William Hahn

Professor Tucker

DATA 340

17 December 2023

iPhone XR Sentiment Analysis

Abstract

My main research question for this project is: does vectorizing our dataset improve the performance of a naïve bayes classifier model? The dataset that I chose contains all iPhone XR reviews from Amazon, and I analyzed this dataset by using a naïve bayes classifier to predict review ratings based on the text of each review. I also used vectorization to improve the model's accuracy. In particular, the vectorization techniques I used include term frequency-inverse document frequency (TF-IDF) and bag-of-words. The accuracy of the naïve bayes model without vectorization is around 73.65 percent. This means that the accuracy of this model is very useful in terms of real-world application because most models tend to be only around 50 percent accurate. The accuracy of the model with TF-IDF is around 73.45 percent. This means that the model's performance didn't improve at all, and that the performance of the model decreased after using TF-IDF. The accuracy of the model with bag-of-words is around 76%. This means that the model's accuracy increased and that using bag-of-words was very useful in improving our model. In conclusion, using bag-of-words proved to be more useful than using TF-IDF due to the differences in how each vectorization method works. However, using bag-of-words only improved the performance of the model by around 2.35 percent, which is still statistically significant in terms of real-world application. This also means that using more vectorization methods is advisable such as using transformations on the dataset. Finally, the goal of this project

is to teach people why using naïve bayes classifiers and vectorization is very important in the real world.

Introduction

The dataset that I chose for this project contains a list of more than 30,000 Amazon reviews on the iPhone XR from Kaggle.com. The main motivation behind my research question is that I wanted to learn the basics on how to use popular NLP techniques such as naive Bayes classifiers on certain datasets. In addition, I chose this particular dataset because I wanted to learn how businesses create predictive models and how they interpret them to further improve their products. Finally, the importance of this project is to not only practice popular NLP skills, but to also give people (especially those who don't have a data science background) an idea of how businesses use predictive models for certain products. It will also give people an idea on how businesses improve their predictive models.

