# Amazon Product Recommendation

Venkata Naga Sai Rama Krishna Pinnimty*
Department of Computer Science
Virginia Tech
*Blacksburg, VA, USA*
ramapinnimty@vt.edu

Sai Krishna Pavan Surya Tej Meda*
Department of Computer Science
Virginia Tech
*Blacksburg, VA, USA*
saikrishnapavanm@vt.edu

## ABSTRACT

Our project includes the implementation of a product recommender system based on product reviews and metadata history (product titles) from the Amazon Product co-purchasing Network metadata dataset. Given a query ID of a product, our system's main purpose is to recommend top-5 similar products efficiently and effectively. More specifically, our methods included a collaborative filtering model with user- and item-based strategies, an Ego-Network model, and a comparison against popular ML-based algorithms. Apart from this, we also performed basic Network Analysis to understand the underlying network structure, analyzed the sentiment of reviews with the help of the VADER model to better learn about the public pulse all while achieving reasonable results. Lastly, we also performed Topic Modeling to understand the product liking and preference patterns of users and thus, to gain a better overall "big-picture" view. In essence, our work is unique in that it includes a thorough qualitative and quantitative analysis across different aspects of social media networks. Finally, we also talked about the benefits and drawbacks of our approach and proposed some future directions.

## KEYWORDS

*Network Analysis, Sentiment Analysis, Topic Modeling, Recommender Systems, Amazon.com, Social Media Analytics.*

## 1 INTRODUCTION

COVID-19 has resulted in an increase in e-commerce and rapid digital transformation in the face of weakening economic activity [1]. As lockdowns became the "new normal," firms and customers gradually increased their online presence by offering and purchasing more goods and services online. Especially now more than ever, people are turning towards various e-commerce platforms like Amazon, eBay, etc., to purchase products that are of their liking. As a result, companies are constantly trying to come up with new and innovative ways to attract customers so that they can establish dominance over their competitors. One such way is for companies to build effective recommendation systems which can identify the underlying patterns based on the shopping behavior of customers and provide personalized recommendations.

Furthermore, millions of individuals use social media sites such as Facebook, Twitter, blogs, etc. to express their emotions, share opinions, and share perspectives about their daily lives. We receive an interactive media through online communities, where we as consumers inform and influence others through forums. However, the amount of content provided by users is just too large for a single person to assess, necessitating automation. Before purchasing a product, buyers can utilize sentiment analysis to determine whether the information provided about it is satisfactory. This analytical data is in turn used by marketers and businesses to learn more about the public pulse.

Therefore many e-commerce and social media companies are now actively trying to analyze the underlying network structure to draw useful insights and better understand their customers. Additionally, firms are trying to create robust Topic models that allow us to swiftly, cheaply, and without human intervention answer big-picture issues. They provide a framework for humans to interpret document collections directly by "reading" models or indirectly by employing topics as input variables for additional analysis once they have been trained.

This work is our joint effort aimed towards understanding Social Networks. This problem is best approached as a Data Analytics task as manual processing requires a significant amount of time and effort. Furthermore, any company that tries to build its brand and reputation through customer feedback can find our work useful to capture the sentiment of its users. For instance, Amazon can use it to separate the positive and the negative reviews for any listed product. Some of the key analyses we perform include (but not limited to):

1) Understand the underlying structure of the Amazon products network through *Network Analysis*.
2) Identify keyphrases with the help of *Topic Modeling* to get a "big-picture" view of the data.
3) Systematic *Sentiment Analysis* to identify the public pulse.

*Denotes Equal Contribution

Google Drive link to our project

4) Comparing and contrasting popular approaches that use Product Title and Reviews to *generate relevant recommendations*.

## 2 RELATED WORK

### 2.1 Network Analysis

Social media networks consist of crucial information which we can use to model effective solutions to make lives better [2]. Multiple aspects, most notably scale and time, can be examined in network graphs. Some research questions are concerned with the structure of the entire graph or large sub-graphs, while others are concerned with identifying specific nodes of interest. Some analysts will want to examine the complete graph across its entire existence, while others will want to segment the network into time units to investigate the network's evolution. Attempts to list the network analysis tasks that the majority of analysts will wish to undertake on their data collection have sparked debate [3]. A good starting point can be to analyze the graph and examine its characteristics as specified in [4]. For our project we analyzed the characteristics of the Amazon Product co-purchasing network and specifically incorporated Ego-Networks, taking inspiration from the work presented in [5].

