Fake Amazon reviews detection

Gianmarco Pastore

Introduction

Addressing the problem

Customers increasingly rely on reviews for product information. However, the usefulness of online reviews is impeded by fake reviews that give an untruthful picture of product quality. The fake reviews usually are generated by using text-generation algorithms to automate the fake review creation, because the effort required is a fraction wrt to human generated reviews. Therefore, detection of fake reviews is needed.

Goal of the work

- Model a classifier capable of detecting fake Amazon Reviews
- Reach good performances with the obtained classifier

Introduction

DATASET

In order to generate the fake reviews, GPT-2 has been applied to the publicly available Amazon Review Data (2018) dataset, which is extensive and reputable. Only the Top-10 Amazon categories with the most product reviews have been considered, accounting for 88.4% of the reviews in the baseline dataset. For each product category, ~2000 fake reviews have been generated, for a total of ~ 40K reviews equally divided in fake and real.

Workflow

Data exploration

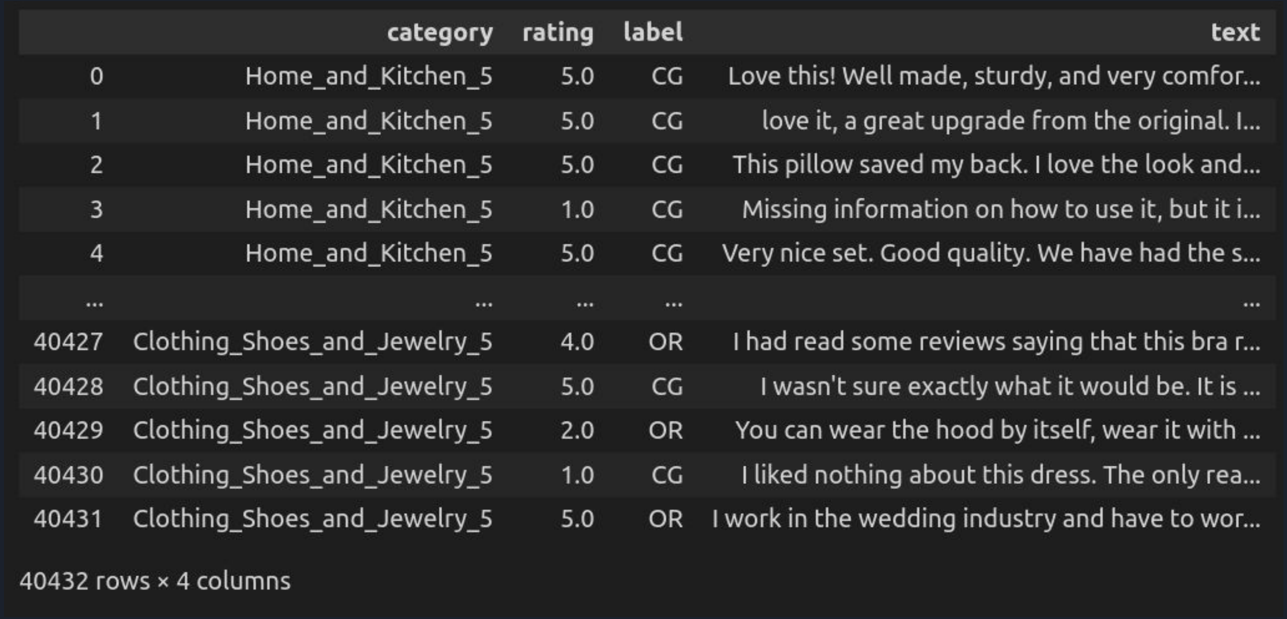
Data preprocessing

Data processing

Performance evaluation

Data exploration

The dataset contain 4 features: category, rating, label and text

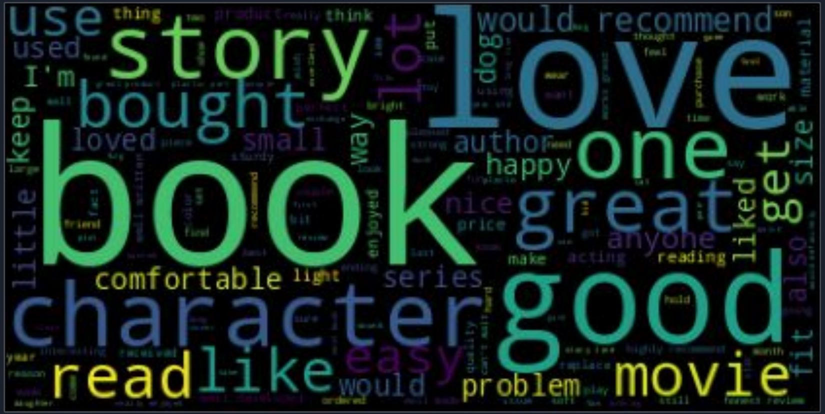


	category	rating	label	text
0	Home_and_Kitchen_5	5.0	CG	Love this! Well made, sturdy, and very comfor...
1	Home_and_Kitchen_5	5.0	CG	love it, a great upgrade from the original. I...
2	Home_and_Kitchen_5	5.0	CG	This pillow saved my back. I love the look and...
3	Home_and_Kitchen_5	1.0	CG	Missing information on how to use it, but it i...
4	Home_and_Kitchen_5	5.0	CG	Very nice set. Good quality. We have had the s...
...	...	...	...	...
40427	Clothing_Shoes_and_Jewelry_5	4.0	OR	I had read some reviews saying that this bra r...
40428	Clothing_Shoes_and_Jewelry_5	5.0	CG	I wasn't sure exactly what it would be. It is ...
40429	Clothing_Shoes_and_Jewelry_5	2.0	OR	You can wear the hood by itself, wear it with ...
40430	Clothing_Shoes_and_Jewelry_5	1.0	CG	I liked nothing about this dress. The only rea...
40431	Clothing_Shoes_and_Jewelry_5	5.0	OR	I work in the wedding industry and have to wor...

