Predictive & Topic Modeling for Virtual Voice Assistants

Muhammad Haris Saleem†
Information Technology & Web Science
Rensselaer Polytechnic Institute
Troy NY USA
saleem@rpi.edu

Lydia Manikonda
Lally School of Management
Rensselaer Polytechnic Institute
Troy NY USA
manikl@rpi.edu

EXECUTIVE SUMMARY

With the continuous increase in development in AI, it is essential to look at how they can be improved. One of the most prominent AI technologies currently are virtual voice assistants namely Amazon Echo, Apple Homepod and Google Home. For any product improvement customer feedback is important. Since most of the products are bought online, we can analyze the reviews posted by the customers to get insights not only related to product performance but main talking points about each product.

We gathered reviews for virtual voice assistants mentioned above from BestBuy. The analysis is done is two major aspects. First, rating prediction using NLP and Machine Learning techniques. Secondly, Topic Modeling using NLP visualizations. The performance of various NLP techniques such as different BERT Embeddings and TF-IDF embeddings are compared to get an accurate model. Important features we used include embeddings from NLP pre trained models, Luke features, POS counts, and the duration for which the product was owned for ranging from less than a week to more than 2 years.

The data asked for a classification approach, however some regression approached were also used, keeping in mind the general applications of our findings. The results we gathered from our algorithms have a good performance giving approximately 80% accuracy, however much difference was not seen amongst performance of different NLP pre-trained models. In terms of Topic Modeling, useful insights were found such as what do people talk about matters a lot on for how long they have owned the product.

KEYWORDS

Natural Language Processing, Voice Assistants, Rating Prediction, Topic Modeling.

2. BENCHMARKING

Martin [1], discusses two different classification approaches for an imbalanced dataset which is Amazon Reviews for a video games dataset. First, binomial approach where Ratings are either considered good (3-5) or bad (1-2). Second is multi-class classification which has 5 classes (1-5). Models used are Support Vector Machines, Naïve Bayes, and Random Forrest. **Accuracy of 90% is achieved for binomial approach while 83% for multi-class classification.**

Another approach used by various people to do Rating prediction is by converting this problem from classification to prediction. Michael [2] used Linear Regression to find mean squared error for predicting user rating for Amazon Gourmet Food reviews mean squared error of 60 was achieved. **Latent Dirichlet Allocation** was used but not great results were achieved. Ridge Regression gave **mean squared error of 0.93.**

For Topic Modeling, Zhao, Lei and Qian [3] used sentiment similarity, sentimental influence and items reputation similarity as a feature to predict the ratings.

3. DATA DESCRIPTION AND PROCESSING

The data is gathered from BestBuy for the products, Amazon Alexa, Apple Homepod and Google Home. Main features of the data are, Posted Date, Review Content, Owned Time and Rating of the product.

Now let's dig a little deep into data description. Fig 1 shows the total counts for each product category. It can be seen Amazon has the highest number of reviews. Later we will also calculate time taken for some algorithms to run, hence for explanation of results this is important.

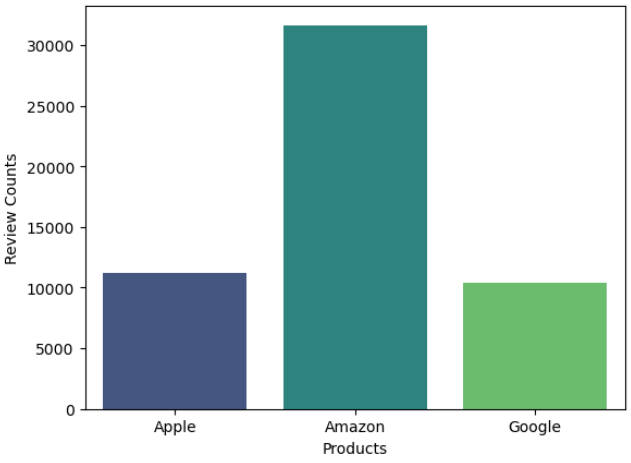

Fig1: Product Review Counts

If we analyze the average length of review for each category, we can see in Fig 2 that Apple product seems to have the highest average review length, followed by Google and then Amazon.