### 2.2 Topic Modeling

Previous works explored using text mining to analyze subject trends and model topics in journal papers [7, 13, 14]. Another popular work discussed Topic modeling and clustering of research publications using a keyword-set based approach [16]. We made the choice to use LDA [15] over other topic modeling approaches because of the generative nature of the model. Thus, it can apply the model it employs to split documents into topics to even documents outside the corpora.

### 2.3 Sentiment Analysis

The authors in [10] considered doing sentiment analysis using a dictionary-based method and compared the results with those obtained using machine learning algorithms in their literature review. There are also ongoing efforts to use Deep Learning models to accomplish this task [11, 12]. For our task however, we chose to employ a simpler model to ease the process of predicting the sentiments of product reviews. Hence, we used the VADER [6] model since it gives predictions almost instantaneously with a reasonable accuracy.

### 2.4 Recommender Systems

In today's recommender systems, a variety of approaches and techniques are being developed and applied, the most prominent of which are collaborative filtering and content-based filtering approaches. Some previous attempts at integrating both these approaches include content-boosted collaborative filtering [8], weighted, mixed, switching and feature combination of different types of recommender systems [9]. In our work, we decided to explore an item-based Collaborative recommender system to generate recommendations based on Product Reviews. Additionally, we used Ego-Networks to generate recommendations when the only information we have is the title of the products.

## 3 APPROACH

We take a step-by-step approach where we first curate the data (section 3.1), pre-process the obtained data (section 3.2), perform basic Network Analysis (section 3.3) along with Topic Modeling (section 3.4) followed by Sentiment Analysis (section 3.5), and lastly compare our results using some popular recommendation strategies (section 3.6).

### 3.1 Data Collection

The dataset we use is Amazon product co-purchasing network meta-data which has been made publicly available on the Stanford Network Analysis Project (SNAP) community [2] since 2006. It was collected by crawling the Amazon website and contains useful information about product metadata and reviews about products from diverse categories like Books, Music, CDs, DVDs and VHS Video Tapes.

| Type | Count |
|---|---|
| Products | *548,552* |
| Product-Project Edges | *1,788,725* |
| Reviews | *7,781,990* |
| Product Category memberships | *2,509,699* |

Table 1: *Dataset Statistics*

Each product has information about Title, Salesrank, List of similar products, Category, and Reviews. Through our preliminary analysis we found that amongst all the available products, Books, Music CDs and Videos are the predominant categories.

### 3.2 Data Pre-processing

We first parsed the raw data and stored a cleaner version of it in a CSV file. Next, we converted all the characters in the reviews to lowercase, tokenized them and performed stop-word removal using the words specified within the

English language in the NLTK package. Although not fail-proof, we also attempted to remove words that are specific to Amazon. We also discarded punctuation, as keeping them will not add much value to the overall performance. Lastly, we used lemmatization to convert the tokens into their base form and generated bi-grams that can be used as auxiliary features.

### 3.3 Network Analysis

We checked to see if the data followed a “power-law distribution”. We mainly used the methods available in the `networkx` package in Python to calculate Degree centrality, Clustering coefficient and Avg. rating of every product to form a weighted connection amongst the products and lastly performed *ego-network analysis* to retrieve the top-k recommendations.

We found that there are `459` products with a sales rank of `-1`. We dropped them since they don’t contribute much to improve the performance of the models. Also we considered only the products with a sales rank value less than or equal to `100K` (i.e, only the top-`100K`) products for computational reasons. After this step, we are left with roughly `52K` products. The minimum and maximum ratings for books were found to be `0` and `5` respectively while the majority of the books had a rating of `4.5`.

There are almost `123K` edges to work upon. The minimum in-degree was `1` and the maximum in-degree was `82`. Similarly, the minimum out-degree was `1` and the maximum out-degree was `5`. Hence the total degree is between `2` and `87` both inclusive. The book *"The Great Gatsby"* has the highest degree of `87` and we cross-checked it on Amazon.com. The final graph we obtained is *not connected* since we are only considering a subset of nodes to expedite the overall analysis. This sub-graph has roughly `45K` nodes and `92K` edges. The sub-graph had a density of `8.7` and `2766` connected components.