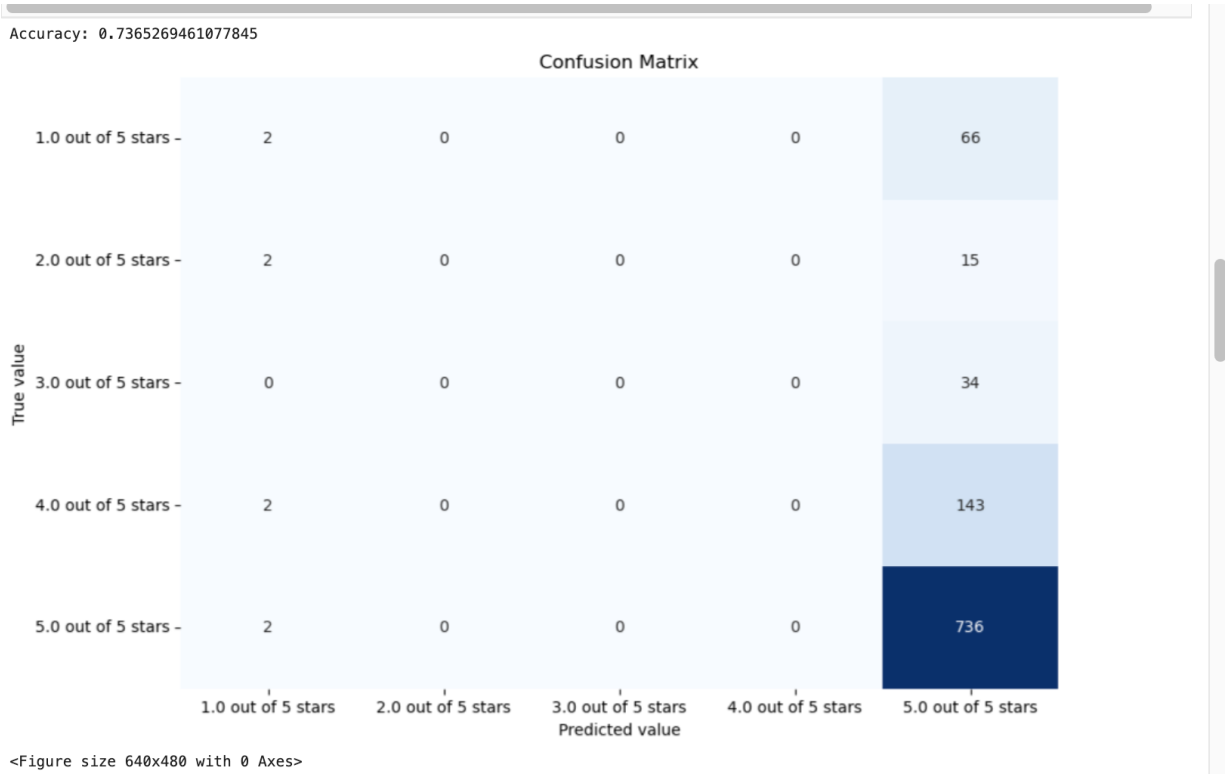
Methodology/Dataset

In order to analyze the chosen dataset, I will be using a naive Bayes classifiers model to predict review ratings on a scale from 1 to 5 based on the text of each review. A naive Bayes classifier is a statistical classification technique that assumes that each predictor is independent from each other. I will also be using two vectorization techniques on the dataset to further improve this model and they are: TF-IDF and bag-of-words. Vectorization is a technique that converts textual data into numerical vectors, which can be used as input for machine learning models such as naive Bayes classifiers.

TF-IDF is a technique that involves the importance of a word in a document relative to a collection of documents. TF-IDF takes into account the frequency of each word in a document and the uniqueness of each word across the collection of documents. Bag-of-words is a technique

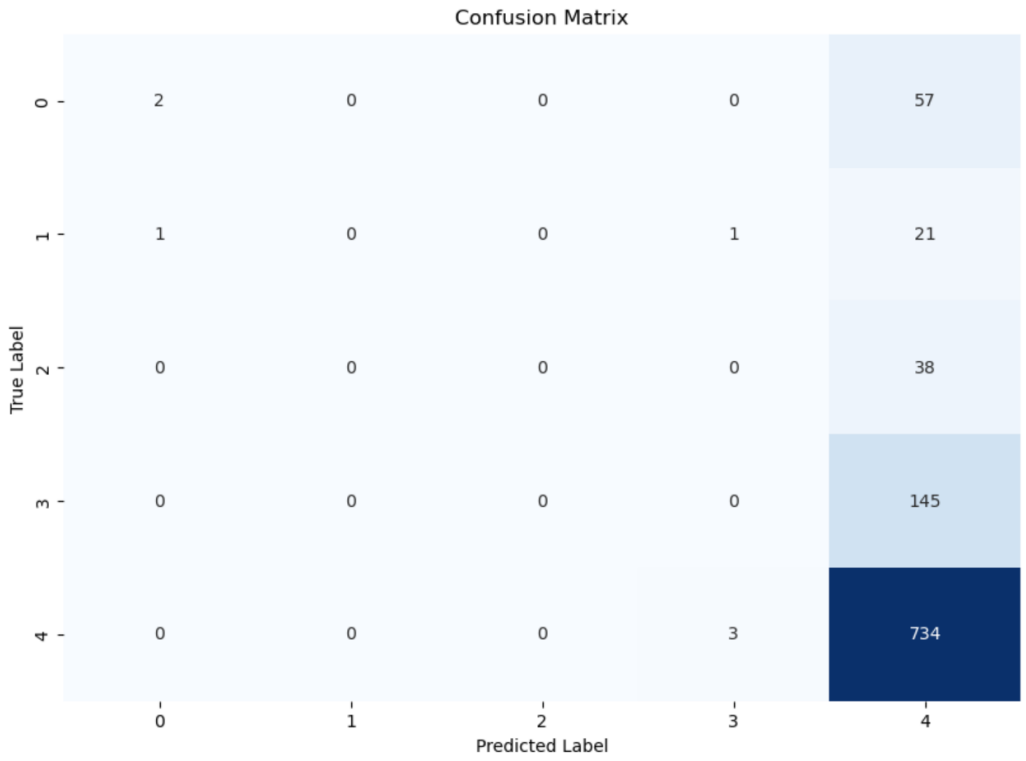
that involves a set of unordered terms that takes into account the frequency of each term in the document. The dataset contains 30,000 Amazon reviews on the iPhone XR, and this data was collected from a website called Kaggle.com. I chose the two vectorization methods and the naive Bayes classifier model because I wanted to learn how to apply those methods and models.

Results



The naive Bayes classifier model showed an accuracy score of around 73.65 percent. This percentage is very good in terms of real-world application because most models in the business world are barely 50 percent accurate and that models over 50 percent are generally rare and sometimes skewed due to bias. However, improving the model's performance can be difficult, and we used two vectorization techniques called: bag-of-words and TF-IDF.

```
Accuracy: 0.7345309381237525
```