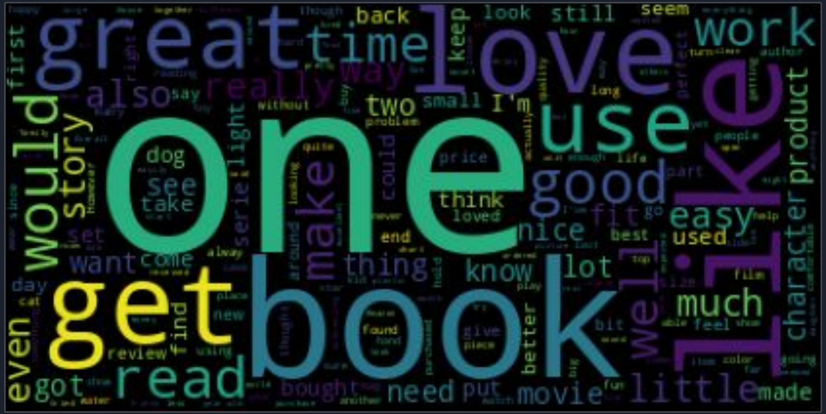
40432 rows × 4 columns

Data exploration

20216 fake reviews and 20216 (hopefully) authentic reviews



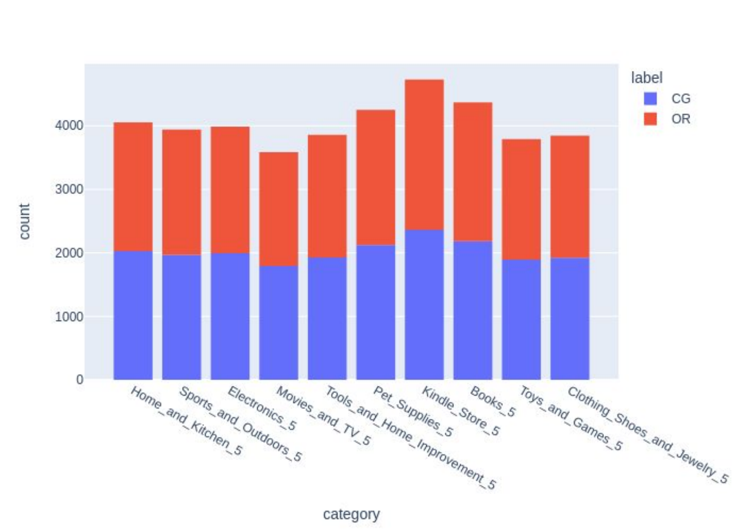
Fake reviews

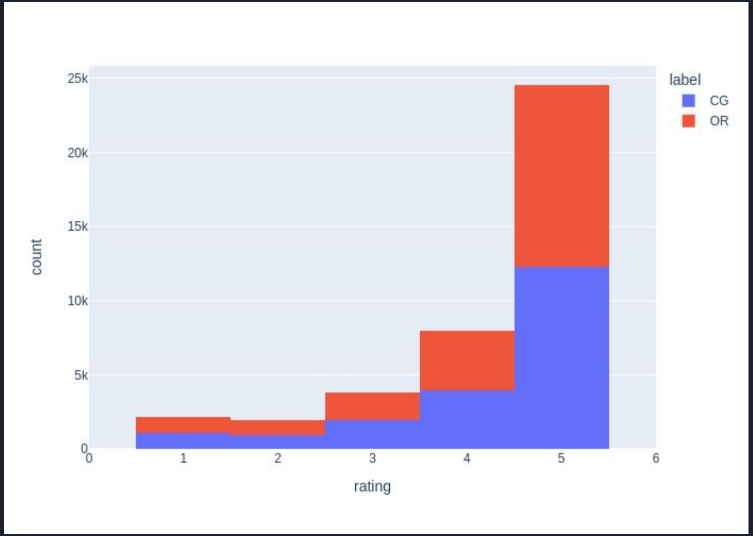


Real reviews