### 3.4 Topic Modeling

One of the prominent machine learning techniques under unsupervised learning is Topic Modeling. Topic Modeling helps in clustering word groups by recognizing prominent patterns and expressions amongst the given set of documents. Given a collection of text documents, topic modeling performs word counting and groups equivalent word patterns from the latent information present in the unstructured data.

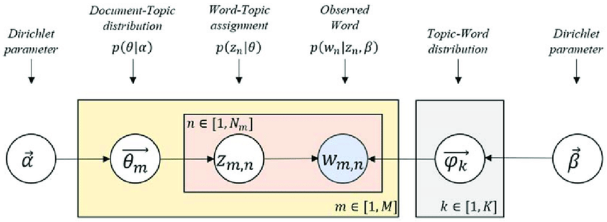


Fig1: *Top-level view of LDA and the dataset (source)*

The topic modeling technique we used for our analysis is the Latent Dirichlet Allocation (LDA). The core assumption of LDA is that “similar words are utilized to form similar topics”. LDA maps every document to a particular topic, whose topic words best cover the overall document. The generated topics might not be human interpretable, so to validate the results Coherence score is used as a metric. The coherence score is based on pointwise mutual information(PMI) amongst the word pairs.

$$PMI\ (x\ ;\ y) \equiv log \frac{p(x, y)}{p(x)\ p(y)}$$

where $x$ and $y$ are random variables.

### 3.5 Sentiment Analysis

Sentiment analysis is a text analysis technique that finds polarity (e.g., a positive or negative opinion) in text whether it's a whole document, paragraph, or sentence. It is a difficult subject, despite its simple appearance on paper. This is because a text can have multiple interpretations at the same time. For our study, we used the VADER `SentimentIntensityAnalyzer` that uses a dictionary to map lexical information to emotion intensities, which are referred to as sentiment scores. A text's sentiment score is calculated by adding the intensity of each word in the text and focus is to determine if a piece of text is *positive*, *negative*, or *neutral*.

### 3.6 Product Recommendation

We explored two ways of recommending similar products -- one involves *Ego-Networks* and the other involves *Collaborative-based Filtering*. We used Ego-Networks to give recommendations based on Product Titles and found matching titles that are in close proximity to the query product. This technique is especially useful when there are products with no reviews (can be new or unpopular) and we still want to include them when generating recommendations. On the other hand, we used item-based Collaborative Filtering to give out recommendations based on Product Reviews.

# 4 EXPERIMENTS

Before starting any evaluation, it is important to first understand the characteristics of the text in the product titles and reviews. We used the `TextBlob` package to perform Part-Of-Speech (POS) tagging to see the tag distribution. We found that the words in product titles are nouns in most of the cases. We also generated bi-grams and did a simple analysis to identify the `20`-most significant words in the titles.

## 4.1 Network Analysis

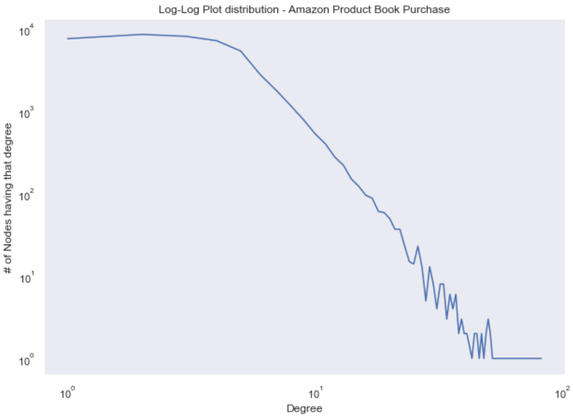


Fig2: *log-log plot for the Amazon co-purchase network*

It is evident from the plot shown in Fig2 that the network follows a heavy-tail distribution and hence obeys power law.

