BUZZ
MARKETING
STRATEGE MARKETING
STRATEGEMARKETING.COM

Big data Marketing

Group project: Penguin app

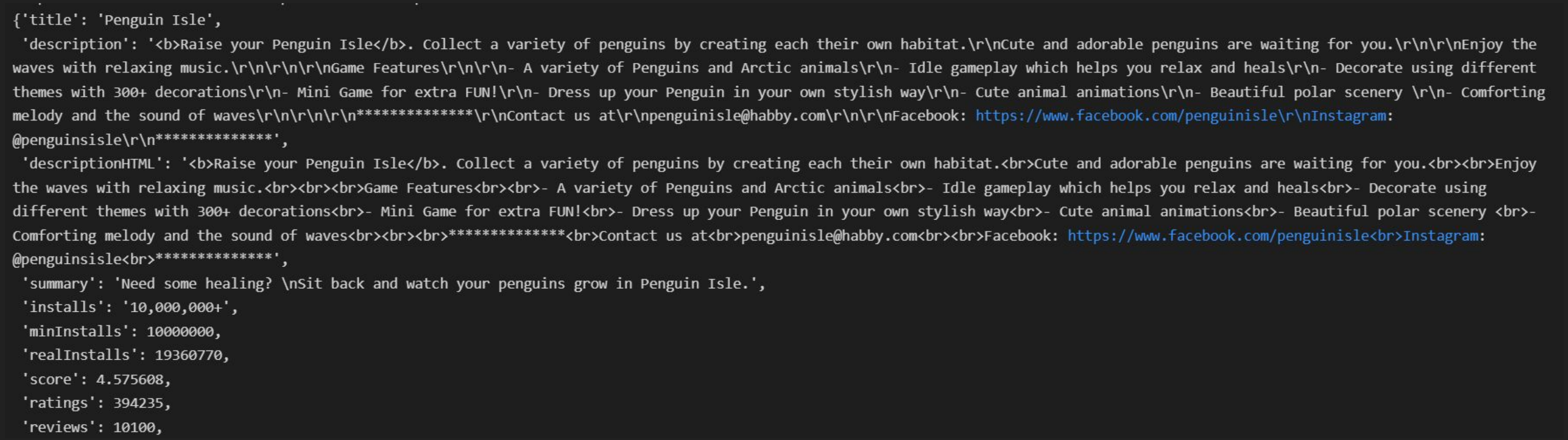
Egor Nikishin

Marina Nemtsova

Hossain Md Al Amin

App info

The product we have chosen is the **penguin app**. We have found out that Bouncemasters: Jumping Games (by **Casual Azur Games**), Penguin Isle (by **Habby**) and Super Penguins (by **Supersolid**) are the most popular penguin apps (in terms of number of downloads). The dataset was parsed from Google Play with the help of google-play-scrape library.



```
{'title': 'Penguin Isle',
 'description': '<b>Raise your Penguin Isle</b>. Collect a variety of penguins by creating each their own habitat.\r\nCute and adorable penguins are waiting for you.\r\n\r\nEnjoy the
waves with relaxing music.\r\n\r\n\r\nGame Features\r\n\r\n- A variety of Penguins and Arctic animals\r\n- Idle gameplay which helps you relax and heals\r\n- Decorate using different
themes with 300+ decorations\r\n- Mini Game for extra FUN!\r\n- Dress up your Penguin in your own stylish way\r\n- Cute animal animations\r\n- Beautiful polar scenery \r\n- Comforting
melody and the sound of waves\r\n\r\n\r\n**************\r\nContact us at\r\npenguinisle@habby.com\r\n\r\nFacebook: https://www.facebook.com/penguinisle\r\nInstagram:
@penguinsisle\r\n**************',
 'descriptionHTML': '<b>Raise your Penguin Isle</b>. Collect a variety of penguins by creating each their own habitat.<br>Cute and adorable penguins are waiting for you.<br><br>Enjoy
the waves with relaxing music.<br><br><br>Game Features<br><br>- A variety of Penguins and Arctic animals<br>- Idle gameplay which helps you relax and heals<br>- Decorate using
different themes with 300+ decorations<br>- Mini Game for extra FUN!<br>- Dress up your Penguin in your own stylish way<br>- Cute animal animations<br>- Beautiful polar scenery <br>-
Comforting melody and the sound of waves<br><br><br>**************<br>Contact us at<br>penguinisle@habby.com<br><br>Facebook: https://www.facebook.com/penguinisle<br>Instagram:
@penguinsisle<br>**************',
 'summary': 'Need some healing? \nSit back and watch your penguins grow in Penguin Isle.',
 'installs': '10,000,000+',
 'minInstalls': 10000000,
 'realInstalls': 19360770,
 'score': 4.575608,
 'ratings': 394235,
 'reviews': 10100,
```