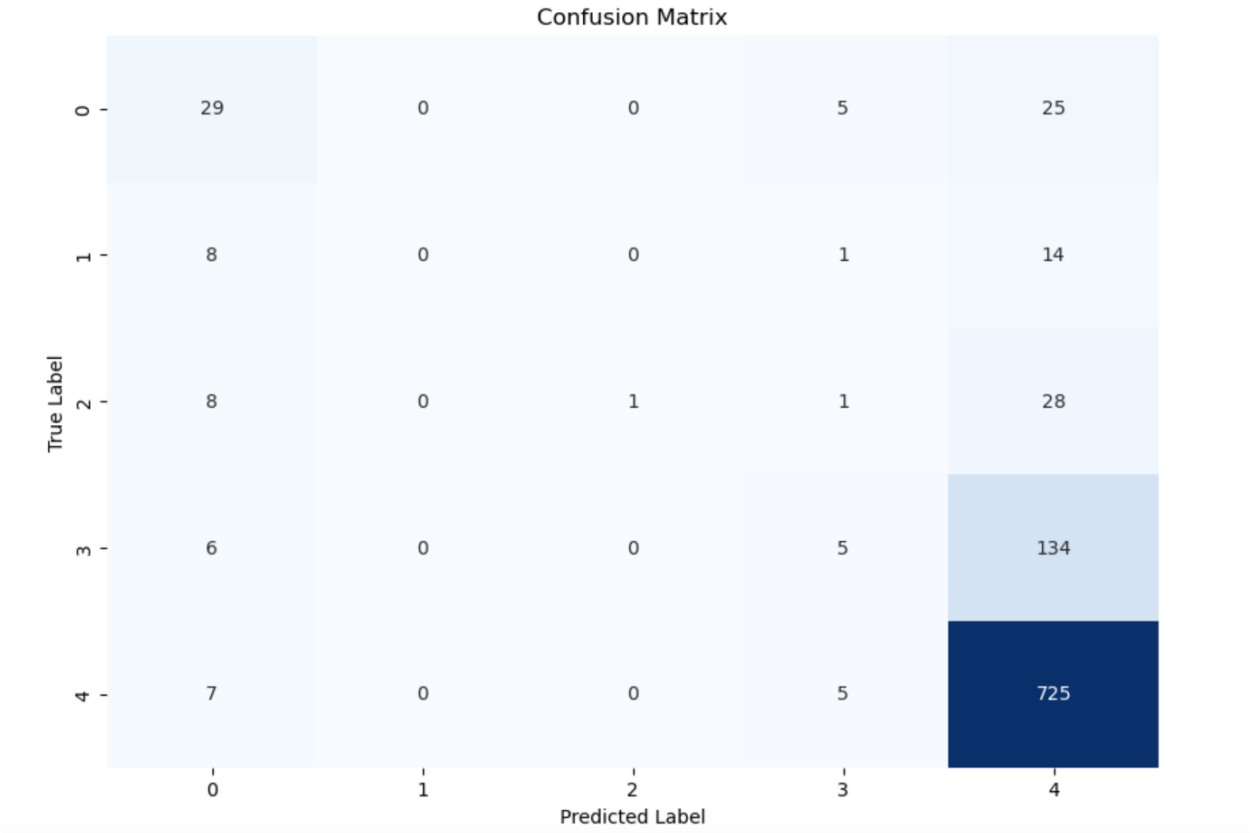



```
<Figure size 640x480 with 0 Axes>
```

After applying the TF-IDF technique, the model's performance is 73.45 percent accurate. This means that the performance has decreased. The model's performance most likely decreased because there is no uniqueness in the data, which means that the frequency of the words in terms of uniqueness was not significant for the model's overall performance.

Accuracy: 0.76

	precision	recall	f1-score	support
1.0 out of 5 stars	0.50	0.49	0.50	59
2.0 out of 5 stars	1.00	0.00	0.00	23
3.0 out of 5 stars	1.00	0.03	0.05	38
4.0 out of 5 stars	0.29	0.03	0.06	145
5.0 out of 5 stars	0.78	0.98	0.87	737
accuracy			0.76	1002
macro avg	0.72	0.31	0.30	1002
weighted avg	0.71	0.76	0.68	1002



The results of the bag-of-words vectorization shows that the improved model's performance is around 76 percent accurate. This means that applying the bag-of-words vectorization significantly improved the naive Bayes classifer model by 2.35 percent. This also means that the model's performance most likely increased because the frequency of each word in the text seems to statistically significant in determining the model's overall performance. The 2.35 percent increase in performance may not seem like a big increase, but in terms of real-world application

this is statistically significant because a two percent increase can mean that it is two percent more reliable in predicting a review rating.

Discussion

The results of study prove that vectorization techniques such as bag-of-words and TF-IDF can either improve or not improve a naive bayes classifier model or any model in general. This is the case because the vectorization technique bag-of-words improved the accuracy of the naïve bayes classifier model by 2.35 percent. However, the vectorization technique TF-IDF didn't improve the model. In fact, the model performed slightly worse in terms of accuracy after we applied the TF-IDF. This also suggests that not all vectorization techniques on datasets will improve the performance of a naïve bayes classifier model.

