Nokia Amazon Reviews

Mikołaj Olesiński

March 2025

Contents

1 Data Analysis

1.1 Products

The dataset contains information about 7,438 unique products. The distribution of reviews across these products is highly skewed:

- On average, each product has about 10.6 reviews.
- The median number of reviews per product is only 2.
- The most reviewed product has 3,443 reviews.
- 75% of products have 6 or fewer reviews.

This indicates that most products receive very few reviews, while a small number of products attract the majority of reviews. Only 741 products (approximately 10% of the total) have received 20 or more reviews.

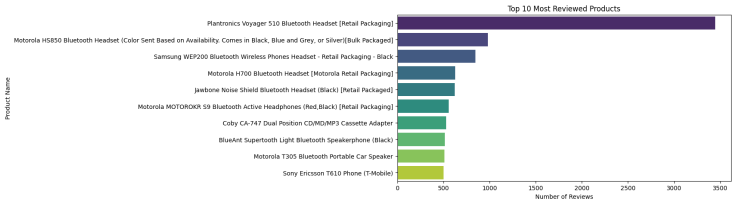

Figure 1: Top 10 most reviewed products

The most reviewed product is the Plantronics Voyager 510 Bluetooth Headset, with over 3,400 reviews. The top reviewed products are primarily Bluetooth headsets and wireless devices, suggesting these categories generate significant customer engagement. Motorola appears frequently in the top 10 list with multiple products, indicating a strong market presence.

Among products with at least 20 reviews, the highest rated is the Nokia CA-53 Connectivity Cable with an average rating of 4.80 out of 5. Other highly rated products include Motorola batteries, Plantronics accessories, and various charging cables. This suggests that accessories and replacement parts tend to receive higher satisfaction ratings than primary devices.

1.2 Users

The dataset presents a challenge in analyzing user behavior due to the presence of users with unidentified usernames. There are 2,276 reviews attributed to "unknown" users, which could potentially interfere with the analysis. To obtain more accurate scores we excluded the "unknown" users.

After filtering out the "unknown" users, the dataset contains 68,040 unique identifiable users who have submitted reviews. The distribution of review activity among these users shows:

- The average number of reviews per user is approximately 1.13
- The median is 1 review per user, meaning most users submit only a single review
- 75% of users have submitted only 1 review
- Only 251 users (about 0.37% of all users) have submitted 5 or more reviews
- The most active reviewer has submitted 47 reviews

This distribution reveals that the vast majority of users are one-time reviewers for specific products, while a very small percentage of users are responsible for a disproportionate number of reviews.

The top 10 most active reviewers have each submitted between 34-47 reviews, making them significantly more engaged than typical users. This pattern of participation is common in online review platforms, where a small group of dedicated users contribute much more content than the average user.

1.3 Ratings

The dataset contains information about 78,930 ratings. The analysis of ratings in the dataset reveals several interesting patterns:

- The average rating across all products is 3.52 out of 5, with a standard deviation of 1.52
- The distribution of ratings is bimodal, with peaks at 5 stars (38.3% of all ratings) and 4 stars (22.4%)
- 1-star ratings are the third most common (18.6%), indicating a significant portion of highly dissatisfied customers
- The least common ratings are 2 stars (9.6%) and 3 stars (11.0%)

This distribution suggests a pattern common in online review systems, where extremely positive ratings are most frequent, followed by extremely negative ratings, with fewer moderate opinions. This phenomenon may be attributed to users being more motivated to write reviews when they have had either very positive or very negative experiences with a product.

The temporal analysis of ratings shows a clear evolution in the volume of reviews over the years:

1. The number of reviews started very low in 1999 and remained minimal until 2003
2. A steady increase began in 2004, accelerating dramatically through 2005-2007
3. The peak occurred in 2007 with nearly 20,000 reviews