Data parsing

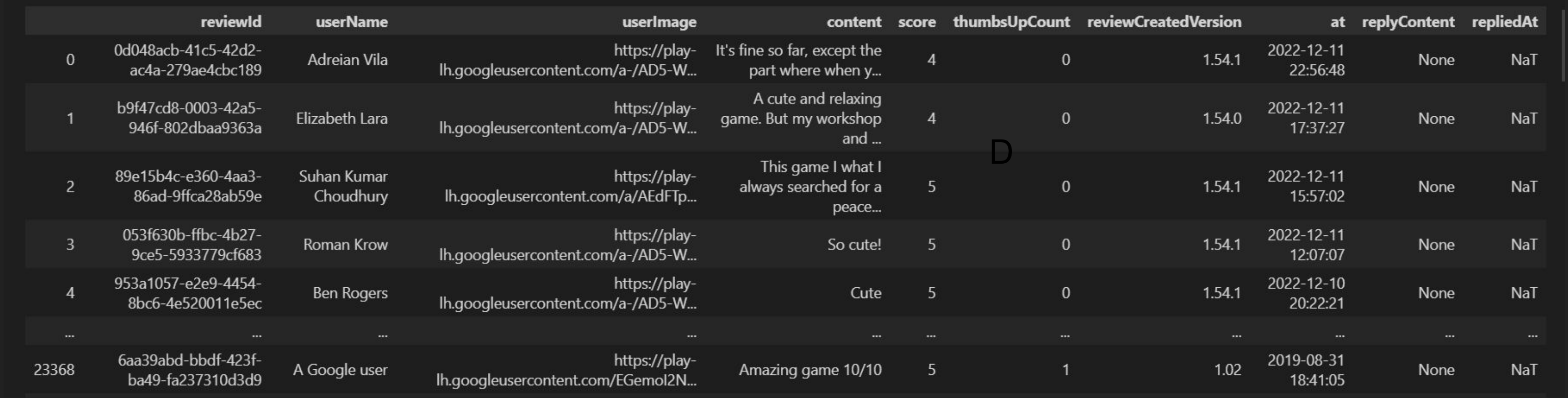
We have received the following parameters of app's reviews:

- at - review date
- content - review text translated to English
- score - review or product rating out of 5
- userName - author account name
- separate table info - info about the product