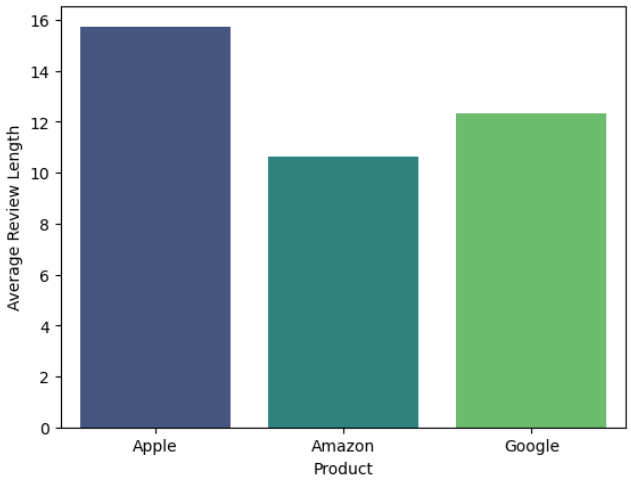

Fig 2: Average Review Length

Now we will discuss important processing which was required for our modeling. An important feature in specially topic modeling is going to be the duration for which product was owned. Fig 3 shows the counts of reviews against each category of duration. Fig 3 is for Apple product; most reviews are from customers who have owned the product for 1 month. We have so many small subcategories.

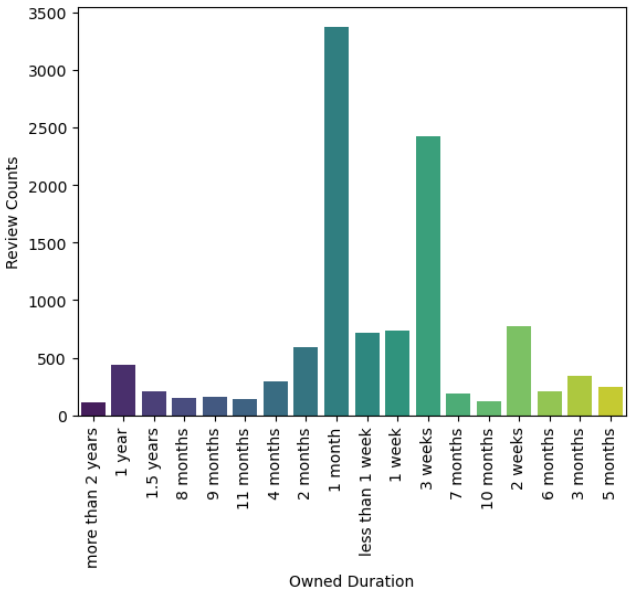

Fig 3: Original Duration Categories

For an effective analysis, we decided to categorize the duration into various bins. Fig 4,5,6 show review counts for Apple, Amazon and Google respectively. Similar trend is followed, most of the reviews are from people who have owned the product for duration between 1 and 6 Months.