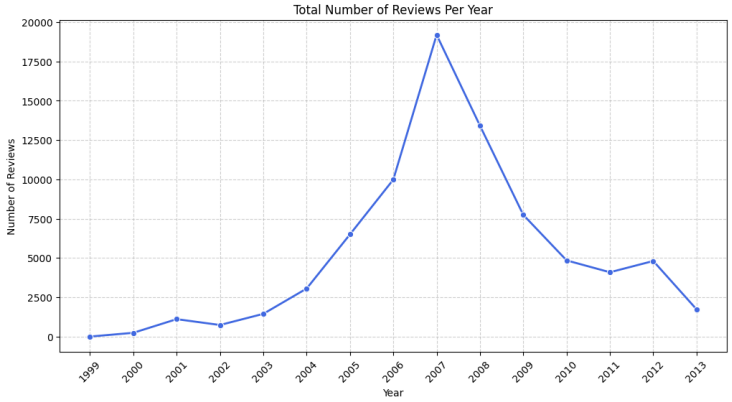

Figure 2: Number of reviews per year

4. After 2007, there was a consistent decline in review volume
5. Between 2010-2012, the number of reviews stabilized around 4,000-5,000 per year
6. In 2013, review numbers dropped to approximately 2,000

This temporal pattern may reflect changes in the popularity of the platform, shifts in consumer behavior, or changes in the marketplace for the products being reviewed. The peak in 2007 coincides with significant growth in e-commerce and online review culture, while the subsequent decline might indicate shifts to other platforms or review sources.

1.4 Reviews length

The dataset contains 78,930 reviews with varying lengths and levels of helpfulness. Analysis of these reviews reveals several interesting patterns:

- The average review length is 103 words, with a standard deviation of 123 words
- The median review length is 65 words, indicating that most reviews are relatively concise
- Review length varies considerably, ranging from extremely brief 1-word reviews to detailed reviews containing up to 2,885 words
- 25% of reviews contain 34 or fewer words, while 75% contain 123 or fewer words

A notable aspect of the dataset is the distinction between helpful and unhelpful reviews. Using a threshold of at least 10 total votes with at least 70% of those votes being helpful, we classified:

- 5,804 reviews (7.4%) as "helpful"
- 73,126 reviews (92.6%) as "not helpful"

The boxplot comparison between helpful and unhelpful reviews shows a clear difference in length distribution:

- Helpful reviews tend to be significantly longer, with a median length approximately 3 times greater than unhelpful reviews
- The interquartile range (middle 50% of data) for helpful reviews spans from about 150 to 350 words
- Unhelpful reviews are much shorter, with most containing fewer than 100 words
- Both categories contain outliers, but helpful reviews have more extreme outliers with some exceeding 2,000 words

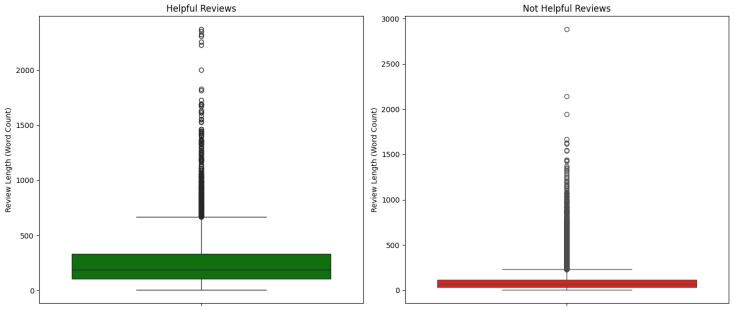

Figure 3: Box plot of review length, helpful reviews, and not helpful reviews

This pattern suggests that consumers find more value in detailed, comprehensive reviews that provide substantial information about the product. Brief reviews, while more common, tend to receive fewer helpful votes, possibly because they provide insufficient detail for making informed purchasing decisions. The distinction between helpful and unhelpful reviews has important implications for both consumers and sellers. For consumers, identifying and prioritizing helpful reviews can improve decision-making. For sellers, encouraging more detailed reviews could potentially increase the perceived value of the review system.

2 Sentiment Analysis

The sentiment of each review was determined based on its rating, with ratings of 4-5 stars classified as "Positive" and ratings of 1-3 stars classified as "Negative". This binary classification allows for easier analysis of customer satisfaction. Analysis of the sentiment distribution reveals:

- 47,970 reviews (60.8%) were classified as "Positive"
- 30,960 reviews (39.2%) were classified as "Negative"

2.1 Text Preprocessing

To prepare the review text for sentiment analysis, the following preprocessing steps were applied:

- Converting all text to lowercase
- Removing text within brackets
- Removing punctuation
- Removing words containing numbers
- Removing quotation marks and ellipses
- Removing newline characters

This cleaning process helps to standardize the text and remove elements that might introduce noise to the analysis, while preserving the content.

