Analysing Interaction of Fake Review on E-Commerce Platform Using Text-Based Network Analytics

1. INTRODUCTION

Online Fake information in the form of reviews or news has gained the attention of the research and the industry in recent years. Detecting such information becomes even more crucial when we are incredibly dependent on such information due to a lack of resources, for instance, online product reviews. However, companies and researchers have shown much growth in developing better models to help predict whether reviews are fake in an eCommerce setting.

Considering the difficulties for the developer, we have proposed a method of studying the properties of the reviews. We created a text-based network and analysed the properties of the graphs. Understanding the properties of these nodes and doing a differential analysis of the fake reviews from the original reviews help us understand the unique properties while giving us an additional value add for using these internal characteristics to generate better classifier models.

According to Gu, Park, & Konana (2012), Word of mouth for online businesses has a significant impact on the sales and purchase intention of the consumers. Consumers need information before purchasing any form of product, especially for non-products where the evaluation of the product isn’t viable online like the experiential goods category. Online reviews hold much additional value for consumers.

With such high value of online reviews for organisations and the difficulty of establishing a positive internet reputation, various strategies, including illegal ones, have been utilised to increase online visibility. Fake informational content degrades information found on sites like Amazon and TripAdvisor. However, according to Jindal and Liu (2007), not all bogus reviews are equally detrimental. Phoney negative reviews on high-quality items are incredibly damaging to businesses. At the same time, fake positive ratings on low-quality products are equally damaging to customers. Fake positive ratings on low-quality items are equally detrimental to rivals that sell average or good-quality products with fewer reviews.

This research aims to examine and analyse the features of false reviews. We would demonstrate the use of this analysis in the construction of crucial prediction systems that may be developed utilising the features of such an analysis and properties obtained by online fake (and genuine) review networks.

2. DATA AND METHODS

The fake review dataset has been obtained from Open Science Foundation, containing approximately 20,000 fake reviews and 20000 real product reviews. Every review in the dataset has four attributes associated with it, which are:

1. Category: It represents the type of product for which the review has been given. The categories are

    a. Home_and_Kitchen_5

b. Sports_and_Outdoors_5

c. Electronics_5

d. Movies_and_TV_5

e. Tools_and_Home_Improvement_5

f. Pet_Supplies_5

g. Kindle_Store_5

h. Books_5

i. Toys_and_Games_5

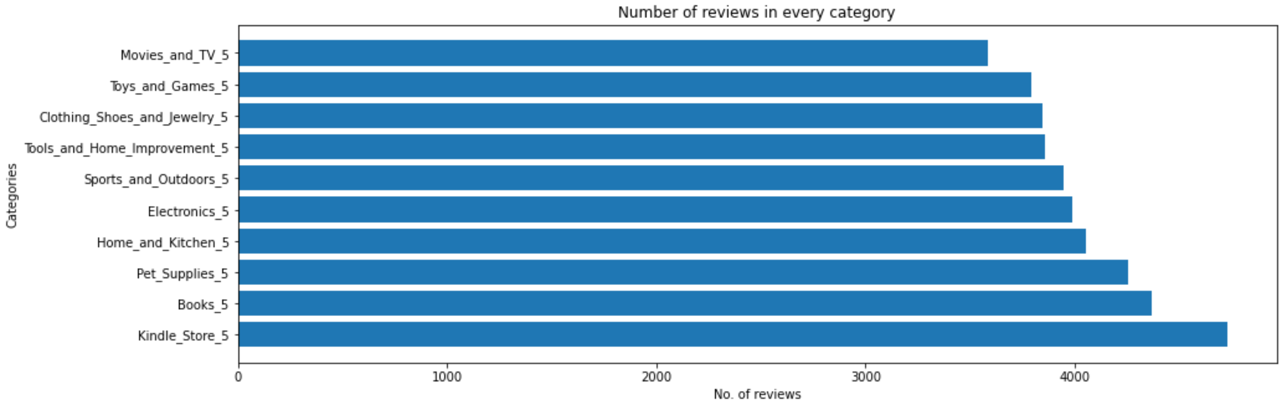
j. Clothing_Shoes_and_Jewelry_5

2. Rating: The rating given by the user in the review is between the range of 1-5. The higher the rating value, the more satisfied the customer is with the product.