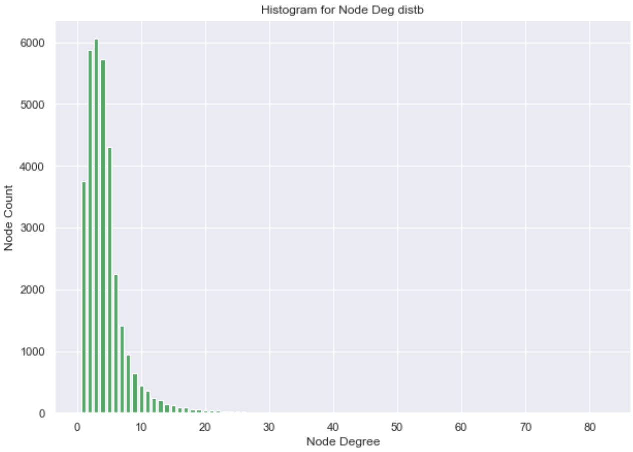


Fig3: *Histogram plot showing the degree distribution*

As we can clearly see from Fig3, the majority of the nodes have a degree under `20`.

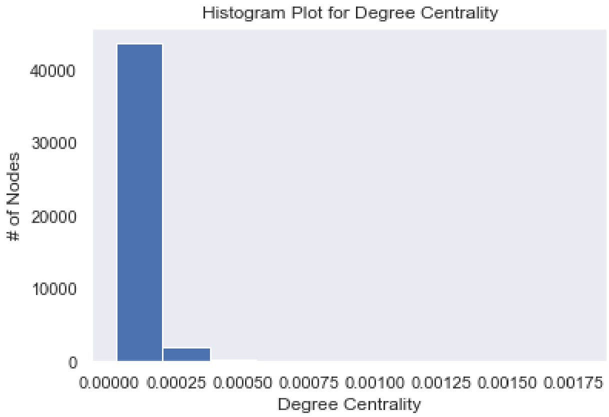


Fig4: *Histogram plot showing degree centrality*

To facilitate faster computation of the metrics, we only considered the largest component and created a sub-graph. We found the density of the sub-graph to be extremely small (`0.0001`) which indicates very low interconnections. Also, the degree centrality of the subgraph was found to be `0.0017`.

*NOTE: Many other visualizations and figures are provided in the supplementary material accompanying this report.*

## 4.2 Topic Modeling

We employed LDA for two different sets of documents. The first one being topic modeling applied to book titles with a goal to identify the top-`4` topics. The key observations we obtained from Fig2 are neatly summarized in Table-2 below-

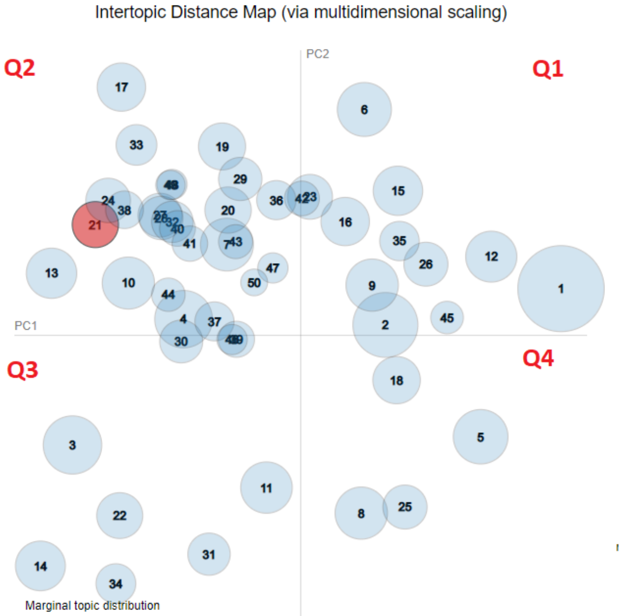


Fig5: *Topic Modeling on Book/Magazine Reviews*

| | |
|---|---|
| Topic-0 | American History, World Wars, World Maps |
| Topic-1 | Bible, Chirstianity, Compassion and Love. |
| Topic-2 | Work force, Man Power, Work life balance. |
| Topic-3 | Business Strategies, Secrets, and Guide to successful life. |