Data exploration

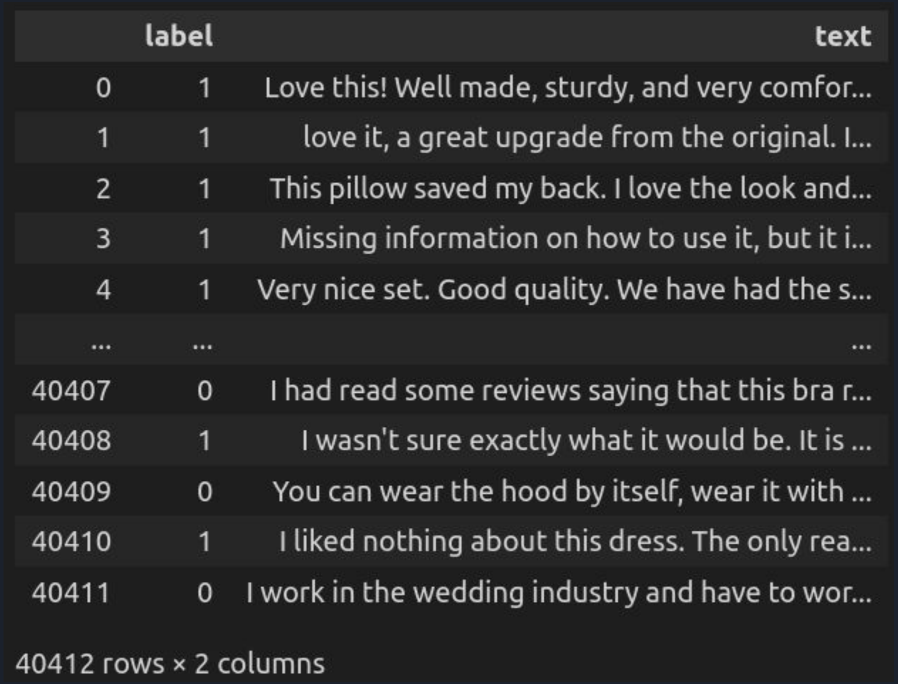
Checked that the number of fake/real reviews is equal regardless of the category. If we consider the rating they are almost equal. These columns will be removed since they are not useful.