2.2 Model Comparison

Multiple classification models were compared to find the most effective approach:

Model	**Accuracy**	**F1-score (Negative)**	**F1-score (Positive)**
Logistic Regression	85%	0.80	0.88
Multinomial Naive Bayes	82%	0.76	0.86
Linear SVC	83%	0.78	0.86
Random Forest	81%	0.71	0.86

Table 1: Performance comparison of different classification models

Note: The Random Forest model has high precision (87%) but low recall (59%) for the negative class.

The Logistic Regression model achieved the best overall performance across all metrics, making it the preferred model for this sentiment analysis task. The Random Forest model showed interesting characteristics with high precision for negative reviews, but at the cost of low recall.

2.3 Logistic Regression Model

A logistic regression model was trained on the preprocessed review text using a bag-of-words approach (CountVectorizer). The model achieved:

- 85% overall accuracy
- 82% precision and 78% recall for negative reviews
- 87% precision and 89% recall for positive reviews

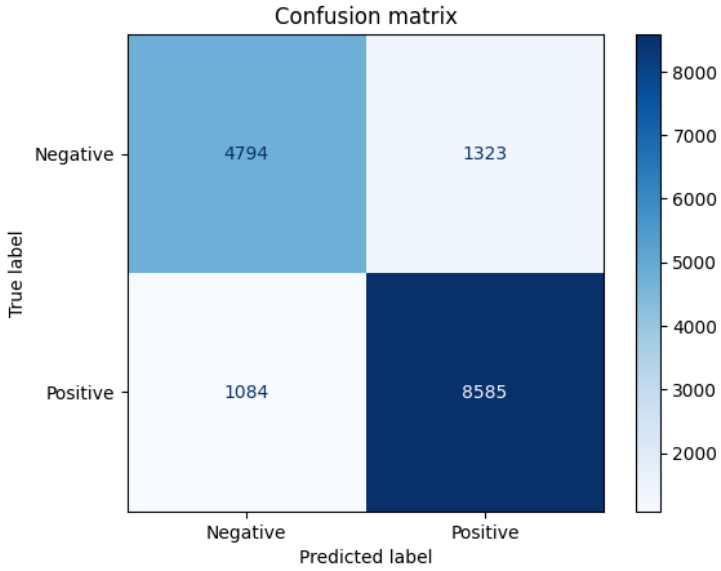

Figure 4: Confusion Matrix

2.4 Feature Importance Analysis

The logistic regression model provides insights into which words most strongly indicate positive or negative sentiment.

.

These words provide valuable insights into the specific aspects of products that drive customer satisfaction or dissatisfaction.

Top Positive Sentiment Words	Coefficient
excellent	1.94
hesitate	1.90
lipstick	1.85
restored	1.84
flawless	1.79

Table 2: Words strongly associated with positive sentiment

Top Negative Sentiment Words	Coefficient
worst	-2.22
worthless	-2.14
inconsistent	-2.01
disappointment	-1.97
dissatisfied	-1.94

Table 3: Words strongly associated with negative sentiment

2.5 Model Improvement Experiments

2.5.1 Removing Frequent Misclassified Words

Analysis of misclassified reviews revealed that certain high-frequency words appeared often in reviews that the model struggled to classify correctly. The top 30 such words were removed from the vocabulary, and a new model was trained.

2.5.2 Removing Stop Words

Another experiment involved removing English stop words (common words like "the", "and", "is) from the vocabulary. We used ready set from sklearn.

2.5.3 Conclusion

Neither approach led to noticeable improvements in model performance. Removing frequently misclassified words might have reduced useful contextual cues, while removing stop words had little effect, possibly because the model was already handling them appropriately. This suggests that the misclassifications were likely driven by deeper linguistic patterns rather than the presence of specific words.

3 Text Clustering Analysis of Product Reviews

3.1 Data Preprocessing

- Used TF-IDF vectorization
- Removed English stop words
- Selected top 10,000 features
- Transformed text into numerical matrix

3.2 Clustering Technique

The clustering was performed using the K-Means algorithm. Five clusters were created, and the random state was set to 42 to ensure reproducibility of the results.