23373 entries

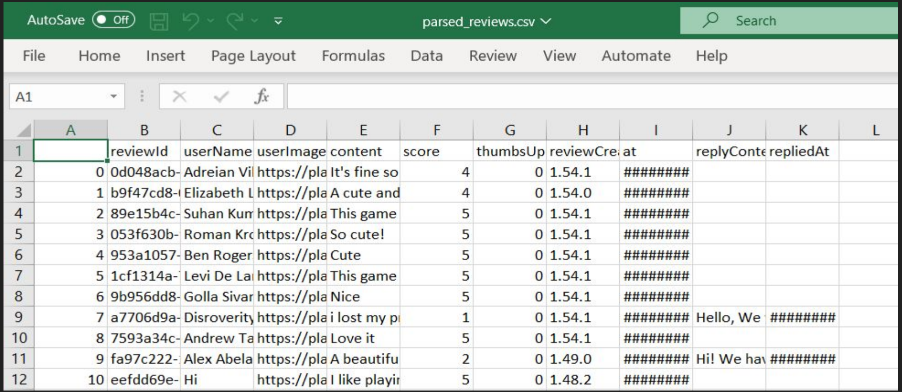
10 columns

memory usage: 1.8+ MB



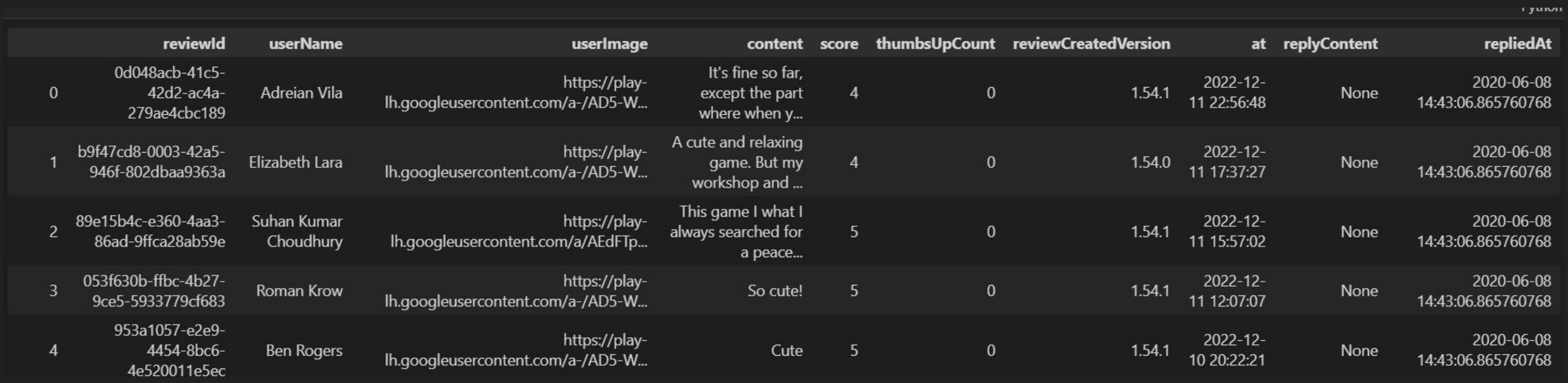
	reviewId	userName	userImage	content	score	thumbsUpCount	reviewCreatedVersion	at	replyContent	repliedAt
0	0d048acb-41c5-42d2-ac4a-279ae4cbc189	Adreian Vila	https://play-lh.googleusercontent.com/a-/AD5-W...	It's fine so far, except the part where when y...	4	0	1.54.1	2022-12-11 22:56:48	None	NaT
1	b9f47cd8-0003-42a5-946f-802dbaa9363a	Elizabeth Lara	https://play-lh.googleusercontent.com/a-/AD5-W...	A cute and relaxing game. But my workshop and ...	4	0	1.54.0	2022-12-11 17:37:27	None	NaT
2	89e15b4c-e360-4aa3-86ad-9ffca28ab59e	Suhan Kumar Choudhury	https://play-lh.googleusercontent.com/a/AEdFTp...	This game I what I always searched for a peace...	5	0	1.54.1	2022-12-11 15:57:02	None	NaT
3	053f630b-ffbc-4b27-9ce5-5933779cf683	Roman Krow	https://play-lh.googleusercontent.com/a-/AD5-W...	So cute!	5	0	1.54.1	2022-12-11 12:07:07	None	NaT
4	953a1057-e2e9-4454-8bc6-4e520011e5ec	Ben Rogers	https://play-lh.googleusercontent.com/a-/AD5-W...	Cute	5	0	1.54.1	2022-12-10 20:22:21	None	NaT
...	...	...	...	...	...	...	...	...	...	...
23368	6aa39abd-bbdf-423f-ba49-fa237310d3d9	A Google user	https://play-lh.googleusercontent.com/EGemoI2N...	Amazing game 10/10	5	1	1.02	2019-08-31 18:41:05	None	NaT

Data preprocessing



	A	B	C	D	E	F	G	H	I	J	K	L
1		reviewId	userName	userImage	content	score	thumbsUp	reviewCre	at	replyConte	repliedAt	
2	0	0d048acb-	Adreian Vil	https://pla	It's fine so	4	0	1.54.1	########			
3	1	b9f47cd8-(	Elizabeth L	https://pla	A cute and	4	0	1.54.0	########			
4	2	89e15b4c-	Suhan Kum	https://pla	This game	5	0	1.54.1	########			
5	3	053f630b-	Roman Krc	https://pla	So cute!	5	0	1.54.1	########			
6	4	953a1057-	Ben Roger	https://pla	Cute	5	0	1.54.1	########			
7	5	1cf1314a-	Levi De La	https://pla	This game	5	0	1.54.1	########			
8	6	9b956dd8-	Golla Sivar	https://pla	Nice	5	0	1.54.1	########			
9	7	a7706d9a-	Disroverity	https://pla	i lost my pi	1	0	1.54.1	########	Hello, We	########	
10	8	7593a34c-	Andrew Ta	https://pla	Love it	5	0	1.54.1	########			
11	9	fa97c222-	Alex Abela	https://pla	A beautifu	2	0	1.49.0	########	Hi! We hav	########	
12	10	eefdd69e-	Hi	https://pla	I like playir	5	0	1.48.2	########			

Some of the rows didn't have a date of review (Nan), so we counted the mean of not-null dates and deleted the nulls. Both versions of the dataframe were saved in csv files