3. label: Specifies whether the review is original and generated by a user (OG), or it has been generated by a computer (CG)

4. Text_: Contains the content of the review

Based on category, the dataset has been divided in this manner:



The dataset filters the category to be used, where network analysis is performed on the Kindle_store_5 category. This category contains the highest number of reviews in comparison to others. Later, Natural Language Preprocessing (NLP) is performed on every review of the filtered dataset to remove stop words, punctuations, numerical values, and expanded contractions. It also converts text to standard form

using lemmatise. As a result, we get the bag of words for each review in the dataset—a list of words in each list.

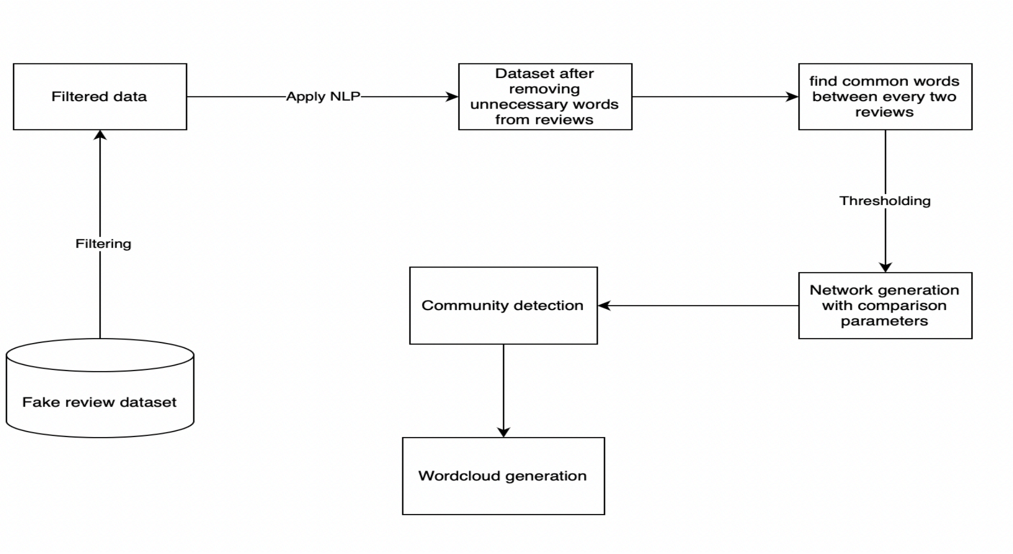

Filtered data
Apply NLP
Dataset after removing unnecessary words from reviews
find common words between every two reviews
Thresholding
Filtering
Community detection
Network generation with comparison parameters
Fake review dataset
Wordcloud generation

Image : Flowchart of methodology

We use the bag list of words to determine the number of common words present for every two reviews. The commonness helps in creating the network graph along with some network parameters. Thresholding also limits the number of connections between two reviews for better visualisations.

To draw out clusters that every network has, we use community detection. Community detection techniques are helpful to discover reviews focused on shared interests and keep them tightly connected in the graph theory. We have used greedy modularity and Girvan Newman techniques.

These clusters are then visualised separately, and their WordClouds are generated. The WordClouds assists in finding out the words that have occurred in the network the highest number of times.

3. RESULTS AND DISCUSSION

As we can observe visually from the two images below, the density of the communities of 4 and 5 star fake reviews is very high, in comparison to the communities drawn from communities of 4 and 5 star real reviews where each node (review) has very less connections with other nodes in their respective communities.

