JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 1644-1654
https://publishoa.com
ISSN: 1309-3452

Data Analysis of online product reviews

Vidya Kamma*, Sridevi Gutta, D. Teja Santosh

[1,2]Research Scholar, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vadeswaram , -A.P., India.& Assistant Professor Department of Computer Science,Neil Gogte Institute of Technology
[3] Professor, Koneru Lakshmaiah Education Foundation, Vaddeswaram, A.P., India
[4]Associate Professor CSE,CVR College of Engineering,Vastunagar, Mangalpalli (V),Ibrahimpatnam, T.S.-501 510.
E-mail: [1,2]kammavidya@gmail.com *, [3]sridevi.gutta2012@gmail.com, [4]*tejasantoshd@gmail.com*

Visual exploration of the content of the online product reviews is one of the most important tasks in text mining and specifically in sentiment analysis research. However, many text explorations have produced and visualized the characteristics and the outputs of the language model namely word count, character length, word sequences and so on. The main focus of this exploratory data analysis is to explore the crucial pieces of reviews information namely nouns and adjectives scripted by the reviewers and to understand the opinion orientations of the aspects. These are useful in determining the statistical recommendations based on sentiments. The research prospective gained from this exploratory analysis is that the statistical recommendations be explained by using expressive semantic web rules of the ontology in the model-agnostic manner.

Keywords: online reviews, reviews profiling, aspects, opinion orientation, sentiment, recommendations

I. INTRODUCTION

The online shopping websites of today are providing one of many information services by writing the consumer review and encouraging rapid growth through the general comments of consumers [1]. For better understanding of these consumer reviews the available relationships among the crucial pieces of the review namely product aspects and opinions are to be explored. Reviewing and implementing content-based recommendations has a major impact on consumer opinions written on product reviews today [2]. Visual exploring these crucial pieces helps to uncover the nous and adjectives written in them. Subsequently, the quantified details of these adjectives are useful in downstream task of machine learning [3] in predicting the product recommendations. However, many text explorations have produced and visualized the characteristics and the outputs of the language model [4] namely word count, character length, word sequences. By profiling the reviews data by using summarization techniques and analyzing the expected relationships the corresponding insights are obtained in the clearer manner. The focus of this Exploratory Data Analysis (EDA) is to explore the crucial pieces of reviews which are nouns and adjectives scripted by the reviewers and to understand the polarities of these pieces so that the statistical recommendations based on polarities (aggregated to sentiments) are useful in explaining the obtained recommendations by using ontological analysis.

II. LITERATURE REVIEW

The survey on Exploratory Data Analysis of e-commerce product reviews is less touched upon by the sentiment analysis research community. According to Indrila Ghosha et al. [5] the different exploratory tools are examined for developing exploratory analyses. Several tools for data exploration have been described. There are two types of data analysis: classical and exploratory, described by author John T. Behrens [6]. A study has been carried out by Chokey Wangmo [7] on the bank lending to SMEs in Bhutan. Matthew. An exploratory study by Ntow-Gyamfi et al. [8] examines credit risk and loan default among Ghanaian banks. Francis Jency et al. [9] have used machine learning techniques to analyze data and make predictions about loan default. In an exploratory study, K. Ulaga Priya1 et al. [10] used random forests to predict loan privileges for customers. Exploratory data analysis was conducted using the R programming language. Bobumil M. Konopka et al. [11] conducted an exploratory analysis of the results of a clinical trial. Developing a method for exploring data in multidimensional space.

Heidi Nguyen and Aravind Veluchamy made [12] Machine-learning approaches and lexicon-based approaches for comparing sentiment analysis and product reviews. The authors performed exploration of product reviews and used the obtained attributes as features for machine learning and lexicon-based analysis. Wanliang Tan et al., considered [13] Reviews and ratings of Amazon products are correlated. Santosh et al., performed [14] EDA on the collected amazon product reviews and used the identified features for formulating and testing the machine learning hypothesis towards recommending products.

The research on knowledge graph based explainable recommendation has provided motivations for advancing the research by opening the avenues to work with advanced knowledge representation and reasoning approaches. Ai et al., constructed [15] knowledge graph by establishing different types of relations analyzed from the reviews collection. The explanation paths are the results of the explicit user queries. Recently, Xie et al., constructed [16] By optimizing several objectives simultaneously, Explainable recommendation frameworks improve precision, diversity, and explainability.

Peake and Wang relied [17] on model-agnostic approach to explain the recommendations. The authors used association rule mining to obtain the association rules from the matrix factorization model. The researchers showed that if a matrix factorization model recommends an item. In the case that it follows association rules, it can also be recommended. In this case, the item will be explained by the rules.

III. PROPOSED WORKS

IV.

The Exploratory Data Analysis of the dataset is carried out by following the steps namely: (i) Distinguish attributes by pre-processing (cleansing) the dataset, (ii) Univariate Data Analysis, (iii) Bivariate Data Analysis, (iv) Detect interaction among the attributes, (v) Feature Engineering. The EDA steps are depicted in figure 1 below.