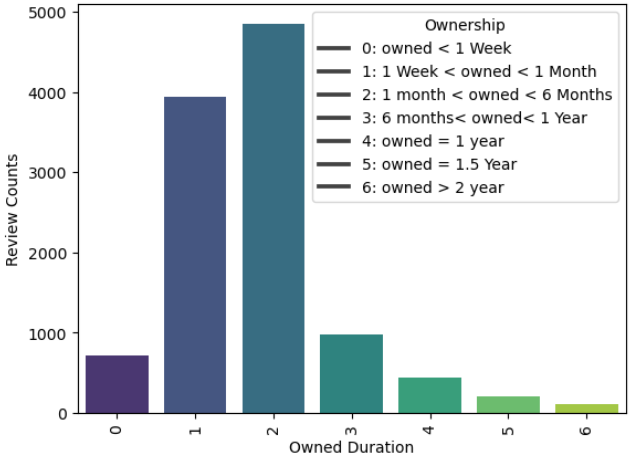

Fig 4: Apple Review Counts with Duration Owned

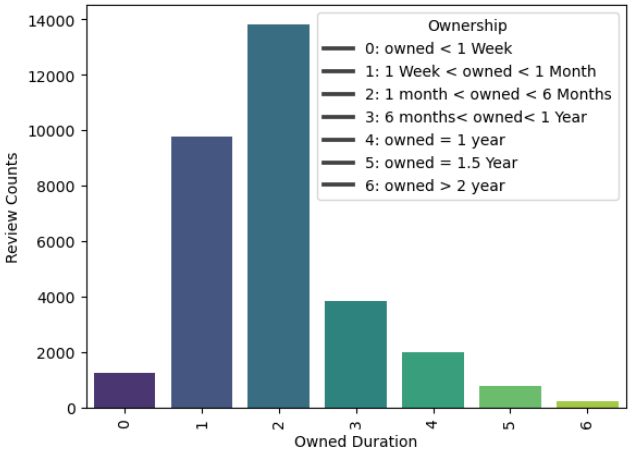

Fig 5: Amazon Review Counts with Duration Owned

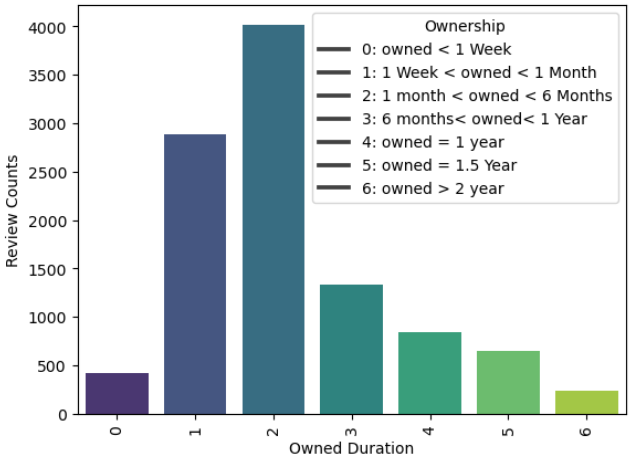

Fig 6: Google Reviews with Duration Owned

Now we will analyze the content of the reviews. Before analyzing what is said in any review, we need to remove the stop words. We are using NLTK library stop words list as a base list. In this list we are going to add product names, as in reviews product name is mentioned again and again. There could be some plurals of product names as well, so using textblob from NLTK we removed any plurals from the content as well.

Followed by removal of stop words, we removed all the non-alphanumeric characters as they generally make the data messy and generally do not signify anything in terms of Machine Learning modeling.

After the removal of stop words and unnecessary words, we can visualize the most used words using the WordCloud function of python. Fig 7,8, 9 represents the word clouds of Apple, Amazon and Google respectively.

Fig 7: Apple WordCloud

Fig 8: Amazon WordCloud

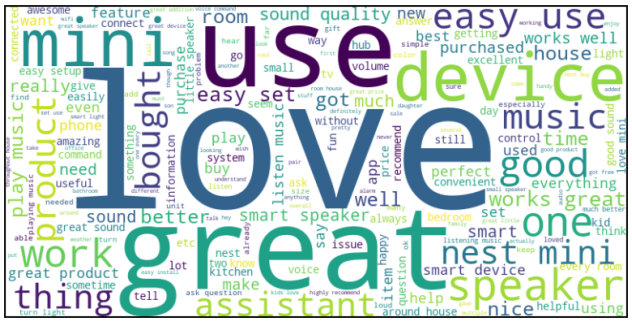

Fig 9: Google WordCloud

4. MODELING

Now or data is cleaned, we will move towards some feature engineering which is a crucial part of the modeling. Initially, using the cleaned reviews from preprocessing. We found the counts of Parts of Speech namely, Noun, Verb and Adjectives used in each review using the spaCy library.

Since, we have already looked at most frequent used words from word clouds in the previous section, we will now look at combination of words called n_grams. In our modeling, we will use Top 100 bigrams. Table 1,2,3 show top 10 bigrams and trigrams for Apple, Amazon and Google respectively.

Bi-Gram	Tri_Gram
'easy set' 'easy use' 'good sound' 'great sound' 'small speaker' 'smart speaker' 'sound great' 'sound quality' 'sounds great' 'works great	'good sound quality' 'great little speaker' 'great sound easy' 'great sound quality' 'great sound small' 'love sound quality' 'sound quality amazing' 'sound quality good' 'sound quality great' 'sound small speaker

Table 1: N_Grams Apple

Bi-Gram	Tri_Gram
'easy set', 'easy use', 'every room', 'great product', 'great sound', 'listen music', 'play music', 'sound quality', 'works great', 'works well'	'easy set use', 'easy use set', 'every room house', 'free tv purchase', 'good sound quality', 'great sound quality', 'love easy use', 'one every room', 'product easy use', 'sound quality great'

Table 2: N_Grams Amazon

Bi-Gram	Tri_Gram
'easy set', 'easy use', 'great product', 'listen music', 'nest mini', 'play music', 'smart speaker', 'sound quality', 'works great', 'works well'	'easy set easy', 'easy set use', 'easy setup use', 'every room house', 'good sound quality', 'great sound quality', 'nest mini 2nd', 'one every room', 'set easy use', 'sound quality good'

Table 3: N_Grams Google

We will be using only the bigrams in our Machine Learning algorithms, as top tri grams are mostly using words of bi-grams. To avoid the curse of dimensionality we will ignore tri-grams.