Data preprocessing

- **Checked null values**: no null values found
- **Removed unuseful features**
- **Removed duplicates**: 20 objects removed, the dataset is still balanced
- Changed the label feature into a binary 0 (real) - 1 (fake)



	label	text
0	1	Love this! Well made, sturdy, and very comfor...
1	1	love it, a great upgrade from the original. I...
2	1	This pillow saved my back. I love the look and...
3	1	Missing information on how to use it, but it i...
4	1	Very nice set. Good quality. We have had the s...
...	...	...
40407	0	I had read some reviews saying that this bra r...
40408	1	I wasn't sure exactly what it would be. It is ...
40409	0	You can wear the hood by itself, wear it with ...
40410	1	I liked nothing about this dress. The only rea...
40411	0	I work in the wedding industry and have to wor...

40412 rows × 2 columns

Data preprocessing

Stemming

Is a process that removes the suffix from a word, obtaining a stem which is equal for the inflected variants of the same word

WORD	STEM
Studying	Studi
University	Univers
Better	Better

Lemmatization

Considers the context and converts the word to its meaningful base form, which is called Lemma

WORD	LEMMA
Studying	Study
University	University
Better	Good

Data preprocessing

Text preprocessing with TFIDF vectorizer

- Punctuation removal
- All text to lowercase
- Tokenization: after it ~47K features
- Stop words removal and Stemming / Lemmatization: after it ~24K / ~28K features
- Bag of words representation and TFIDF transformation

Example

- **Original**: Love this! Well made, sturdy, and very comfortable.
- **Tokenization**: [‘love’, ‘this’,’ well’, ‘made’, ‘sturdy’, ‘and’, ‘very’, ‘comfortable’]
- **Stopwords removal**: [‘love’, ‘well’, ‘made’, ‘sturdy’, ‘comfortable’]
- **Stemming**: [‘love’, ‘well’, ‘made’, ‘sturdi’, ‘comfort’]

Data processing

The classification has been performed with 3 different classifiers:

- Multinomial Naive Bayes
- Linear SVC
- Logistic Regression

Moreover three types of preprocessing have been compared:

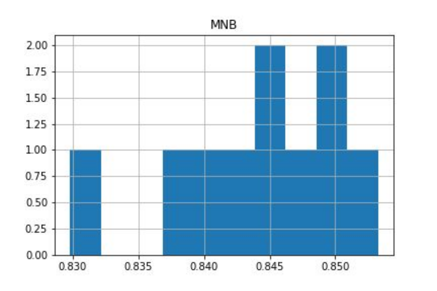
- Stemming with Porter Stemmer
- Stemming with Snowball Stemmer
- Lemmatization

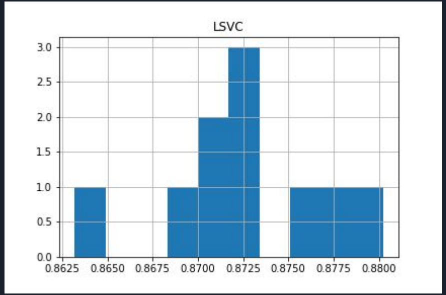
Stratified K Cross-Validation has been used, in order to perform statistically significant comparisons between the results of the classifiers

