Machine Learning Foundation- Batch 03

CLOTHES REVIEWS ANALYSIS WITH NLP

A Dissertation By

Yohan Hemal De Silva

DSA_0303

Instructed By

Dr. Sumudu Tennakoon

Contents

INTRODUCTION

Natural Language Processing (NLP) is a field of Artificial Intelligence whose purpose is finding computational methods to interpret human language as it is spoken or written. The idea of NLP goes beyond a mere classification task that could be carried on by ML algorithms or Deep Learning NNs. Indeed, NLP is about interpretation: we want to train our model not only to detect frequent words, but also to count them or to eliminate some noisy punctuations; we want it to tell whether the mood of the conversation is positive or negative, whether the content of an e-mail is mere publicity or something important.

In here I'm going to analyze the Women's Clothing E-Commerce dataset which contains text reviews written by customers. The idea is predicting the sentiment of each review and see whether it is consistent with the reviewed text by the customer.

DATA

This is a Women's Clothing E-Commerce dataset revolving around the reviews written by customers. Its nine supportive features offer a great environment to parse out the text through its multiple dimensions.

Content

This dataset includes 23486 rows and 10 feature variables. Each row corresponds to a customer review, and includes the variables:

- Clothing ID: Integer Categorical variable that refers to the specific piece being reviewed.
- Age: Positive Integer variable of the reviewers age.
- Title: String variable for the title of the review.
- Review Text: String variable for the review body.
- Rating: Positive Ordinal Integer variable for the product score granted by the customer from 1 Worst, to 5 Best.

- Recommended IND: Binary variable stating where the customer recommends the product where 1 is recommended, 0 is not recommended.
- Positive Feedback Count: Positive Integer documenting the number of other customers who found this review positive.
- Division Name: Categorical name of the product high level division.
- Department Name: Categorical name of the product department name.
- Class Name: Categorical name of the product class name.

Data set reference - kaggle

METHODOLOGY

Tools Used

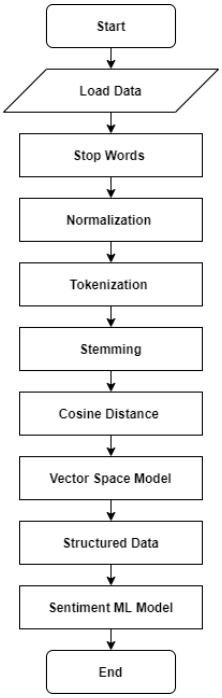
Note – Used SQL Server 2019 to generate Sentiment Column (Y variable) for testing data set using (Rating, Recommended IND, Positive Feedback Count) existing columns from data set.

Solution Approached

For the validate and verify the accuracy level, there are four machine learning models were deployed and tested for sentiment analysis with final data vocabulary.

- **Support vector machine**
- **KNN Algorithm**
- **Random Forest Algorithm**
- **Multinomial Naive Bayes Algorithm**

High – Level Process Flow



RESULTS

- **Apply Random Forest Algorithm After TF/IDF**

	Precision	**Recall**	**F1-Score**	**Support**
Negative	0.57	0.06	0.10	413
Natural	0.65	0.16	0.25	615
Positive	0.80	1.00	0.89	3501
Avg / Total	0.76	0.80	0.73	4529
Accuracy – 0.7970854493265621				

- **Apply Support Vector Machine After TF/IDF**

	Precision	Recall	F1-Score	Support
Negative	0.31	0.15	0.20	413
Natural	0.51	0.38	0.44	624
Positive	0.87	0.97	0.92	3492
Avg / Total	0.77	0.81	0.79	4529
Accuracy – 0.8114374034003091				

- **Apply KNN Algorithm After TF/IDF**

	Precision	Recall	F1-Score	Support
Negative	0.30	0.19	0.23	413
Natural	0.43	0.21	0.28	624
Positive	0.84	0.95	0.89	3492
Avg / Total	0.73	0.78	0.75	4529
Accuracy – 0.7803047030249504				

- **Apply Multinomial Naive Bayes Algorithm After TF/IDF**

	Precision	Recall	F1-Score	Support
Negative	0.23	0.01	0.01	413
Natural	0.56	0.13	0.21	624
Positive	0.79	1.00	0.88	3492
Avg / Total	0.71	0.79	0.71	4529
Accuracy – 0.7853830867741223				

CONCLUSION

Since SVM (Support Vector Machine) has the highest accuracy level it is the best model for sentiment analysis among these deployed models. And also all of above models reach 0.78 of accuracy level with processed reviewed data set.

DISCUSSION

Sentiment analysis plays a pivotal role in market research and similar fields. Indeed, keeping track of customers' satisfaction and, most importantly, isolate those drivers which determine their sentiment towards items, could lead to winning marketing campaigns and selling strategies.