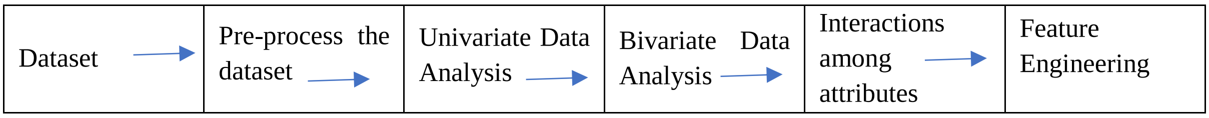


Fig 1. Steps in EDA process

We have collected five categories of Amazon product reviews for this task. In addition to laptops, smart phones, cameras, books, and wristwatches, we consider laptop reviews for analysis. (342) Three hundred forty-two products are considered from the E-commerce applications

There are 67,986 reviews in total. There is an opinion for each item in each review. One product is used as a query case [3] in this dataset, while the remaining two products are used as recommendations based on the query case. Detailed information about the collected reviews can be found in table 1

Table 1. Collected Reviews Details

Document Attributes	Values
Minimum sentences per review	1
Maximum sentences per review	43
Minimum number of words written in the review	20
Maximum number of words written in the review	574

The primary task to be performed on these product reviews before carrying out EDA is reviews pre-processing. First, the reviews are cleansed by eliminating stop words. A stop word list based on its relative importance for aspect identification is manually compiled. Next, the contractions present in the reviews are expanded. The contractions are expanded for better analysis of the reviews. Finally, Parts-of-Speech (PoS) tagging of the lemmatized reviews is carried out. To gain insights from these tagged reviews, rating data is extracted to provide further analysis. As shown in figure 2, below is the distribution of reviews' star ratings.

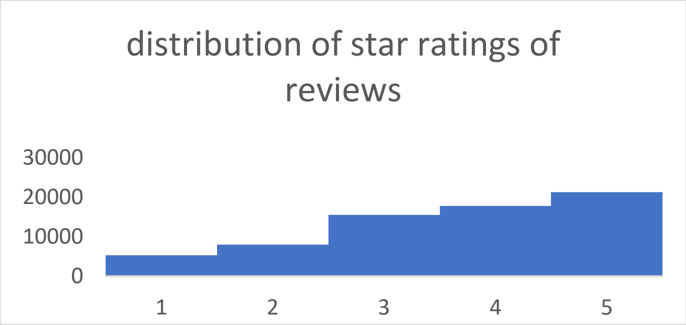
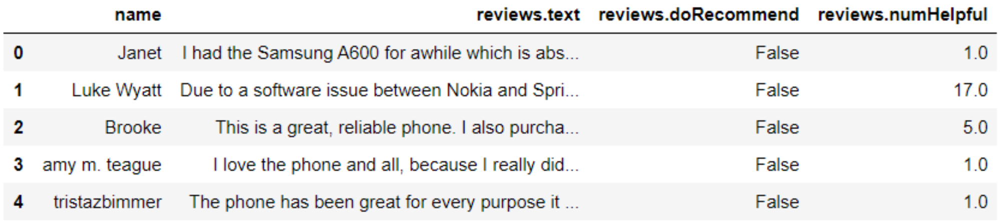


Fig 2. Review ratings distribution

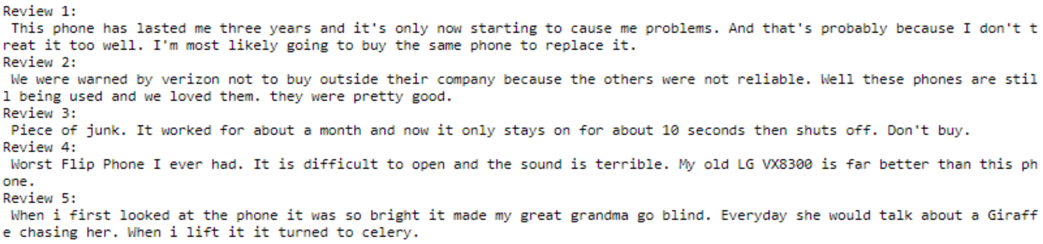
It is observed from the above figure that majority of the reviews are given with high ratings of 4 and 5 respectively. The head (first five reviews) from the collected reviews is shown in figure 3 below. The shape of the review's dataset is 67896 rows and 11 columns.



	name	reviews.text	reviews.doRecommend	reviews.numHelpful
0	Janet	I had the Samsung A600 for awhile which is abs...	False	1.0
1	Luke Wyatt	Due to a software issue between Nokia and Spri...	False	17.0
2	Brooke	This is a great, reliable phone. I also purcha...	False	5.0
3	amy m. teague	I love the phone and all, because I really did...	False	1.0
4	tristazbimmer	The phone has been great for every purpose it ...	False	1.0

Fig 3. Head of the reviews dataset

Some of the smartphone reviews from the reviews section in the dataset are highlighted in Figure 4 below.



```
Review 1:
 This phone has lasted me three years and it's only now starting to cause me problems. And that's probably because I don't t
reat it too well. I'm most likely going to buy the same phone to replace it.
Review 2:
 We were warned by verizon not to buy outside their company because the others were not reliable. Well these phones are stil
l being used and we loved them. they were pretty good.
Review 3:
 Piece of junk. It worked for about a month and now it only stays on for about 10 seconds then shuts off. Don't buy.
Review 4:
 Worst Flip Phone I ever had. It is difficult to open and the sound is terrible. My old LG VX8300 is far better than this ph
one.
Review 5:
 When i first looked at the phone it was so bright it made my great grandma go blind. Everyday she would talk about a Giraff
e chasing her. When i lift it it turned to celery.
```

Fig 4. Smartphone reviews

The stop words are removed from the reviews. The count of top 15 stop words across the collected reviews dataset is plotted are bar chart and is shown in Figure 5 below.