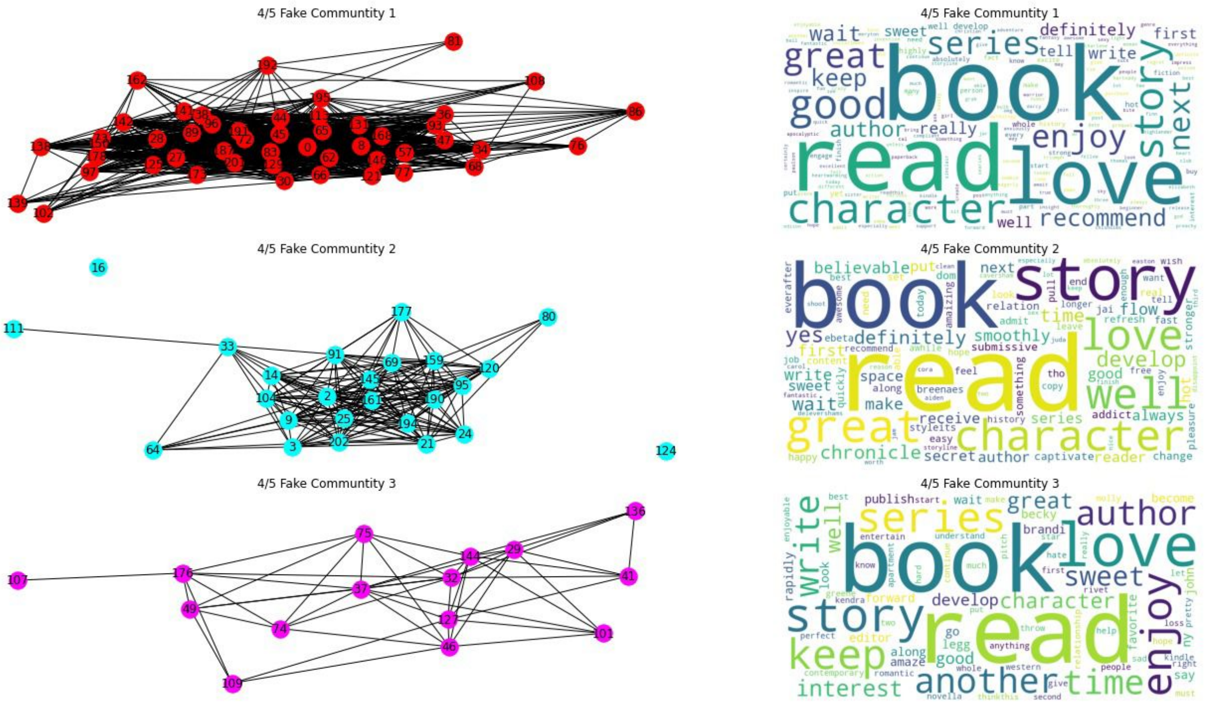


Image : Top 3 Communities of 4 and 5 star fake reviews

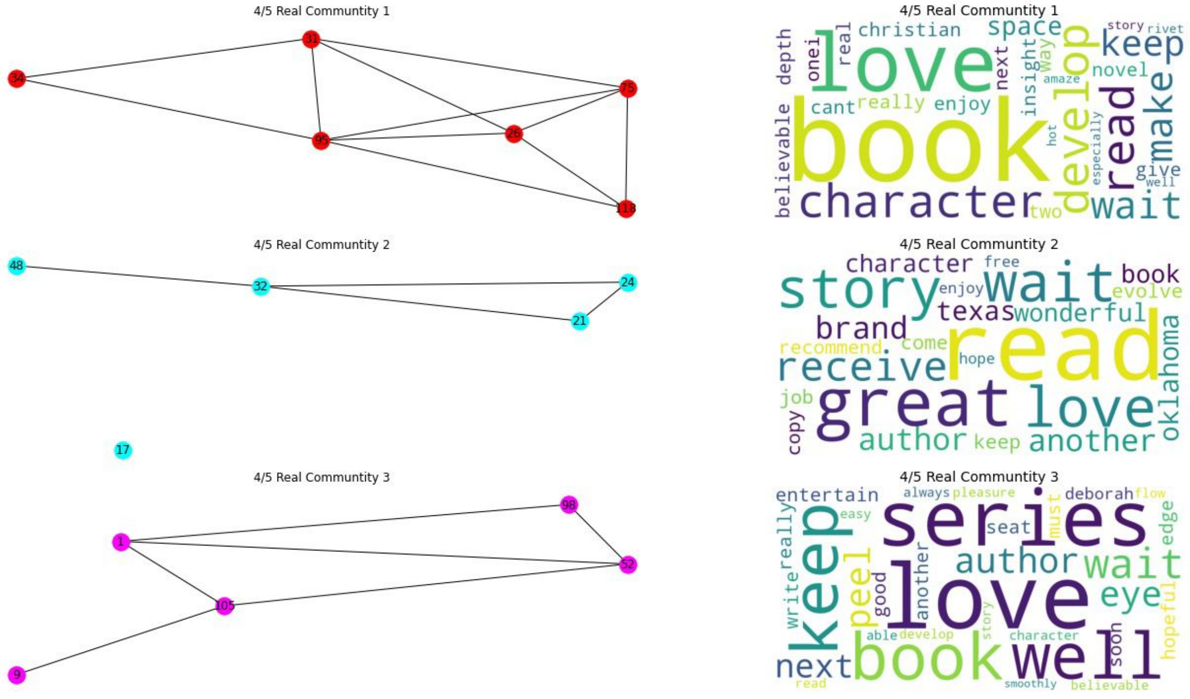


Image : Top 3 Communities of 4 and 5 star real reviews

The below network graphs and corresponding community barplots of 4 and 5 star real reviews represent that the real reviews consists of very small communities in terms of size, where the largest community has only 6 nodes. In comparison to the fake reviews graph and barlpot, the communities are larger in size relatively, indicating how well connected fake reviews are with each other.