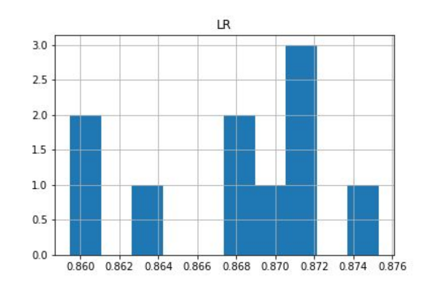
Performance evaluation

Stratified K cross-validation with K=10

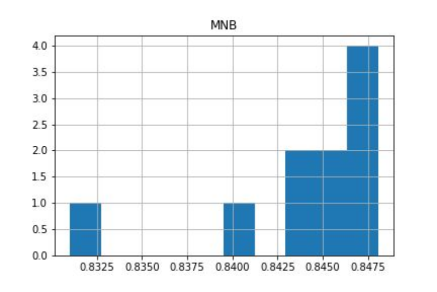
LEMMATIZATION

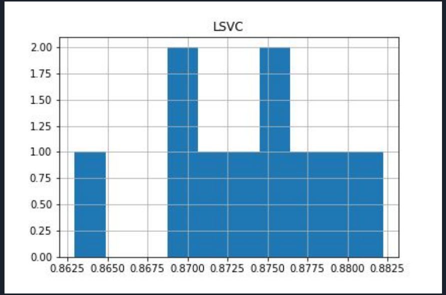


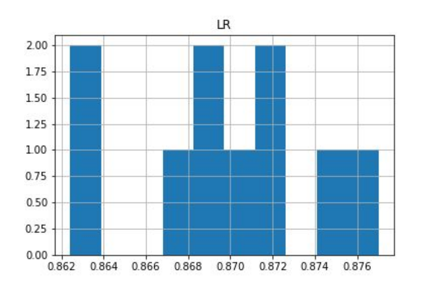




STEMMING
(Porter)



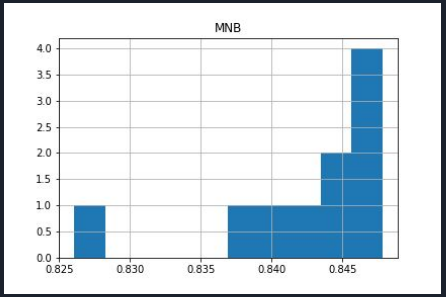


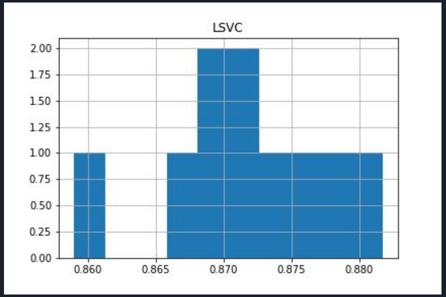


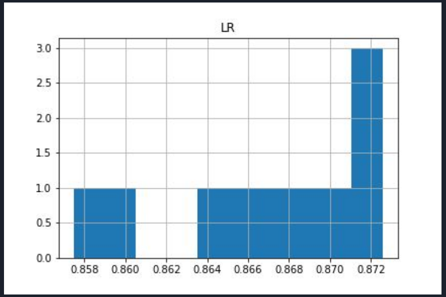
Performance evaluation

Stratified K cross-validation with K=10

STEMMING
(Snowball)







Performance evaluation

ACCURACY	Porter Stemmer	Snowball Stemmer	Lemmatizer
Multinomial NB	84.3%	84.3%	84.4%
Linear SVC	87.2%	87.1%	87.3%
Logistic Regression	86.7%	86.7%	86.8%

Performance evaluation

F1 SCORE	Porter Stemmer	Snowball Stemmer	Lemmatizer
Multinomial NB	85.0%	85.0%	85.2%
Linear SVC	87.3%	87.2%	87.3%
Logistic Regression	86.5%	86.5%	86.6%