4.1 Rating Prediction:

Table 4 below shows the features we are using the algorithms.

Features	Details
Embeddings	BERT or TF-IDF
POS Counts	VERB, NOUN, ADJ
bi-grams	Top 100
Review Length	-
Liwc	Luke features
Ownership Time	6 categories
Ratings	1-5

Table 4: Features for Modeling

Most important feature or list of features in our technique are BERT embeddings. BERT uses transform based approach to convert text to embeddings. These embeddings, however, cannot be interpreted. We used 3 different BERT pre-trained models from the transformers packagen amed as follows:

1. distilbert-base-nli-mean-tokens
2. bert-large-nli-mean-tokens
3. bert-large-nli-cls-token

The embeddings from each pre trained model depend upon the transformers used to train each model. We will compare the results of each model with base case TF-IDF vectorization which created vectors based on their importance in the text.

The algorithm implemented is Random Forrest Classifier. Train Test split of 70-30 and Random State = 12 was used. Due to huge dimensionality cross validation couldn't be performed. The results for Apple, Amazon and Google are mentioned in Table 5, 6 and 7 respectively.

	TF-IDF	BERT BASE MEAN	BERT LARGE MEAN	BERT LARGE CLS
Accuracy (%)	80.85	81.03	80.97	80.94
F1	20.91	26.31	27.33	26.36
Precision	47.51	46.75	49.37	45.79
Recall	21.56	25.36	26.22	25.69
Time (s)	21.98	43.80	49.89	49.38

Table 5: RF Apple

	TF-IDF	BERT BASE MEAN	BERT LARGE MEAN	BERT LARGE CLS
Accuracy (%)	83.13	83.12	83.17	83.13
F1	19.35	19.57	19.73	20.36
Precision	60.75	55.93	61.17	66.64
Recall	20.59	20.68	20.77	21.09
Time (s)	54.39	171.20	188.11	190.92

Table 6: RF Amazon

	TF-IDF	BERT BASE MEAN	BERT LARGE MEAN	BERT LARGE CLS
Accuracy (%)	78.67	79.41	79.38	79.38
F1	18.02	20.97	22.74	22.97
Precision	21.34	49.08	71.34	66.27
Recall	20.17	21.7	22.61	22.75
Time (s)	18.90	42.63	46.11	45.59

Table 7: RF Google

The results are identical across all models. BERT Large and BERT Large mean generally have high precision scores, this indicates a class imbalance and a bias. Almost 70% of reviews for each product are of rating 5. Making hard for the models to capture the data effectively. Literature suggests [2] oversampling can improve the accuracy results and precision scores. F1 score was improved from 40% to 90% by oversampling the data.

Even though this seems to be a classification problem, we can implement a regression algorithm as well, as the idea is to identify bad performing products, which if regression is used can be separated using a threshold.

The results of Random Forest Regression were calculated using Principal Component Analysis as the dimensionality was very high. Approximately 1000+ features for each product.

The results are mentioned in Table 8 and 9 showing with just 10 components and components covering at most 85% of variance.

MSE	BERT BASE MEAN	BERT LARGE MEAN	BERT LARGE CLS
Apple	0.491	0.457	0.453
Amazon	0.326	0.326	0.328
Google	0.403	0.403	0.404

Table 8: RF Regression

MSE	BERT BASE MEAN	BERT LARGE MEAN	BERT LARGE CLS
Apple	0.483	0.458	0.466
Amazon	0.322	0.319	0.324
Google	0.413	0.409	0.405

Table 9: RF Regression 85%Variance

The results from both regression and classification are good. However, from classification results an interesting result is BERT performs like TF-IDF and is computationally expensive.

From Regression results a low Mean Squared Error is achieved which can be explored in detail.