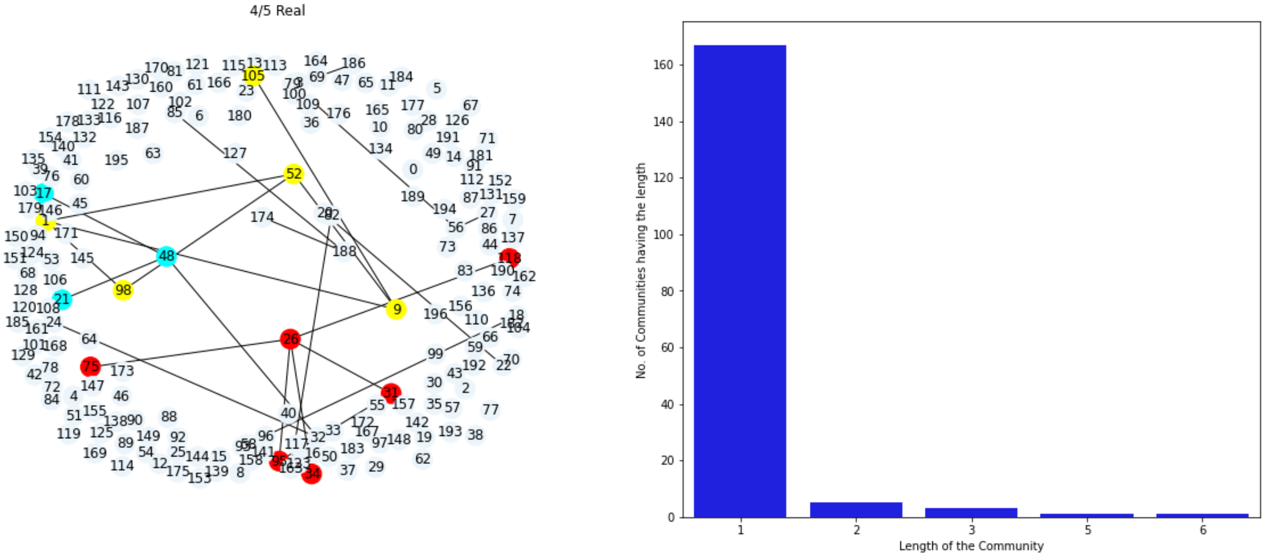


Image : Network graph for 4,5 star real reviews and community length barplot

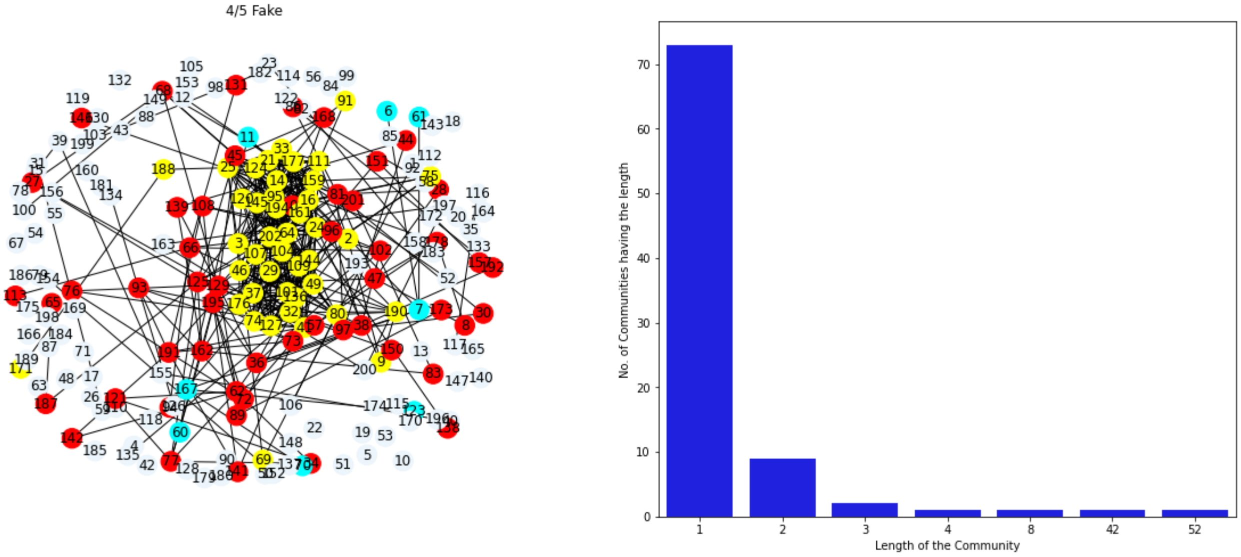


Image : Network graph for 4,5 star real reviews and community length barplot

From the network graphs and corresponding degree distributions, we observe that fake reviews are highly connected in comparsion to those of real reviews. 90% of nodes of real reviews have only one degree, whereas the degrees in fake reviews are more varied. Moreover , 4 and 5 star fake reviews are highly connected , with degree distribution spread to 30 nodes.