3.3 Visualization

For visualization, Principal Component Analysis (PCA) was applied to reduce the data dimensions to 2. A scatter plot was used to display the distribution of the clusters.

Figure 5: Clusters

3.4 Cluster Characteristics

3.4.1 Cluster 0: Phone Cases and Accessories

- **Main topics:** protective cases, phone accessories
- **Key terms:** case, clip, phone, belt, leather
- **Focus:** case quality, fit, and design

3.4.2 Cluster 1: Chargers and General Products

- **Main topics:** chargers, general product experiences
- **Key terms:** product, charger, phone, price
- **Focus:** product utility and pricing

3.4.3 Cluster 2: Phones and Electronics

- **Main topics:** phone features, electronic devices
- **Key terms:** phone, camera, screen, phones
- **Focus:** broad discussions about phone characteristics

3.4.4 Cluster 3: Bluetooth Headsets

- **Main topics:** bluetooth audio devices
- **Key terms:** headset, ear, bluetooth, sound
- **Focus:** sound quality, comfort, usage

3.4.5 Cluster 4: Phone Batteries

- **Main topics:** battery performance
- **Key terms:** battery, charge, life, original
- **Focus:** battery replacement, longevity

3.5 Potential Use Cases

Clustering analysis of product reviews can be applied to product categorization, sentiment analysis, feedback patterns, and targeted marketing. By categorizing products based on reviews, businesses can better understand customer preferences. Sentiment analysis helps identify customer opinions, while feedback patterns reveal areas for improvement. Additionally, clustering can aid in creating more effective marketing strategies tailored to specific customer segments.

4 Product Recommendation on Clusters

Product recommendation systems are designed to suggest items to users based on their preferences, behaviors, and the characteristics of similar products. By leveraging clustering techniques, such as the K-Means algorithm, we can enhance the recommendation system by identifying products that share similar features. The key idea behind this approach is to group similar products into clusters and recommend products from these clusters to users based on their previous interactions with other products.

4.1 Approach

4.1.1 Clustering Products

Products are grouped into clusters using the K-Means algorithm based on their review data. Each product is assigned to a specific cluster, and the centroid of each cluster represents the central point of all the products within it.

4.1.2 User's Interaction with Products

We track the products that each user interacts with, either through reviews or ratings. By identifying the clusters of products they have already engaged with, we can recommend new products from the same cluster or similar clusters.

4.1.3 Similarity Measurement

To recommend products similar to a given one, we calculate the similarity between clusters based on the distance between the centroids. Products from clusters that are closer to the given product's cluster are considered similar.

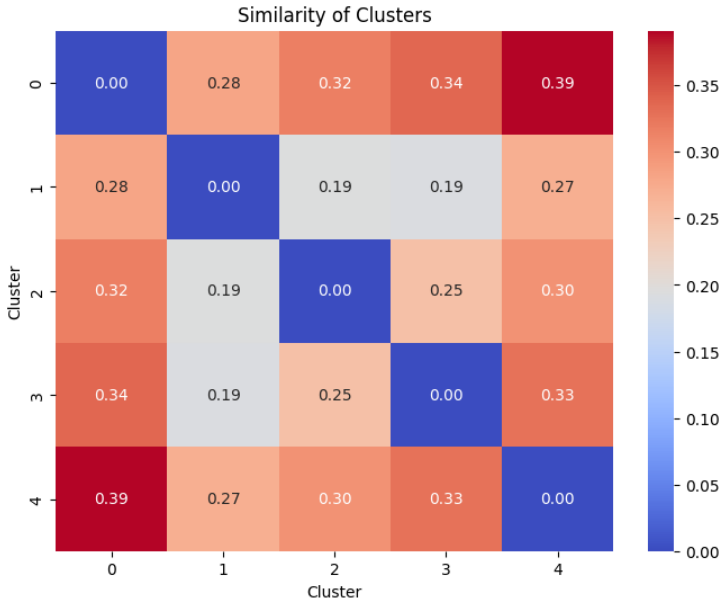

Figure 6: Clusters similarity

4.1.4 Recommendation Score

For each candidate product in the neighbor clusters, a composite score is calculated based on several factors:

- **Rating:** The mean rating of the product, which indicates its overall popularity and satisfaction.
- **Rating Count:** The number of reviews, reflecting the product's exposure and reliability.
- **Helpful Votes:** The sum of helpful votes, providing a measure of how useful others found the reviews.