Table 2: *Topic clusters obtained using Book Titles*

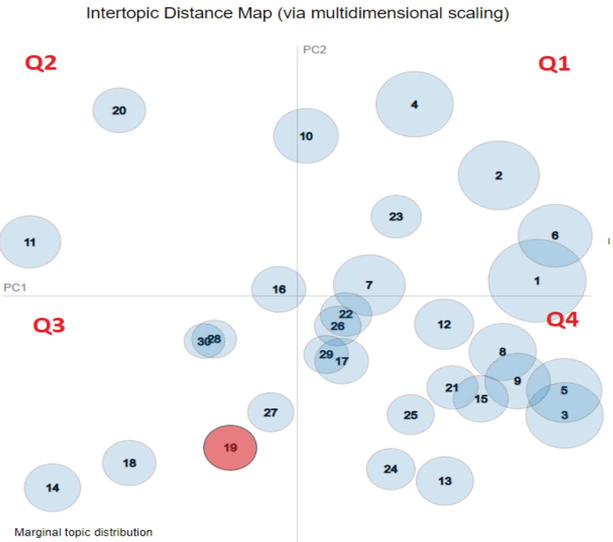


Fig6: *Topic Modeling on Appliances Reviews*

Next, we applied topic modeling for product reviews related to electronic appliances (as shown in Fig6) and compared them to those obtained from books/magazines (as shown in Fig5) as summarized in Table-3. The topic inferences are as below.

| | **Books** | **Appliances** |
|---|---|---|
| *Quadrant (Q1)* | Subscription Issues, Renewal/cancellation, Competitor worthiness (Economist, Newsweek etc.) | Appliance repair, ware & tear, Warranty & customer service, Similar product issues (water filter, fridge, etc.) |
| *Quadrant (Q2)* | American - World Culture, History, Art, Technology, Health, Sports | Samsung refrigerator, Whirlpool washing machine, Sony music system |
| *Quadrant (Q3)* | Sentiment, Product satisfaction, Pricing, Subscriptions | Product recommendation, Appliance satisfaction, Quality, Delivery |
| *Quadrant (Q4)* | Holidays, Celebrations, Acceptance among various age groups | Product defects, efficiency, defects, under age restrictions. Maintenance & Safety |

Table 3: *Comparison of Topic clusters obtained using Reviews*

It is important to note that we observed similar clustering results across the four quadrants for reviews of both Books & Electronic appliances.

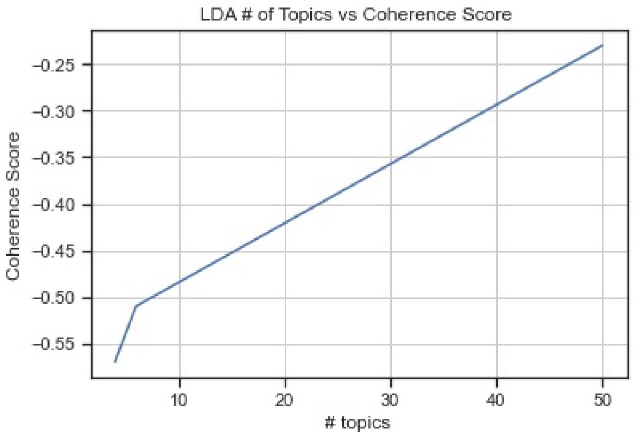


Fig7: *No. of Topics vs Coherence score*

After performing Topic Modeling, we noticed that as we increased the no. of topics from `4` to `50`, it resulted in a better coherence score. To validate this statement further, we tried to increase the `#topics` and observe the change it brings in the coherence score, but due to computational limits, we decided to attempt this in future.

### 4.3 Sentiment Analysis

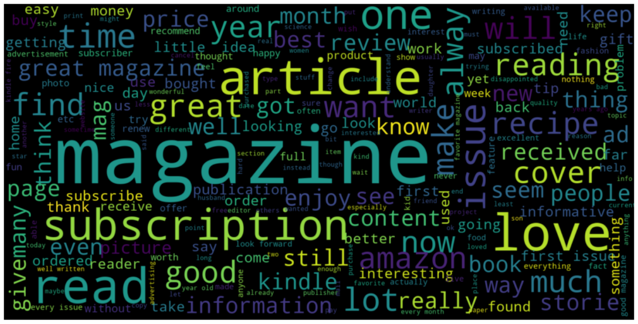
We analyzed reviews for two distinct categories -- *Books* and *Appliances*. The motivation behind our choice of the categories comes from the hypothesis that reviews for

Appliances are more skewed when compared to the reviews about Books which are often fairly neutral.