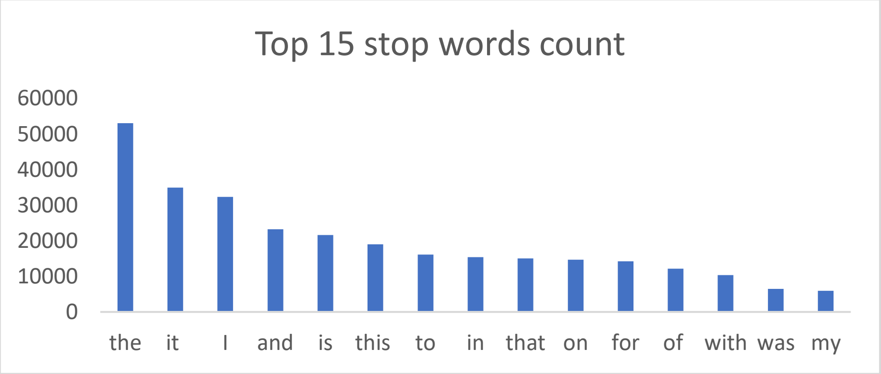


Fig 5. Top 15 stop words count

The exploration of the review text provides crucial insight about the number of review words will be useful as prospective nouns and adjectives. The comparison among the bigrams before and after removing the stop words has been carried out. Below are distributions of top bigrams before and after removing stop words in figures 6- 7.

