into a single dataset.

2) SENTIMENT ANALYSIS ON REVIEWS:

- In the preprocessed data we have a feature column ‘review text’ which contains the customer reviews for the products.
- To analyse the sentiment of the reviews a customer sentiment analysis function is implemented. These sentiments will be either ‘positive’ or ‘negative’ or ‘neutral’ class[5].
- Sentiment score for positive is 2, neutral is 1 and for negative is 0.

3)GRAPH GENERATION:

- The need for graph generation arised due to the scale of the data and its domain[1].
- The relation between product and customer is a bipartite relationship, no two products are related and no two users are related. Graph representation are the fastest way to interpret these type of data[2].
- The preprocessed and sentiment analyzed data points are then converted into a weighted directed bipartite graph.
- First set is the Product ID and second set is the Reviewer ID. If a product is reviewed by a user then there is an edge between these two nodes and their corresponding sentiment score, ratings {W = <S,R>} will be the Edgeweight



4)BIASED REVIEW DETECTION:

- Reviewers with biased reviews can create a greater impact while buying. These biased reviewers should be identified and reviews given by them are removed.
- A Biasedness score (β) is calculated for each reviewer per product

$$\beta_{Ui} = \frac{\sum r_{Ui}}{\left\{\frac{\sum r_i - r_{Ui}}{(N_i - 1)}\right\}}$$

Where,

β_{Ui} Biasedness score of user(U) for product(i)
r_{Ui} Ratings by user(U) for product(i)
r_i Ratings for product(i)
N_i Number of ratings for product(i)

- A Biasedness score (ɣ) is calculated for each reviewer per seller

$$\gamma_{Us} = \frac{\sum r_{Us}/N_{Us}}{\left\{\frac{\sum r_s - r_{Us}}{N_s - N_{Us}}\right\}}$$

Where,

γ_{Us} Biasedness score of user(U) for seller(s)
r_{Us} Ratings by user(U) for seller(s)
r_s Ratings for seller(s)
N_s Number of ratings for seller(s)
N_{Us} Number of ratings by user(U) for seller(s)

There are 4 main categories for biased reviews we are concentrating for now,

1. Type 1 - User will always give a positive review, ratings for all the products bought irrespective of its quality.
2. Type 2 - User will always give a negative, ratings review for all the products bought irrespective of its quality.

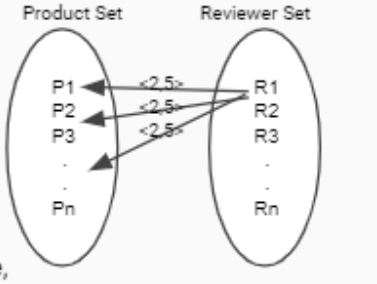
$$\beta_U = \frac{\sum_i \beta_{Ui}}{i}$$

Where,

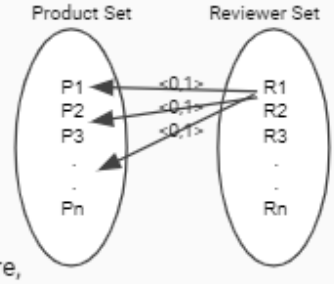
β_U Combined Biasedness score of user(U)

β_{Ui} Biasedness score of user(U) for product(i)

i Number of product(i)



Where,
<2,5> = <positive review, max rating> ∀ products reviewed by User 1



Where,
<0,1> = <negative review, min rating> ∀ products reviewed by User 1

A value of 0.25 is set as threshold limit, If the β_U of user is lesser than the threshold then the user is potentially a biased reviewer.

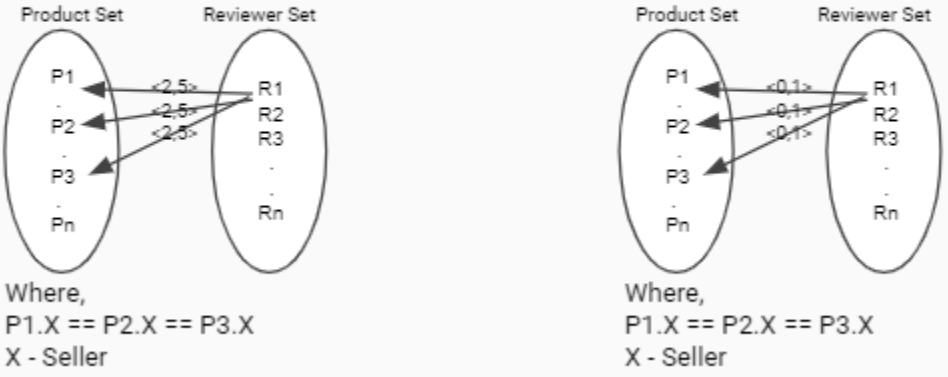
3. Type 3 - User will give positive reviews, ratings for all the products bought from a specific seller.
4. Type 4 - User will give negative reviews, ratings for all the products bought from a specific seller.

$$\beta_U = \gamma_{Us}$$

Where,

β_U Combined Biasedness score of user(U)

γ_{Us} Biasedness score of user(U) for seller(s)



Potentially Biased Reviewer - R1

A value of 0.25 is set as threshold limit, If the β_U value is lesser than the threshold then the user is potentially a biased reviewer.

5)FRADULENT SCORE CALCULATION:

1. Unbiased and Helpful reviews are given as input to this module. These reviews are considered as loyal reviews.
2. Fraudulent score (F) will be calculated for each product based on the helpfulness score, sentiment score and ratings.

$$F_{Ui} = Avg(h_{Ui} + s_{Ui} + r_{Ui})$$

Where,

F_{Ui} Fraudulent score for product(i) by user(U)

h_{Ui} Helpfulness score for product(i) by user(U)

s_{Ui} Sentiment score for product(i) by user(U)

r_{Ui} Ratings for product(i) by user(U)

3. Based on the fraudulent score, the probability of a product being a ‘good’ or ‘bad’ can be calculated and known by the future buyers or the platform owner.
4. Overall fraudulent score for a product is the average of all the fraudulent score for the given product by all the users.

REFERENCES

Base Paper: Li, A., Qin, Z., Liu, R., Yang, Y. and Li, D., 2019, November. Spam review detection with graph convolutional networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (pp. 2703-2711).

[1] T. U. Haque, N. N. Saber and F. M. Shah, "Sentiment analysis on large scale Amazon product reviews," 2018 IEEE International Conference on Innovative Research and Development (ICIRD), 2018, pp. 1-6, doi: 10.1109/ICIRD.2018.8376299.

[2] Akoglu, L., McGlohon, M., Faloutsos, C. (2010). oddball: Spotting Anomalies in Weighted Graphs. In: Zaki, M.J., Yu, J.X., Ravindran, B., Pudi, V. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2010. Lecture Notes in Computer Science(), vol 6119. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-13672-6_40

[3] Zhangyu Cheng, Chengming Zou, and Jianwei Dong. 2019. Outlier detection using isolation forest and local outlier factor. In Proceedings of the Conference on Research in Adaptive and Convergent Systems (RACS '19). Association for Computing Machinery, New York, NY, USA, 161–168. DOI:https://doi.org/10.1145/3338840.3355641

[4] https://jmcauley.ucsd.edu/data/amazon/

[5]https://towardsdatascience.com/step-by-step-twitter-sentiment-analysis-in-python-d6f650ade58d

[6]https://www.businessinsider.in/business/startups/news/shopclues-sold-heres-a-timeline-of-how-the-unicorn-startup-went-down/articleshow/71847099.cms

[7] https://www.statista.com/statistics/792047/india-e-commerce-market-size/