### *Analysis of Book reviews*

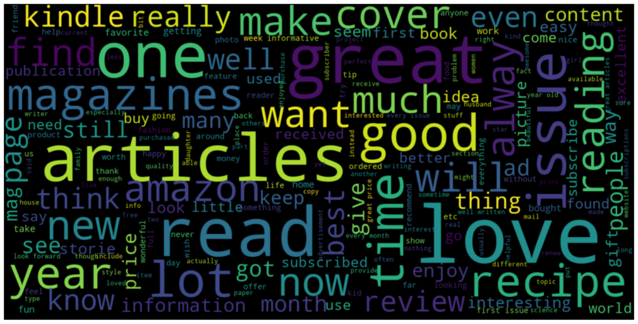
In total we gathered around 90K reviews where each review consisted of 12 columns. For our analysis we decided to keep only the columns ASIN (i.e, unique product ID), review text, overall rating, brief summary. There were around 54K reviews with an overall 5-star rating and around 5K reviews with an overall 2-star rating. We performed all the necessary pre-processing steps as discussed in section-3.2.

After we trained the VADER model using the data above, we randomly sampled reviews to validate the predicted sentiments and fine-tuned the model if necessary.



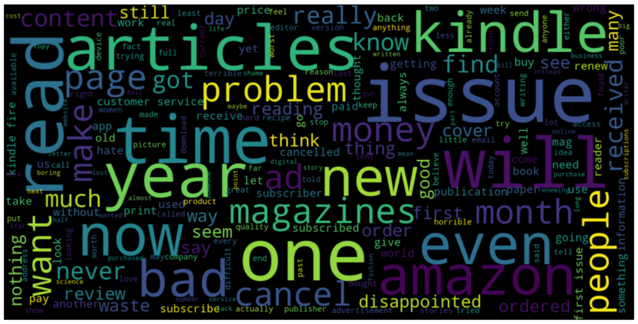
Most Frequent words in the product User Reviews

The above figure shows prominent words that appeared most frequently in the reviews under the *Books* category.



Most Frequent words in the Positive Product User Reviews

The above figure shows prominent words that appeared most frequently in the **positive** reviews under the *Books* category.



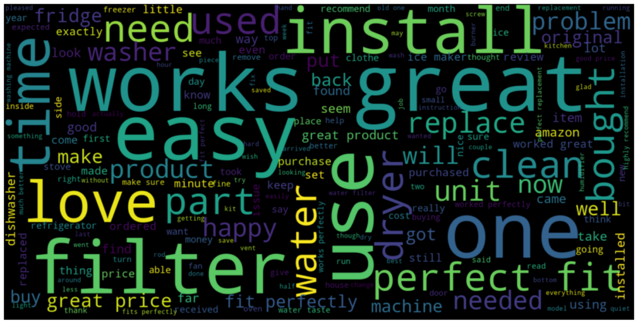
Most Frequent words in the Negative Product User Reviews

The above figure shows prominent words that appeared most frequently in the **negative** reviews under the *Books* category.

### *Analysis of Appliances reviews*

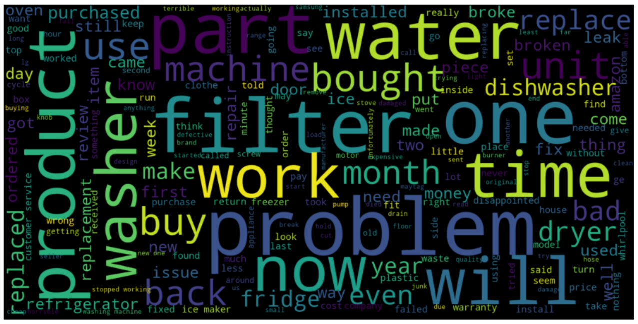
We roughly collected around 600K reviews where each review consisted of 13 columns. Once again, for our analysis we decided to keep only the ASIN (i.e, unique product ID), review text, overall rating, and brief summary columns. There were around 416K reviews with an overall 5-star rating and around 20K reviews with an overall 2-star rating. Note that we performed all the necessary pre-processing steps which we already discussed in section-3.2.

After we trained the VADER model using the data above, we randomly sampled reviews to validate the predicted sentiments and fine-tuned the model if necessary.



Most Frequent words in the Positive Product User Reviews - Appliances

The above figure shows prominent words that appeared most frequently in the **positive** reviews under the *Appliances* category.



Most Frequent words in the Negative Product User Reviews - Appliances

The above figure shows prominent words that appeared most frequently in the **negative** reviews under the *Appliances* category.