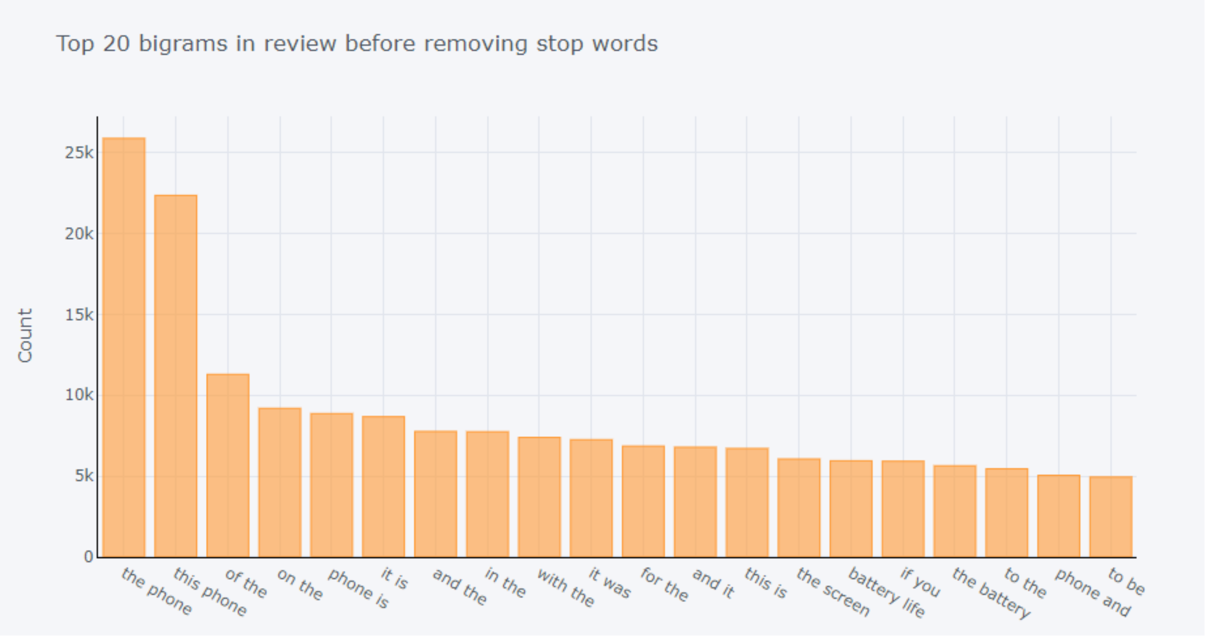


Fig 6. 20 bigrams before removing stop words

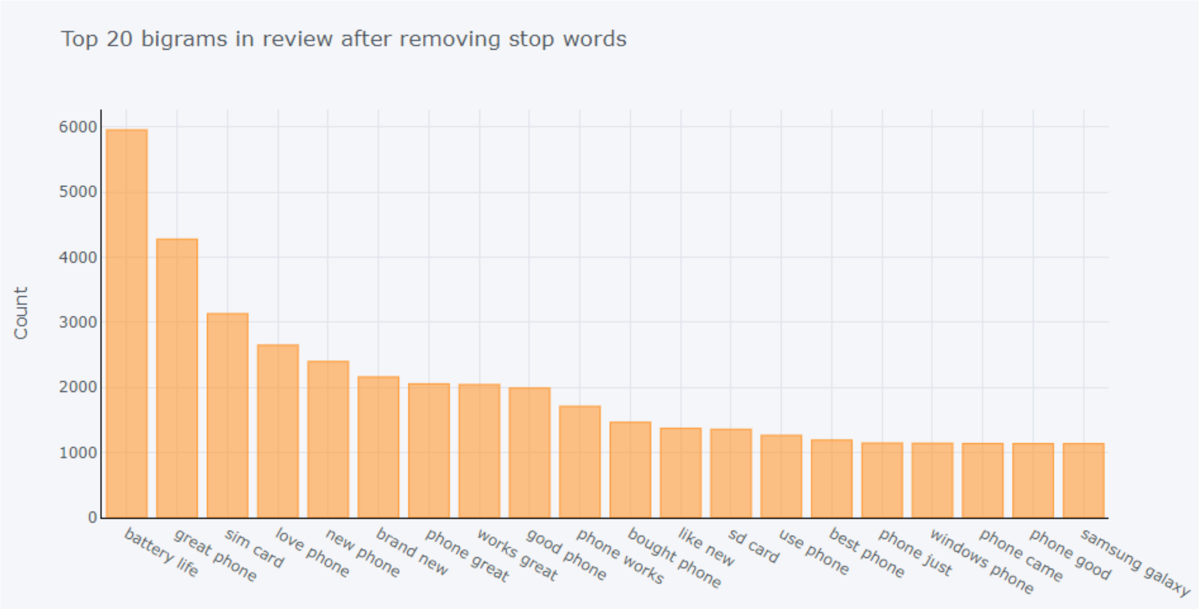


Fig 7. 20 bigrams after removing stop words

Further, the reviews after removing the stop words are PoS tagged. The top 10 PoS tags from the tagged reviews are shown in Figure 8 below.

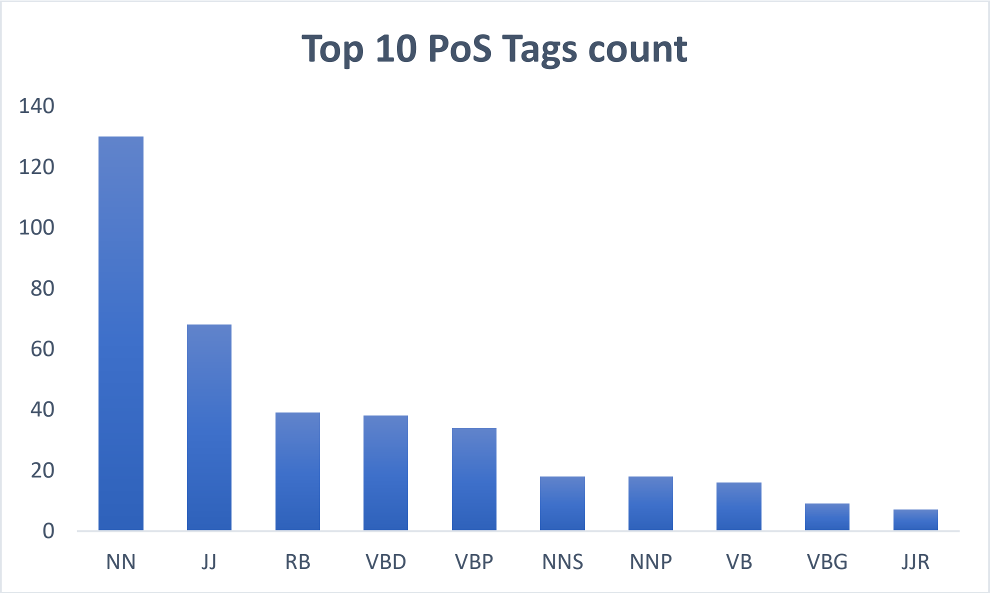


Fig 8. Top 10 PoS Tags count from the pre-processed reviews

It is observed that from the figure that nouns (NN) and adjectives (JJ) are written more in the reviews. The frequent nouns and their associated adjectives are deemed as product aspects and opinions. In order to find the frequent aspects written across various smartphones, word cloud is utilized to visualize the aspects and opinions. These are shown in figure 9 below.