The composite score is calculated as:

$$\text{Composite Score} = (\text{rating mean} \times 1.2) + (\log(\text{rating count}) \times 0.3) + (\log(\text{helpful votes}) \times 0.1)$$

This ensures that products with higher ratings, more reviews, and more helpful votes are ranked higher.

4.2 Personalized Recommendations

For personalized recommendations, we consider the user's previously rated products and the clusters they belong to. We then recommend products from

these clusters or from nearby clusters with similar characteristics, ensuring that the recommendations are tailored to the user's preferences.
These recommendations are based on the similarity of clusters where the user has rated products highly.

4.3 Recommended Products for a Specific Product

For example, for the product Motorola A630 Phone (T-Mobile) (Product ID: B0003RA29O), the following similar products are recommended:

- Motorola Cigarette Lighter Adapter for Motorola Phones (Product ID: B0007N08NO) with an average rating of 4.38
- Plantronics Voyager 510 Bluetooth Headset [Retail Packaging] (Product ID: B0009B0IX4) with an average rating of 4.16
- Sunforce 44447 900W Whisper Wind Turbine (Product ID: B000FIUSCC) with an average rating of 5.00
- Nokia CA-53 Connectivity Cable (Product ID: B000JJH2VW) with an average rating of 4.80
- Multi-Use Vehicle Charger with Dual USB Ports and Dual 12 Volt Sockets - Magnadyne (Product ID: B000PB8CQI) with an average rating of 4.61

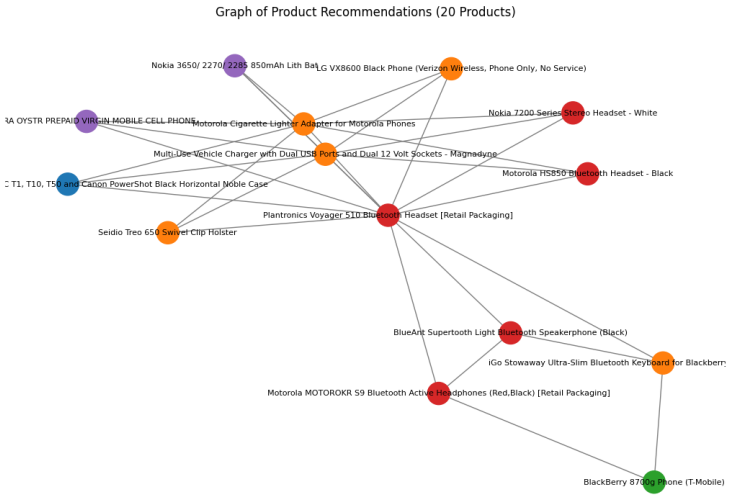

Figure 7: Graph of recommended products

5 NCF Rating Prediction Model with PyTorch

In addition to the cluster-based recommendation approach, we developed a neural collaborative filtering (NCF) model to predict ratings by users. This deep learning-based method leverages neural networks to capture complex, non-linear interactions between users, products and ratings.

5.1 Model Architecture

The Neural Collaborative Filtering (NCF) model consists of several key components:

- **User and Product Embeddings:** Learned low-dimensional representations of users and products
- **Multi-Layer Perceptron (MLP):** A neural network that captures non-linear interactions
- **Bias Regularization:** Incorporated to prevent overfitting and improve generalization

The model architecture includes:

1. User and product embedding layers with 32-dimensional embeddings
2. A multi-layer neural network with:
    - First hidden layer: 128 neurons with batch normalization and ReLU activation
    - Second hidden layer: 64 neurons with batch normalization and ReLU activation
    - Dropout layers to prevent overfitting
    - Final output layer predicting the rating
3. Bias terms for users and products to capture individual rating tendencies

5.2 Training Process

The training process involved several key steps:

- **Data Preprocessing:** Filtering users and products with minimum review thresholds
- **Feature Engineering:** Creating user and product indices, calculating global mean and bias terms
- **Model Training:** Using Adam optimizer with learning rate scheduling
- **Loss Function:** Mean Squared Error with bias regularization
- **Early Stopping:** Preventing overfitting by monitoring validation loss

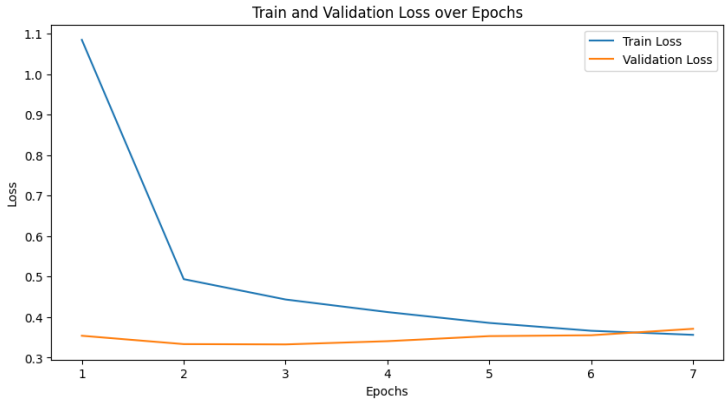

Figure 8: Train Process

5.3 Performance Metrics

The model's performance was evaluated using standard regression metrics:

Metric	**Value**
Root Mean Square Error (RMSE)	0.4554
Mean Absolute Error (MAE)	0.2338

Table 4: Neural Collaborative Filtering Model Performance

The model demonstrates promising performance in predicting user ratings, with relatively low error rates. The low RMSE and MAE indicate that the model can effectively capture user-product interactions.

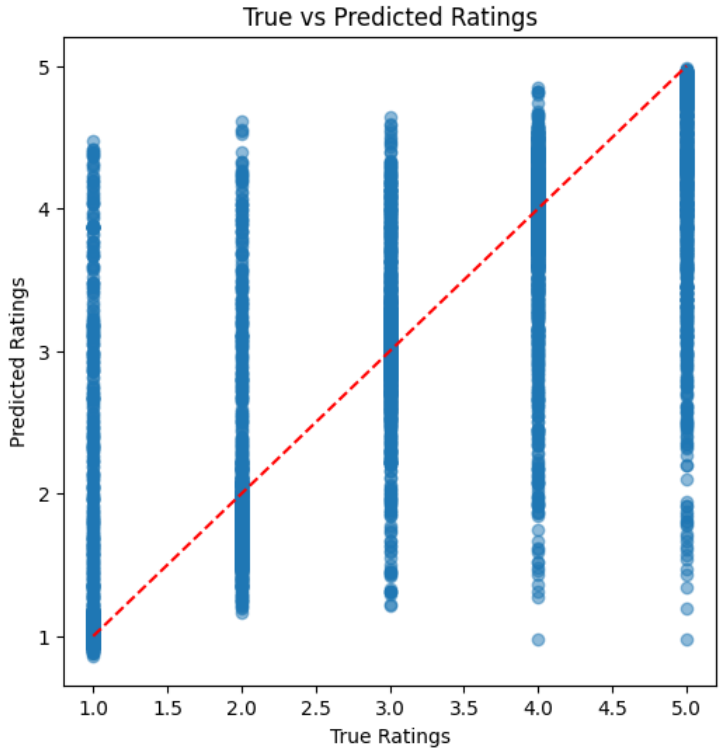

Figure 9: True vs Preticted rating

5.4 Key Advantages

The NCF approach offers several advantages over traditional recommendation methods:

- **Non-linear Interactions:** Captures complex user-product relationships
- **Learned Representations:** Automatically learns meaningful embeddings
- **Scalability:** Can handle large datasets with multiple features

5.5 Limitations and Future Work

While the model shows promising results, potential improvements could include:

- Incorporating more features beyond ratings
- Can be used in recommendation model

- Experimenting with deeper network architectures
- Implementing advanced regularization techniques
- Exploring alternative embedding dimensionalities

6 Word Embedding NLP

Word embeddings are a powerful technique in Natural Language Processing (NLP) that transform words into dense vector representations, capturing semantic relationships between words.