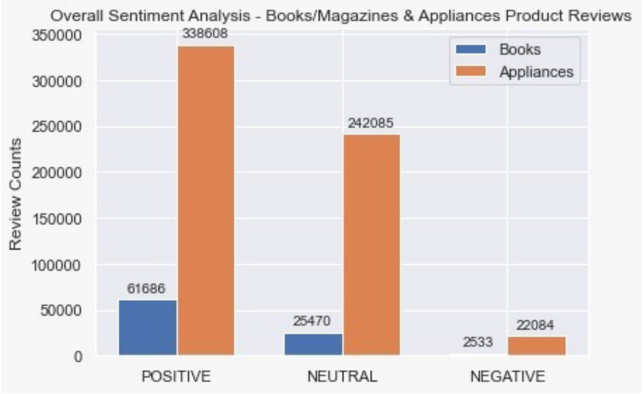


Fig8: *Sentiment comparison between Books & Appliances categories*

As we can clearly see from Fig8, the `#reviews` in the *Appliances* category significantly outnumber `#reviews` in the *Books* category. Furthermore, the VADER model predicted most of the reviews to be *positive* which indicates that the customers were happy with their purchases from either *category*.

## 4.4 Product Recommendation Analysis

### 4.4.1 Ego-Graph based recommendations

As conveyed in section-3, ego-networks provide rich information which is a strong correlation between tie strength and information diffusion in online social networks. Using the *ego_graph* in `networkx` package, for a given *product_id*, a sub_graph of close proximal neighbor nodes centered around the current *product_id* node is obtained which necessarily contains the essential recommendations. The tunable parameter here is the *radius* within which the neighbors need to be included and by increasing it one can obtain even more additional recommendations. To generate an ego-graph, network features like weighted edges, centrality scores and clustering coefficients are used.

| Product_ID - Title | Recommendations |
|---|---|
| *0805047905 - Brown Bear, Brown Bear, What Do You See?* | Amazon suggested products: -<br>● *Goodnight Moon*<br>● *Where is Baby*<br>● *Jamberry Board*<br><br>Ego-graph suggested products: -<br>● *Goodnight Moon*<br>● *Board Book & Slippers.* |
| *0486220125 - How the Other Half Lives* | Amazon suggested products: -<br>● *The Battle with the Slum*<br>● *Five Points*<br>● *Jacob Riis*<br><br>Ego-graph suggested similar products: -<br>● *The Battle with the Slum*<br>● *Old New York in Early Photos*<br>● *Jacob Riis*<br>● *Five Points* |

Table 4: *Comparing recommendations provided by Amazon and those predicted by our model*

### 4.4.2 ML-based Collaborative Filtering

While Ego-graph is useful in scenarios where the product is not prominent or doesn't have enough user reviews, for the products which do have enough ratings or comments, Machine Learning techniques are the best way to identify the nearest neighbors. On the user reviews data by utilizing the review information and the overall rating, machine learning techniques such as k-Nearest Neighbors (KNN) and Singular Value Decomposition (SVD) are applied. Root Mean Squared Error is used as a performance measure. The baseline RMSE value was `1.39` obtained when average rating is given to all products, this is compared against kNN and SVD for improvements.

| Model Type | RMSE Error |
|---|---|
| Avg. Rating-based Popularity Model | *1.39* |
| kNN (`k=10`) | *1.03* |
| SVD (`n_factors=25`) | *1.05* |
| Best Params: SVD (`n_factors=85`) | ***0.71*** |

Table 5: *Comparison of error rates given by ML models*

## 5 LIMITATIONS

During the course of the project, two major limitations hindered us from achieving what we specified in the proposal (Project Step-1 stage). Firstly, due to the massive number of nodes and high sparsity, our current machines couldn't handle processing such data and hence it was really hard to perform advanced network analysis like Community Detection and Link Prediction. The other roadblock was having limited access to fetch user opinions on Twitter about products they bought through Amazon.com. Twitter's *essential access* limits us from pulling tweets past 30 days. Since the metadata available on the SNAP community is at least a decade old, gathering tweets from that time is an impossible task. For those reasons, even detecting Cascades in the network is difficult.