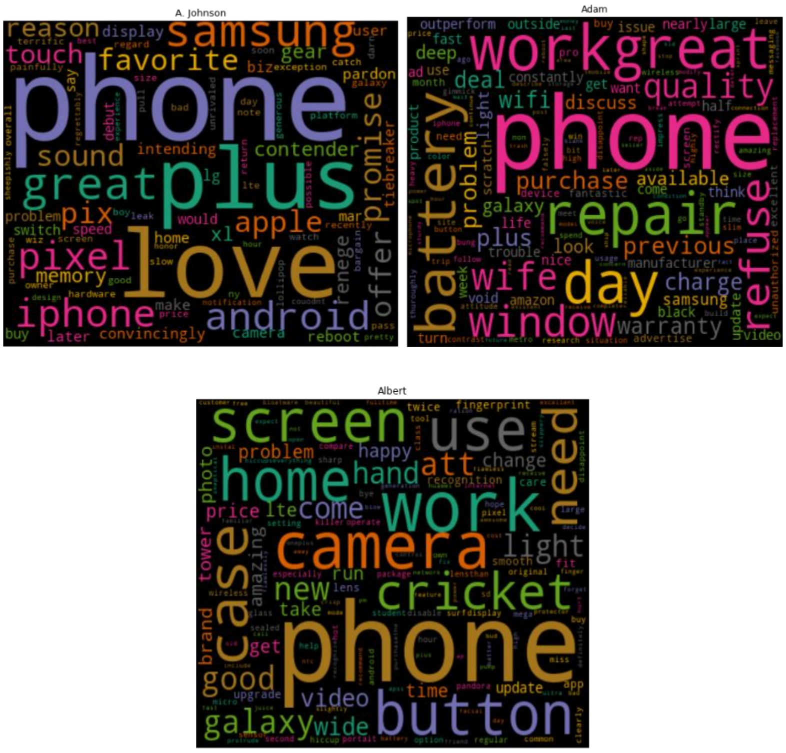


Fig 9. Frequent nouns and the adjectives of smart phones

Comparing the sentiment polarity score, the rating, and the length of reviews is depicted in Figure 10. As we can observe, Nokia brand scored the highest in sentiment polarity, and Motorola scored the lowest. As well, the Samsung's polarity score is the lowest. Review numbers for Samsung mobiles are the lowest. Due to this reason, there isn't as much variety in its score distribution as the other brands.

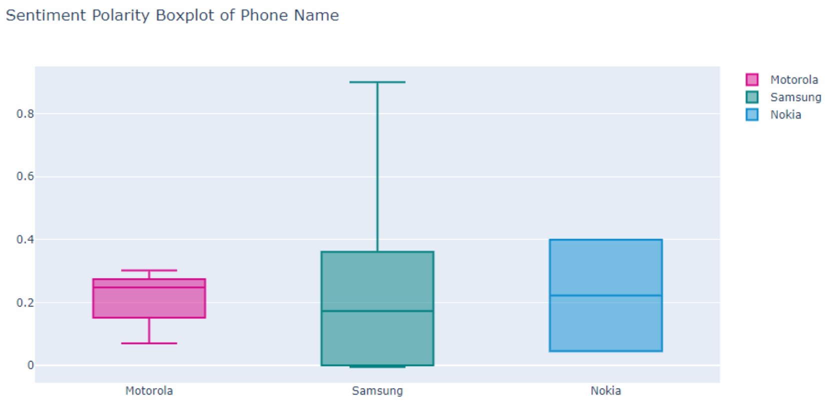


Fig 10. Sentiment Polarity Boxplot of Phone Name

In figure 11, the median review length of Motorola and Samsung brands are visualized. It is observed that the median review length of Motorola and Samsung brands are relative lower than Nokia brand.

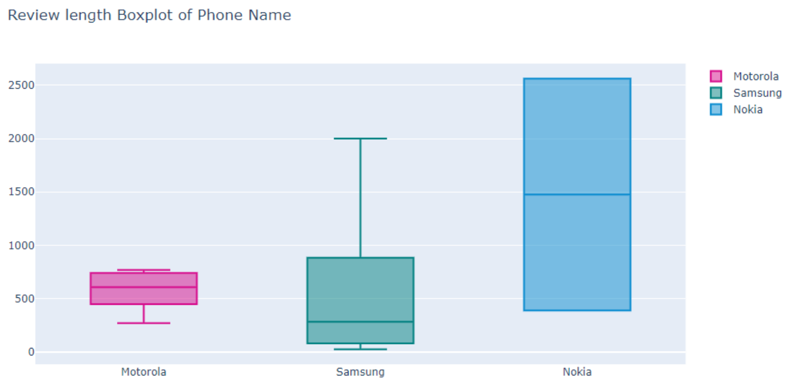


Fig 11. Review Length Boxplot of Phone Name

It is tacitly implicit that opinions expressed in reviews about product aspects provide knowledge about recommendations [3]. As certain parts of the reviews describe aspects and opinions of the products, harnessing these critical pieces of review content can solve the problem of rating-based product recommendations. Reviewing POS tags across query case and base case results is aggregated for nouns, adjectives, adverbs, and verbs. Adjectives, adverbs, and verbs are counted as product aspects [18], and nouns are counted as opinions [1]. A shift in valence is also evaluated in these opinions [19]. As a result of the interaction between factors and opinions, we have been able to develop variables that summarize and illuminate the relationships between these crucial pieces.