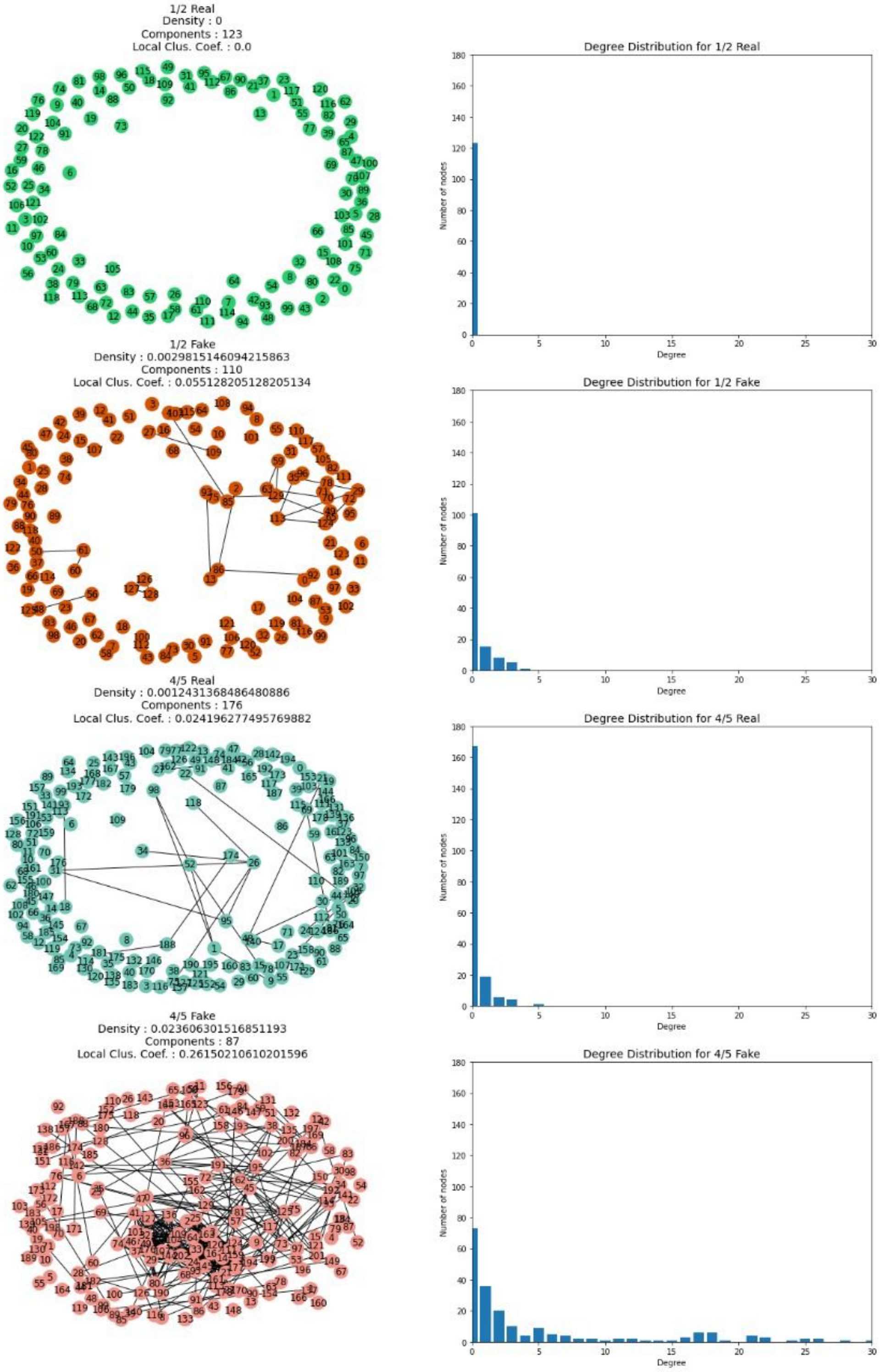


Image: Network graphs with corresponding degree distribution.

Networks graphs	**Density**	**No of components**	**Local clustering coeff**
1-2 star real	**0**	**123**	**0.0**
1-2 star fake	**0.0029**	**110**	**0.0551**
4-5 star real	**0.0012**	**176**	**0.0241**
4-5 star fake	**0.0236**	**87**	**0.2615**

4. REFERENCES

B. Gu, J. P., & Konana, P. (2012). Research note-the impact of external word-of-mouth sources on retailer sales of high-involvement products. *Information Systems Research*.

Cui, G., Lui, H.-K., & Guo, X. (2012). The effect of online consumer reviews on new product sales. *International Journal of Electronic Commerce*.

Jindal, N., & Liu, B. (2007, May). Review spam detection. In *Proceedings of the 16th international conference on World Wide Web* (pp. 1189-1190).