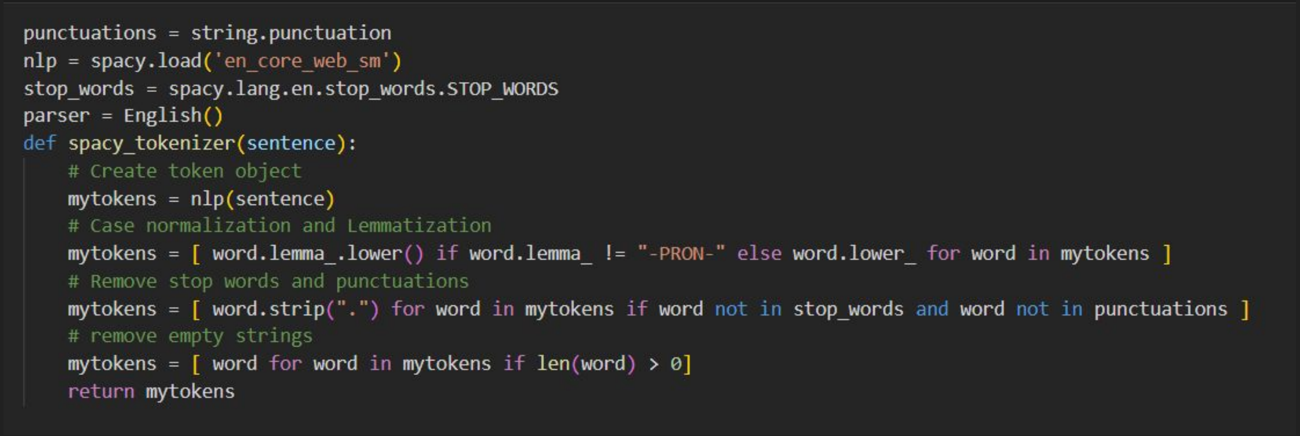
	reviewId	userName	userImage	content	score	thumbsUpCount	reviewCreatedVersion	at	replyContent	repliedAt
0	0d048acb-41c5-42d2-ac4a-279ae4cbc189	Adreian Vila	https://play-lh.googleusercontent.com/a-/AD5-W...	It's fine so far, except the part where when y...	4	0	1.54.1	2022-12-11 22:56:48	None	2020-06-08 14:43:06.865760768
1	b9f47cd8-0003-42a5-946f-802dbaa9363a	Elizabeth Lara	https://play-lh.googleusercontent.com/a-/AD5-W...	A cute and relaxing game. But my workshop and ...	4	0	1.54.0	2022-12-11 17:37:27	None	2020-06-08 14:43:06.865760768
2	89e15b4c-e360-4aa3-86ad-9ffca28ab59e	Suhan Kumar Choudhury	https://play-lh.googleusercontent.com/a/AEdFTp...	This game I what I always searched for a peace...	5	0	1.54.1	2022-12-11 15:57:02	None	2020-06-08 14:43:06.865760768
3	053f630b-ffbc-4b27-9ce5-5933779cf683	Roman Krow	https://play-lh.googleusercontent.com/a-/AD5-W...	So cute!	5	0	1.54.1	2022-12-11 12:07:07	None	2020-06-08 14:43:06.865760768
4	953a1057-e2e9-4454-8bc6-4e520011e5ec	Ben Rogers	https://play-lh.googleusercontent.com/a-/AD5-W...	Cute	5	0	1.54.1	2022-12-10 20:22:21	None	2020-06-08 14:43:06.865760768

Custom Tokenizer

We wrote our own function to tokenize and clean our dataset appropriately to remove unhelpful noise from our dataset.

This function will:

- Tokenize each word.
- Lemmatize each token. E.g. going → go, went → go
- Convert everything to lowercase
- Remove stop words Stop words are extremely common words that are irrelevant for our analysis and can be removed e.g. if, and, but, or