Derived variables are those based on the opinion orientations of the extracted product aspects. As shown below, semantic orientation (SO) can be determined using this expression.

$$SO(opinion) = \frac{dist(opinon, bad) - dist(opinion, good)}{dist(good, bad)} \qquad (1)$$

Dist(term1, term2) is defined as the difference between the sentiwordnet score of the opinion term (term1) and the sentiwordnet score of the seed term (term2). In this case, good and bad terms are considered. Distance is calculated as follows:

dist (term1, term2) = sentiwordnetscore(term1) – sentiwordnetscore(term2)   (2)

To assign a sentiment score to an aspect, the SO scores are aggregated at individual aspect level. The sentiment score variable is a normalized variable as product aspects have different types of opinion orientation counts. This is performed with the help of following formula.

$$Sent(Ai,P) = \frac{Pos\_Opinion\_Count(Ai,P) - Neg\_Opinion\_Count(Ai,P)}{Pos\_Opinion\_Count(Ai,P) + Neg\_Opinion\_Count(Ai,P) + Neu\_Opinion\_Count(Ai,P)} \qquad (3)$$

Generally, Ai determines sentiment, while P describes the current product. Senti(Ai, P) denotes the sentiment of the aspect, Pos_Opinion_Count(Ai, P) denotes the number of positive opinions obtained on the aspect A, Neg_Opinion_Count(Ai, P) the number of negative opinions obtained on the aspect A and Neu_Opinion_Count(Ai, P) the number of neutral opinions obtained on the aspect A.

An E-commerce search is a query case, and a base case is a similar product category. In order to allow for comparisons between cases, it is essential to guarantee some minimal set of aspects are shared between them. Performing the comparison requires a k-compatibility. k aspects must be shared by Bu and Bv within the same case. This is a boolean property of two cases Bu and Bv. Equation 1 illustrates how base cases that are at least k-comparable (share at least k aspects with the target query case Q) are selected during retrieval. In product cases, BC represents the base case.

Retrievek (Q) = {B $^{\varepsilon}$ BC: k – comparable (Q, B)}   (4)

In order to determine product similarity, the traditional cosine similarity is used.

$$Cos(Q,B_i) = \frac{\sum_{i=1}^{n} Q.B_i}{\sqrt{\sum_{i=1}^{n} (Q*Q)*(Bi*Bi)}} \qquad (5)$$

Customers can access the list of recommended products through the recommender system. The polarity score of reviews is an indicator of whether it will be recommended or not. On figure 12, you can see the sentiment polarity of reviews that were based on recommendations.

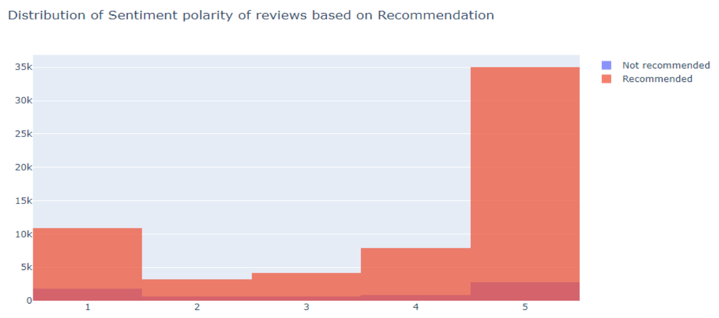


Fig 12. Distribution of Sentiment polarity of reviews based on Recommendation

The precision, recall, and F1-score metrics are used for evaluating the utility of the recommendations produced by the recommender system. The following formulas calculate precision, recall, and the F1-score.

Precision = (#recommended items that are relevant) / (# of recommended items)

Recall = (# of recommended items that are relevant) / (total # of relevant items)

F1-Score=2*(Precision*Recall) / Precision+Recall

By looking at the number of users in the dataset, the 'k' value represents similar preferences regarding products. Using opinion orientation counts, feature level sentiments were calculated. This cold start problem was eliminated due to the case-based recommender system. The 'k' value describes how many features of similar products are considered when recommending similar products.

Table:2 Information retrieval measures on the product recommendations

k-Value	Precision	Recall	F1-Score
20	10	6	75
13	50	100	67

The sentiments are of high importance when they are analysed at aspect level. On Figure 13 below, we present statistical sentiments at the aspect level of common aspects among query cases and base cases.