6.1 Model Creation and Training

The Word2Vec model was trained on the combined text of review summaries and full review texts. Key parameters include:

- Vector size: 100 dimensions
- Window size: 5 words
- Minimum word count: 2
- Training epochs: 15

6.2 Word Similarity Analysis

The model demonstrates interesting semantic relationships:

- Words similar to 'bad': good, terrible, horrible, shabby, poor
- Similarity between 'good' and 'great': 0.79
- Similarity between 'fast' and 'quick': 0.78

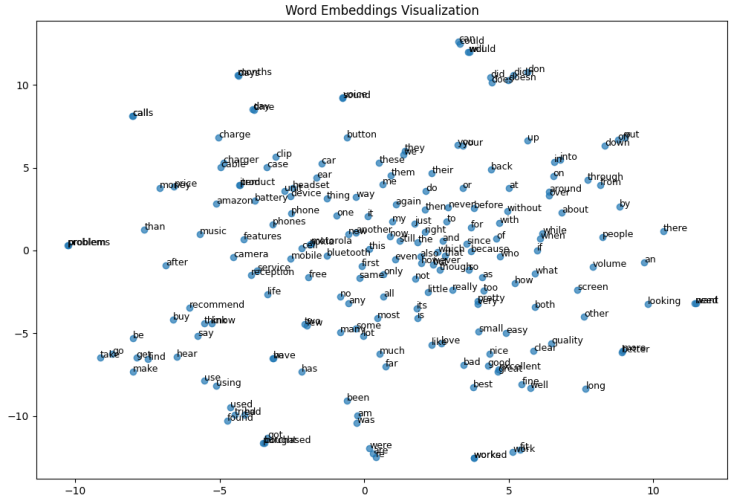

Figure 10: Graph of recommended products

6.3 Sentiment Analysis with Mean Vector Approach

6.3.1 Methodology

This approach differs from traditional text classification by using mean word embeddings to represent entire reviews. Each review is converted to a vector by averaging the word vectors of its constituent words.

6.3.2 Limitations of the Approach

The mean vector method showed less promising results compared to the Word2Vec approach, which could be due to several factors:

- Loss of word order information
- Averaging can blur important meaning
- Simple addition of values might miss complex language patterns

6.3.3 Performance Metrics

Despite limitations, the model achieved:

- Overall accuracy: 79
- Precision for Negative reviews: 0.75
- Precision for Positive reviews: 0.81

Unfortunately, the results are not optimal and the original version of Sentiment Analysys is better.

6.4 Product Feature Extraction

6.4.1 Objective

Product feature extraction aims to automatically identify and analyze key characteristics of products based on review text. This technique helps businesses understand customer perceptions and product strengths/weaknesses.

6.4.2 Methodology

Our approach involves:

- Tokenizing review text
- Filtering out stop words
- Identifying frequent, meaningful features
- Finding semantically related words using word embeddings

6.4.3 Example Analysis: Blackberry Headset

For the OEM Original Blackberry Stereo Headset, analysis shows that common terms like "price," "quality," "echo," and "time" frequently appear in reviews. However, the overall rating is not very high (3.44/5), indicating that the product might have some issues in these areas. Specifically, words like "price" and "quality" suggest some dissatisfaction with the cost and performance, while "echo" points to sound issues. This should be addressed to improve customer satisfaction.

6.5 Potential Applications

Product feature extraction can be used for:

- Identifying product strengths and weaknesses
- Guiding product development
- Understanding customer preferences
- Competitive analysis

6.6 Similar Product Recommendation

6.6.1 Recommendation Approach

The recommendation system leverages word embeddings to measure semantic similarity between products based on their review texts. Key steps include:

- Extracting top features from product reviews

- Calculating semantic similarity between product features
- Ranking products based on feature similarity

6.6.2 Similarity Calculation

Similarity is computed by:

- Comparing word embeddings of product features
- Using cosine similarity between feature vectors
- Averaging similarity scores across multiple features

6.6.3 Example Recommendation

For a sample product (Motorola HS850 Bluetooth Headset), the top similar products were:

1. Motorola HS850 Bluetooth Headset - Black (Similarity: 0.228)
2. Jabra BT150 Bt Headset (Similarity: 0.227)
3. Bt 250v (Similarity: 0.226)
4. Motorola HS810 Bluetooth Headset (Similarity: 0.226)
5. Jabra UJC250 2.5Mm Handsfree (Similarity: 0.220)

6.6.4 Potential Use Cases

Semantic similarity-based recommendation can be applied in:

- E-commerce product suggestion systems
- Personalized recommendation engines
- Cross-selling and upselling strategies

7 Negative Review Knowledge Graph Analysis

7.1 Methodology

We developed a novel approach to analyze negative product reviews using a knowledge graph, which allows us to extract and visualize complex relationships between products, features, and negative sentiments. The methodology involves several key steps:

- **Negative Review Identification:** Select products with at least 10 reviews and an average rating below 2.7

- **Feature Extraction:** Use natural language processing to identify:
  - Nouns representing product features
  - Adjectives describing product characteristics
  - Negative sentiment words
- **Knowledge Graph Construction:** Create a multi-directional graph connecting:
  - Products as central nodes
  - Extracted features and descriptors as connected nodes

7.2 Findings

7.2.1 Top Negative Review Features

Our analysis revealed the most frequently mentioned features in negative reviews:

Feature	Count	Avg Rating
phone	8,542	2.52
time	1,475	2.30
problem	1,147	2.44
battery	1,085	2.49
call	1,059	2.41
product	1,032	2.20
service	857	2.30

Table 5: Top Features in Negative Reviews

7.2.2 Graph Statistics

The knowledge graph provides a comprehensive view of negative product experiences:

- **Total Nodes:** 7,849
- **Total Edges:** 113,310

7.3 Insights and Implications

7.3.1 Product Improvement

The knowledge graph reveals critical areas for product development:

- **Phone Performance:** "phone" and "battery" are top negative features

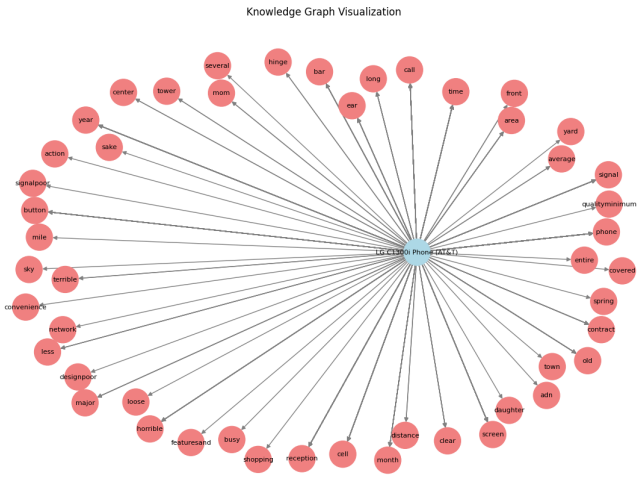

Figure 11: Visualization of graph for one product

- **Service Quality:** "service" and "call" indicate potential telecommunications issues
- **Time and Usability:** "time" suggests reliability and performance concerns

7.3.2 Recommendation Systems

The graph can enhance recommendation systems by:

- Identifying semantically related products
- Understanding user pain points
- Filtering recommendations based on negative feature analysis

7.4 Future Work

Potential extensions of this research include:

- Developing more sophisticated sentiment analysis techniques
- Creating predictive models for product improvements
- Expanding the knowledge graph to include more diverse data sources

8 Dataset Analysis Challenges

1. **Anonymous Reviews**
   - 2,276 reviews lack user identification, potentially skewing analysis.
2. **Limited User Engagement**
   - Average 1.16 reviews per user, making meaningful patterns hard to detect.
3. **Data Incompleteness**
   - Missing crucial product context like pricing, categories, specifications.
4. **Unbalanced Review Distribution**
   - Most products have fewer than 6 reviews.
   - Extreme concentration of reviews on few popular items.
5. **Recommendation Model Barriers**
   - Insufficient data to create personalized recommendations.
6. **Semantic Complexity**
   - Short, often subjective reviews make accurate sentiment analysis difficult.
7. **Platform Bias**
   - Potential self-selection bias from users motivated to write extreme (very positive/negative) reviews.