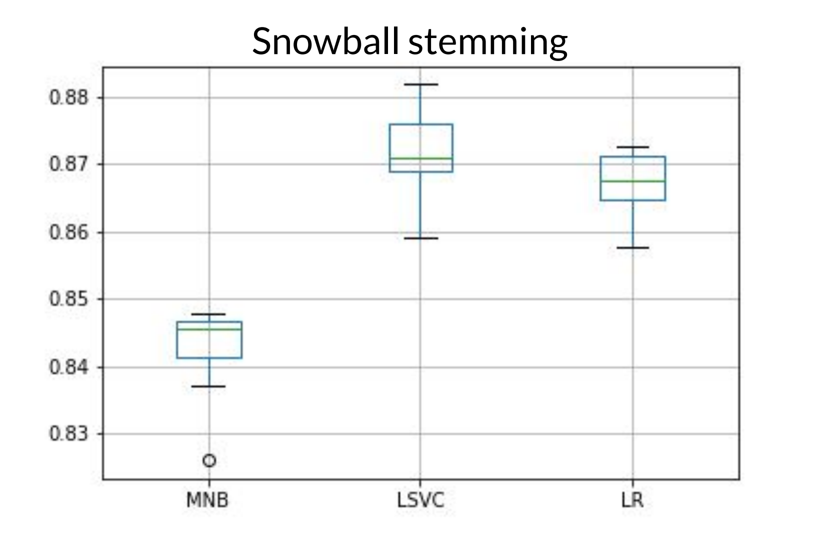
Performance evaluation

Preprocessing time (s)	Porter Stemmer	Snowball Stemmer	Lemmatizer
Time	29.6	20.5	240.1

Stemming, as expected, is faster! For bigger datasets should be preferred, but in this case a few minutes for lemmatization is still reasonable

Testing Null Hypothesis

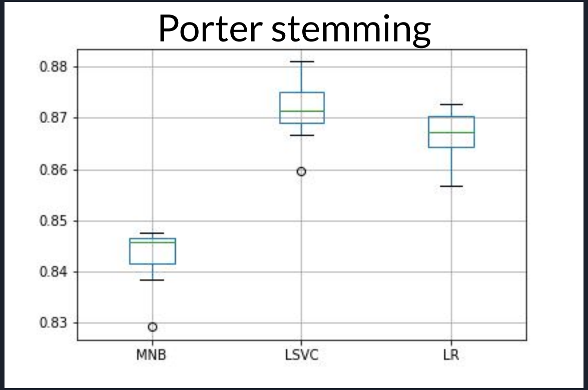
The best results, considering both time and accuracy, have been obtained with LinearSVC and Stemming. To compare them, after the cross validation, the Wilcoxon (nonparametric) statistical test have been used. The confidence used for the the test is α=0.05.



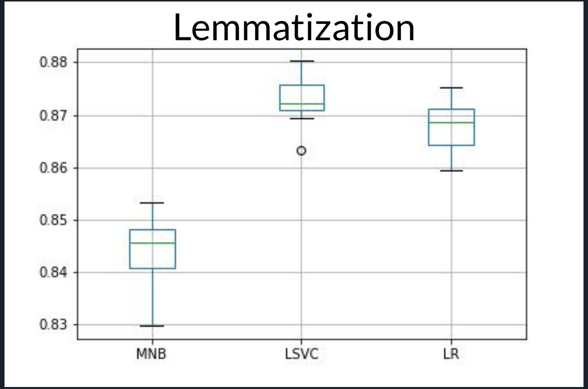
Classifiers Pair	p-value
MNB - LR	0.0051
LR - LSVC	0.0051
LSVC - MNB	0.0051

For each pair we reject the null hypothesis: the difference between the three models is statistically significant!

Testing Null Hypothesis


Porter stemming

Classifiers Pair	p-value
MNB - LR	0.0051
LR - LSVC	0.0051
LSVC - MNB	0.0051


Lemmatization

Classifiers Pair	p-value
MNB - LR	0.0051
LR - LSVC	0.0069
LSVC - MNB	0.0050

Conclusions and Next Steps

Conclusions:

- The best classifier turned out to be the Linear SVC model with Snowball stemming as preprocessing. The goal of creating a classifier for detecting fake reviews, with decent performances, can be considered achieved.
- Stratified K-cross validation have been performed so the results can be considered statistically significant.

Next steps:

- Investigating the generalizability the model on independent data. Unfortunately it really difficult to find other datasets that contains labeled fake reviews.
- Trying more complex ML algorithm. For example: [1] can achieve accuracies up to 98% on the dataset with more complex algorithm

References

- [1] Salminen, Joni, et al. "Creating and detecting fake reviews of online products." Journal of Retailing and Consumer Services 64 (2022): 102771