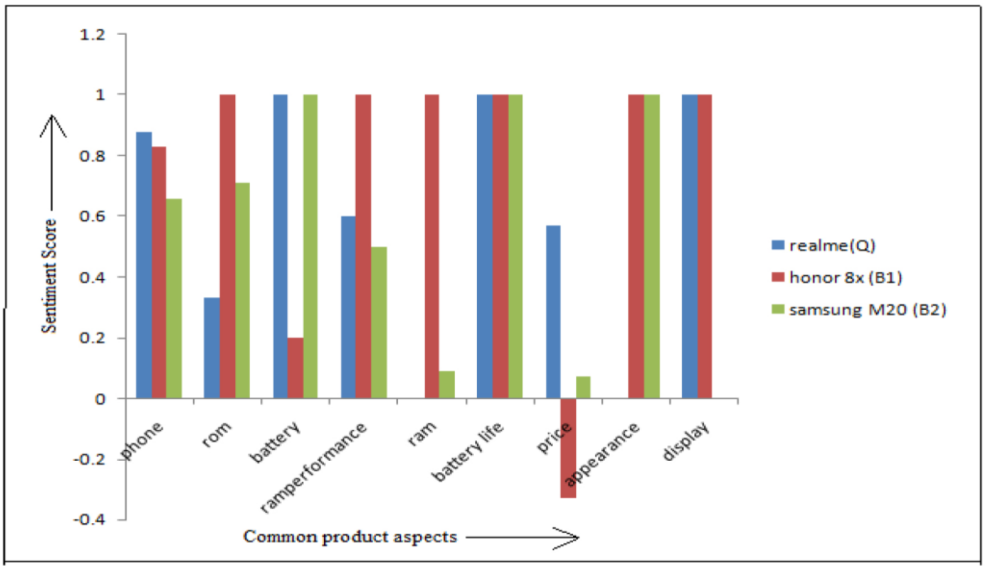


Fig 13. Statistical sentiments among query case and base cases

Now, the analytical learning is performed as the downstream machine learning task. Explanations for potential recommendations are generated by walking the path in the knowledge graph model rendered from the earlier extracted pieces of knowledge. It helps in finding paths with the required reasons to justify the recommendation by relating the higher-level concepts to the extracted knowledge pieces.

Products, their features, as well as statistical sentiments based on those features, have been annotated with ontology concepts that can be used to reason the recommendation using ontological knowledge models. Below is a figure of EPRO (Engineered Explainable Product Reviews Ontology) which is developed for the review's domain.

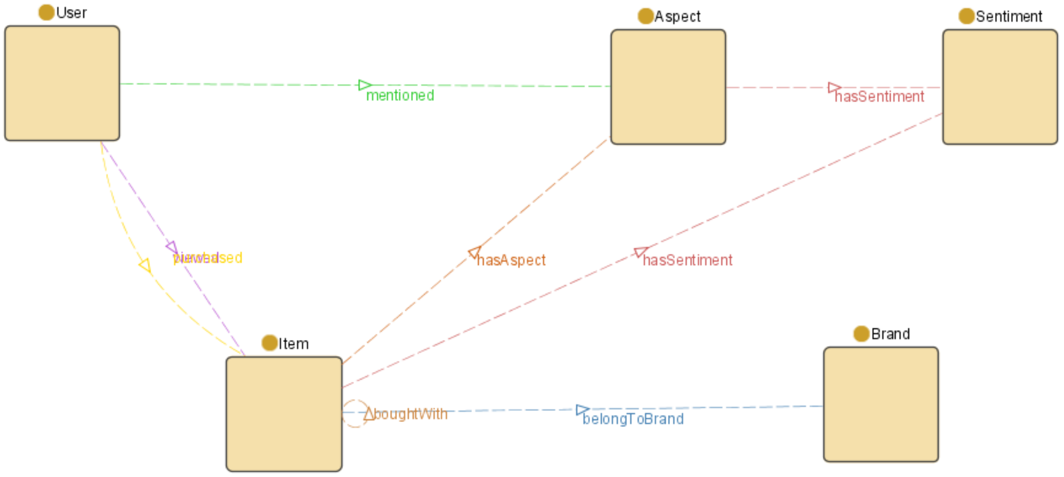


Fig 14. Visualization of Explainable Product Reviews Ontology

Recommendation is also a concept in this ontology. The above ontology visualization does not reflect this since its primary goal is to identify a name for the Recommendation entity instance when it is introspectively explained. As shown in figure 15, the EPRO ontology represents concepts.

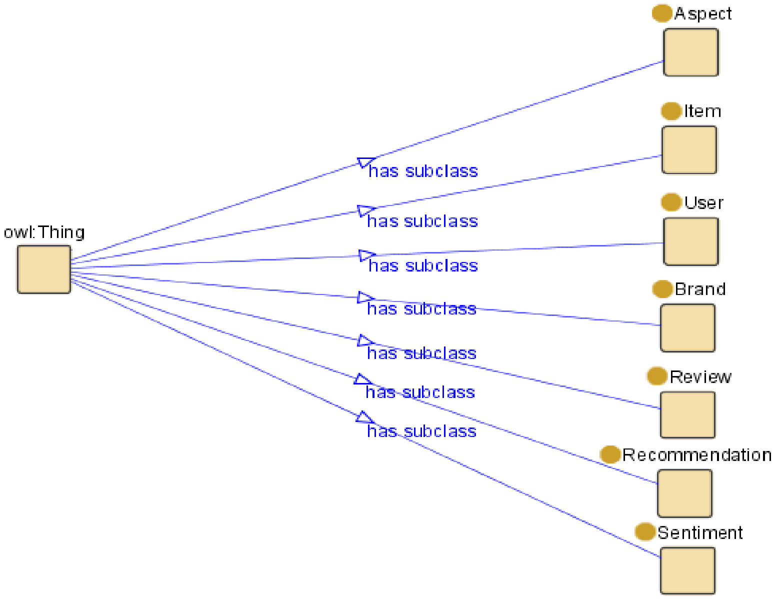


Fig 15. Concepts in EPRO Ontology

Ontologies based in Description Logic (DL) are strongly supported in engineering EPRO, but they do not fully cover the expressive possibilities needed to explain recommendations.

In order to achieve the goal, the research prospective utilizes Semantic Web Rule Language (SWRL), which is a language built to build rules on top of the engineered EPRO ontology consistent and logically to discover new relationships. These SWRL relationships allow for symbolic reasoning of the annotated data in the induced manner by the reasoner. The built SWRL rules will combine the advantages of both causal reasoning and neural logic reasoning. These induced new relationships will thus be useful in materializing the EPRO ontological knowledge graph and be associated as corresponding explanations to the statistical recommendations. This ontological analysis-based explanation serve as a separate model in the model-agnostic approach [20] and will be a new research direction.