4.2 Topic Modeling :

Topic modeling is important application for BERT model. We can analyse the main topics people talk about.

Fig. 10, 11 and 12 shows topics in BERT for Apple, Amazon and Google respectively. Topics are divided amongst the period the product was owned.

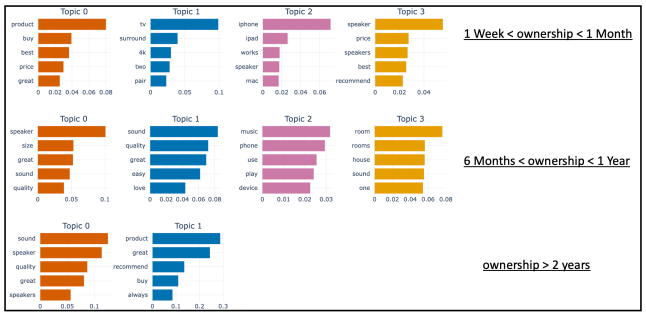

Fig 10 : Apple Topics

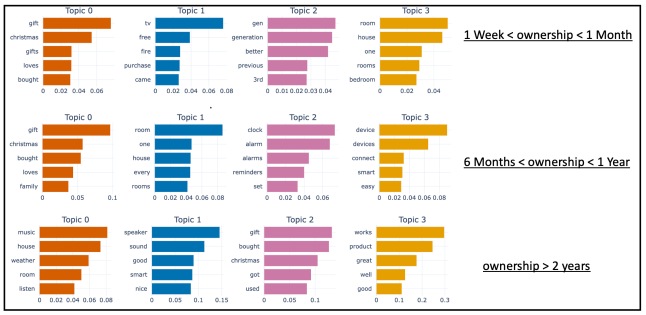

Fig 11 : Amazon Topics

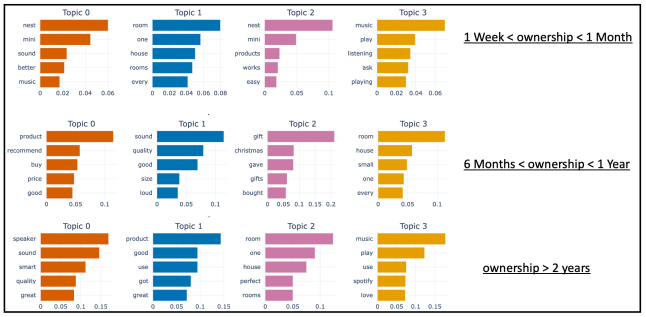

Fig 12: Google Topics

The results from topic modeling show that people who have owned the product for less amount of time talk more about the price and purchase experience as can be observed above, top words are related to **gift, purchase and free**. While people who have owned the product for longer talk about specific features of the product such as **speaker, sound and quality**. After improving the topic modeling further we can put these parameters and finding as input to out predictive model for better understanding of reviews.

5. CONCLUSION

We used several algorithms to predict ratings, error rate and accuracy was good. **Accuracy of 80% and MSE of 0.5** all the models performed around the same metrics value. Which highlights the lack of algorithms to capture the complexity of embeddings. Additionally, the class imbalance discussed is also a problem which needs to be solved to understand the performance of BERT modeling. Future work can include use of additional Classification and Regression algorithms. Another important aspect of Voice Assistants is Topic Modeling in the aspect of **Ethics and Privacy.** Which can also identify interesting patterns and can be used as an important feature in predictive modeling.

ACKNOWLEDGMENTS

The project would not have been possible without Dr. Lydia Manikoda's continuous guidance and support throughout the project.

REFERENCES

[1] Martin, M. (2016) Predicting ratings of Amazon Reviews - techniques for imbalanced datasets. Available at: https://matheo.uliege.be/bitstream/2268.2/2707/4/Memoire_MarieMartin_s112740.pdf

[2] Tran, Michael. Predicting and Recommendation Using Basic Linear Regression Models, cseweb.ucsd.edu/classes/sp15/cse190-c/reports/sp15/029.pdf.

[3] X. Lei, X. Qian and G. Zhao, "Rating Prediction Based on Social Sentiment From Textual Reviews," in IEEE Transactions on Multimedia, vol. 18, no. 9, pp. 1910-1921, Sept. 2016, doi: 10.1109/TMM.2016.2575738.