In this case, the model only improved for the bag-of-words vectorization because this type of vectorization only takes into account the frequency of each term in the text. On the other hand, the model didn't improve for the vectorization technique TF-IDF because it took into account both the frequency and uniqueness of each term in the text. However, there could also be many more explanations on why such techniques such as TF-IDF make the original model perform much worse in terms of accuracy.

One explanation could be that the dataset contains too much noise, which means that there is too much irrelevant information in the text data. Another explanation could be that the naive Bayes classifier model is not complex enough for TF-IDF to improve the model's accuracy by that the documents in the dataset might be too short. In short, all of those explanations should be considered when using any vectorization technique and that one shouldn't assume that one technique is better than the other in terms of model performance.

Conclusion

In conclusion, the vectorization techniques may or may not improve a naive Bayes classifier model's performance in terms of accuracy. This is true because not all vectorization techniques are appropriate for a given dataset. For example, the vectorization technique bag-of-words made the perform much better than using the TF-IDF vectorization technique. It is also good practice to try not only all of the vectorization techniques, but to also explore new ways on how one can improve a naive Bayes classifier without vectorization.

References

Devastator, The. "IPhone Reviews from Amazon.Com." *Kaggle*, 18 Jan. 2023, www.kaggle.com/datasets/thedevastator/apple-iphone-11-reviews-from-amazon-com.

Appendices

Dataset (head and tail):

	index	product	helpful_count	total_comments	url	review_country	reviewed_at	review_text	review_rating	product_company	profile_name	review_title
0	0	Apple iPhone XR (64GB) - Black	5,087 people found this helpful	24	https://www.amazon.in/Apple-iPhone-XR-64GB-Bla...	India	2018-12-12	NOTE:	3.0 out of 5 stars	Apple	Sameer Patil	Which iPhone you should Purchase ? iPhone 8, X...
1	1	Apple iPhone XR (64GB) - Black	2,822 people found this helpful	6	https://www.amazon.in/Apple-iPhone-XR-64GB-Bla...	India	2018-11-17	Very bad experience with this iPhone xr phone....	1.0 out of 5 stars	Apple	Amazon Customer	Don't buy iPhone xr from Amazon.
2	2	Apple iPhone XR (64GB) - Black	1,798 people found this helpful	0	https://www.amazon.in/Apple-iPhone-XR-64GB-Bla...	India	2019-01-27	Amazing phone with amazing camera coming from ...	5.0 out of 5 stars	Apple	A	Happy with the purchase
3	3	Apple iPhone XR (64GB) - Black	1,366 people found this helpful	14	https://www.amazon.in/Apple-iPhone-XR-64GB-Bla...	India	2019-05-02	So I got the iPhone XR just today. The product...	1.0 out of 5 stars	Apple	Shubham Dutta	Amazon is not an apple authorised reseller. Pl...
4	4	Apple iPhone XR (64GB) - Black	536 people found this helpful	5	https://www.amazon.in/Apple-iPhone-XR-64GB-Bla...	India	2019-05-24	I've been an android user all my life until I ...	5.0 out of 5 stars	Apple	Nepuni Lokho	Excellent Battery life and buttery smooth UI
...	...	...	...	...	...	...	...	...	...	...	...	...
5005	5005	Apple iPhone XR (64GB) - Black	0	0	https://www.amazon.in/Apple-iPhone-XR-64GB-Bla...	India	2019-11-13	Dhamaka	4.0 out of 5 stars	Apple	Shreya	Dhamaka phone
5006	5006	Apple iPhone XR (64GB) - Black	0	0	https://www.amazon.in/Apple-iPhone-XR-64GB-Bla...	India	2019-11-15	Goodbye	4.0 out of 5 stars	Apple	murali hv	Good
5007	5007	Apple iPhone XR (64GB) - Black	0	0	https://www.amazon.in/Apple-iPhone-XR-64GB-Bla...	India	2019-12-29	Nothing	5.0 out of 5 stars	Apple	Manish	Fantabulous phone. Easy to use.
5008	5008	Apple iPhone XR (64GB) - Black	0	0	https://www.amazon.in/Apple-iPhone-XR-64GB-Bla...	India	2019-11-10	Superbb	5.0 out of 5 stars	Apple	basil john p	Fantastic
5009	5009	Apple iPhone XR (64GB) - Black	0	0	https://www.amazon.in/Apple-iPhone-XR-64GB-Bla...	India	2019-11-05	Nothing	5.0 out of 5 stars	Apple	Amazon Customer	Best purchase

The only columns used for this analysis are: "review _text" and :review_rating".