V. CONCLUSION

The Exploratory Data Analysis of the e-commerce product reviews towards aspect-based sentiment analysis has been carried out successfully. Simply performing EDA can show us interesting patterns. Before making any assumptions, it is important to understand your data, which will help you avoid creating inaccurate models based on the data. Analysing the datasets properly and interpreting the results that are aligned with the business objectives is essential with the help of predictive analytics and machine learning. The statistical recommendations are determined by using sentiments and similarities with case-based reasoning.

The future prospective for providing the explanations to the statistical recommendations are by using ontological model in the model-agnostic approach. It will be deemed that when the SWRL rule consequent matches with the statistically recommended product then the SWRL rule antecedent is considered as the corresponding explanation to that product recommendation.

REFERENCES

1. Santosh, D. Teja, and B. Vishnu Vardhan. “Obtaining feature-and sentiment-based linked instance RDF data from unstructured reviews using ontology-based machine learning.” International Journal of Technology 2 (2015): 198-206.
2. Stavrianou, Anna, and Caroline Brun. “Expert Recommendations Based on Opinion Mining of User-Generated Product Reviews.” Computational Intelligence 31.1 (2015): 165-183.
3. Dong, Ruihai, et al. “Opinionated product recommendation.” International conference on case-based reasoning. Springer, Berlin, Heidelberg, (2013).
4. Krishnamurthi, Karthik, Vijayapal Reddy Panuganti, and Vishnu Vardhan Bulusu. "Understanding document semantics from summaries: a case study on Hindi texts." ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) 16.1 (2016): 1-20.
5. Ghosh, A., Nashaat, M., Miller, J., Quader, S., & Marston, C. (2018). A comprehensive review of tools for exploratory analysis of tabular industrial datasets. Visual Informatics, 2(4), 235-253.
6. Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. Psychological Methods, 2(2), 131.
7. Wangmo, C. (2017). An Exploratory Study On Bank Lending To SME Sector In Bhutan. International Journal of Scientific & Technology Research, 6(11), 47-51.
8. Ntow-Gyamfi, M., & Boateng, S. (2013). Credit risk and loan default among Ghanaian banks: An exploratory study. Management Science Letters, 3(3), 753-762.
9. Jency, X. F., Sumathi, V. P., & Sri, J. S. (2018). An exploratory data analysis for loan prediction based on nature of the clients. International Journal of Recent Technology and Engineering (IJRTE), 7(4).
10. Priya, K. U., Pushpa, S., Kalaivani, K., & Sartiha, A. (2018). Exploratory analysis on prediction of loan privilege for customers using random forest. Int. J. Eng. Technol, 7(2.21), 339.
11. Konopka, B. M., Lwow, F., Owczarz, M., & Łaczmański, Ł. (2018). Exploratory data analysis of a clinical study group: Development of a procedure for exploring multidimensional data. PloS one, 13(8), e0201950.
12. Nguyen, Heidi, et al. "Comparative study of sentiment analysis with product reviews using machine learning and lexicon-based approaches." SMU Data Science Review 1.4 (2018): 7.
13. Tan, Wanliang, Xinyu Wang, and Xinyu Xu. "SENTIMENT ANALYSIS FOR AMAZON REVIEWS."
14. D. Teja Santosh, V. Kakulapati & K. Basavaraju (2020) Ontology-based sentimental knowledge in predicting the product recommendations: A data science approach, Journal of Discrete Mathematical Sciences and Cryptography, 23:1, 1-18, DOI:10.1080/09720529.2020.1721676
15. Ai, Q., V. Azizi, X. Chen, and Y. Zhang (2018). “Learning heterogeneous knowledge base embeddings for explainable recommendation”. Algorithms. 11(9): 137.
16. Xie, L., Hu, Z., Cai, X. et al. Explainable recommendation based on knowledge graph and multi-objective optimization. Complex Intell. Syst. 7, 1241–1252 (2021). https://doi.org/10.1007/s40747-021-00315-y
17. Peake, G. and J. Wang (2018). “Explanation mining: Post hoc interpretability of latent factor models for recommendation systems”. In: Proceedings of Beyond Personalization 2005: A Workshop on the Next Stage of

Recommender Systems Research at the 2005 International Conference on Intelligent User Interfaces, San Diego, CA,USA. ACM. 2060–2069.

18. Santosh, D. Teja, and B. Vishnu Vardhan. "Automatic Machine Recognition Of Features And Sentiments From Online Reviews." IAENG TRANSACTIONS ON ENGINEERING SCIENCES: Special Issue for the International Association of Engineers Conferences 2015. 2017.
19. Kumar, KC Ravi, D. Teja Santosh, and B. Vishnu Vardhan. "Determining the semantic orientation of opinion words using typed de-pendencies for opinion word senses and SentiWordNet scores from online product." (2017). doi: 10.1504/IJKWI.2017.10010171
20. Molnar, C. (2019). Interpretable Machine Learning. Leanpub.