## 6 FUTURE WORK

Even though our current analysis provided a way to achieve product recommendations using vital information in social media networks, there is still scope for improvement. Some ideas to work on in future include (but not limited to): -

1. Use Transformer models like BERT to understand and recommend out-of-category products that best suit a user purchase preferences.
2. Employ Spark or Hadoop clusters to analyze the massive graph data.
3. Apply Graph Neural Networks to detect latent information and determine interactions that help in generating better recommendations and revenue.

## 7 CONCLUSION

Motivated by collaborative-based filtering systems, we set out to build a unified framework for recommending relevant products to users on Amazon.com. To achieve this, we started by collecting the Amazon product co-purchasing meta-data dataset from the SNAP community and transformed it to a network structure (with products as nodes and co-purchases as neighbors in a directed graph) that can enable us to perform some basic network analysis. We observed that the network followed a power-law distribution. Next we performed Topic Modeling using LDA and evaluated the results by calculating the Coherence score for varying topic counts. We also made an interesting observation that if a product is easy to install and set up (obtained through topic modeling), then people showed willingness (obtained through sentiment modeling) to recommend the product to others. Also, for products like Books if the content appeals to all age groups, users made more purchases and gifted them during celebratory events.

## ACKNOWLEDGEMENTS

We sincerely thank Dr. Patrick Butler, Prof. Naren Ramakrishnan and the TAs for their constant support, feedback, guidance and motivation. We really appreciate their effort in creating this course that immensely helped us in expanding our horizons and learning about Social Media Networks in a practical setting.

## REFERENCES

[1] https://unctad.org/webflyer/covid-19-and-e-commerce-global-review.

[2] M. Ge, F. Persia and D. D'Auria, "Advanced Recommender Systems by Exploiting Social Networks," 2019 IEEE International Conference on Humanized Computing and Communication (HCC), 2019, pp. 118-125, doi: 10.1109/HCC46620.2019.00025.

[3] Lee, B., Plaisant, C., Parr, C., Fekete, J.-D., and Henry, N. (2006) Task Taxonomy for Graph Visualization Proc. ACM BELIV '06, .81-85.

[4] Perer, A. and Shneiderman, B. (2008). Systematic yet flexible discovery: guiding domain experts through exploratory data analysis, Proc. 13th Int'l Conf. on Intelligent User Interfaces, 109-118, New York, NY, USA,. ACM.

[5] Global and Local Feature Learning for Ego-Network Analysis via arXiV https://arxiv.org/pdf/1012.4050.pdf

[6] Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

[7] Han, Yun, et al. "Interactive assigning of conference sessions with visualization and topic modeling." 2020 IEEE Pacific Visualization Symposium (PacificVis). IEEE, 2020.

[8] P. Melville, R.J. Mooney, R. Nagarajan Content-Boosted Collaborative Filtering for Improved Recommendations, Proceedings of the 18th National Conference on Artificial Intelligence (AAAI-2002), July 2002, Edmonton, Canada.

[9] Bruke, R. Hybrid recommender systems: survey and experiments, User Modeling and User Adapted Interaction 12 (2002) 331-370.

[10] Hasan, A., Moin, S., Karim, A., & Shamshirband, S. (2018). Machine Learning-Based Sentiment Analysis for Twitter Accounts. Mathematical and Computational Applications. https://doi.org/10.3390/mca23010011.

[11] Dashtipour, Kia, et al. "Exploiting deep learning for Persian sentiment analysis." International conference on brain inspired cognitive systems. Springer, Cham, 2018.

[12] Medrouk, Lisa, and Anna Pappa. "Deep learning model for sentiment analysis in multilingual corpus." International Conference on Neural Information Processing. Springer, Cham, 2017.

[13] Cho, Kyoung-Won, Sung-Kwon Bae, and Young-Woon Woo. "Analysis on topic trends and topic modeling of KSHSM journal papers using text mining." The Korean Journal of Health Service Management 11.4 (2017): 213-224.

[14] Kherwa, Pooja, and Poonam Bansal. "Topic modeling: a comprehensive review." EAI Endorsed transactions on scalable information systems 7.24 (2020).

[15] Jelodar, Hamed, et al. "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey." Multimedia Tools and Applications 78.11 (2019): 15169-15211.

[16] Shubankar, Kumar, AdityaPratap Singh, and Vikram Pudi. "A frequent keyword-set based algorithm for topic modeling and clustering of research papers." 2011 3rd Conference on Data Mining and Optimization (DMO). IEEE, 2011.