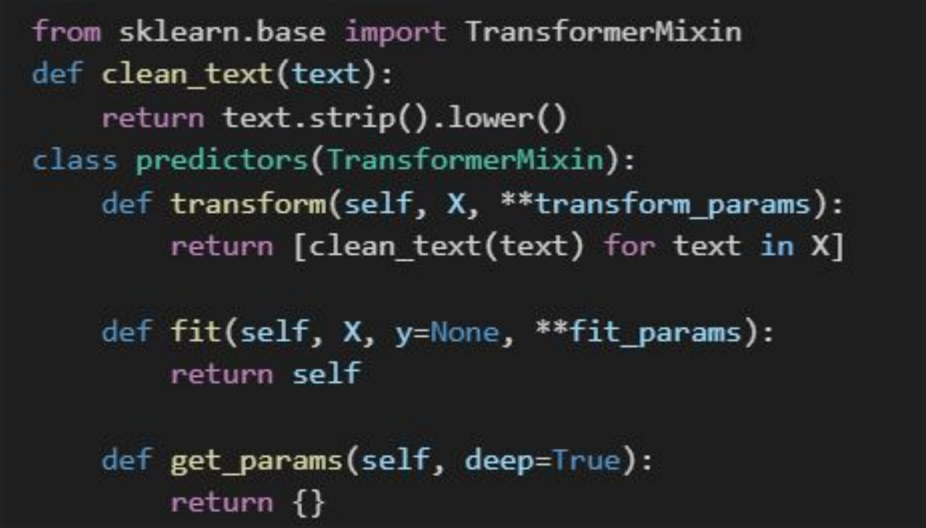


```
punctuations = string.punctuation
nlp = spacy.load('en_core_web_sm')
stop_words = spacy.lang.en.stop_words.STOP_WORDS
parser = English()
def spacy_tokenizer(sentence):
    # Create token object
    mytokens = nlp(sentence)
    # Case normalization and Lemmatization
    mytokens = [ word.lemma_.lower() if word.lemma_ != "-PRON-" else word.lower_ for word in mytokens ]
    # Remove stop words and punctuations
    mytokens = [ word.strip(".") for word in mytokens if word not in stop_words and word not in punctuations ]
    # remove empty strings
    mytokens = [ word for word in mytokens if len(word) > 0]
    return mytokens
```

Custom Transformer

We will be using the class TransformerMixIn from sklearn to create our own class transformer.

Our class will override the transform, fit and getparams from the main function and create our own. We will also pass a function called clean_text() that removes the spaces and converts the text into lowercase for an easier analysis.



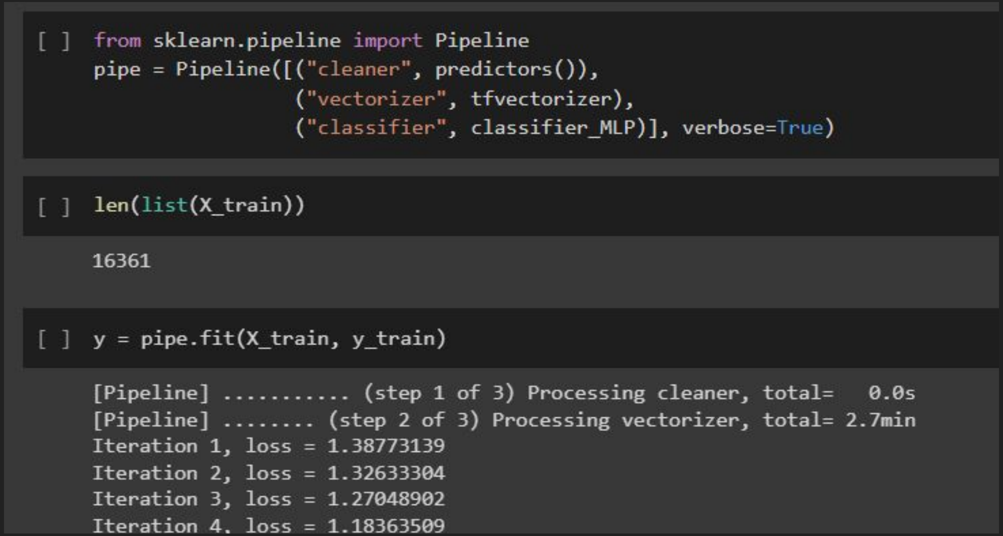
```
from sklearn.base import TransformerMixin
def clean_text(text):
    return text.strip().lower()
class predictors(TransformerMixin):
    def transform(self, X, **transform_params):
        return [clean_text(text) for text in X]

    def fit(self, X, y=None, **fit_params):
        return self

    def get_params(self, deep=True):
        return {}
```

Pipeline

We created a pipeline that cleans data, creates tokens (TfidfVectorizer with our custom tokenizer) and classifies (MLPClassifier with max. 400 iterations) the train data (70%) to make prediction of score (rating out of 5) based on the content of reviews.



```
[ ]  from sklearn.pipeline import Pipeline
     pipe = Pipeline([("cleaner", predictors()),
                      ("vectorizer", tfvectorizer),
                      ("classifier", classifier_MLP)], verbose=True)
```

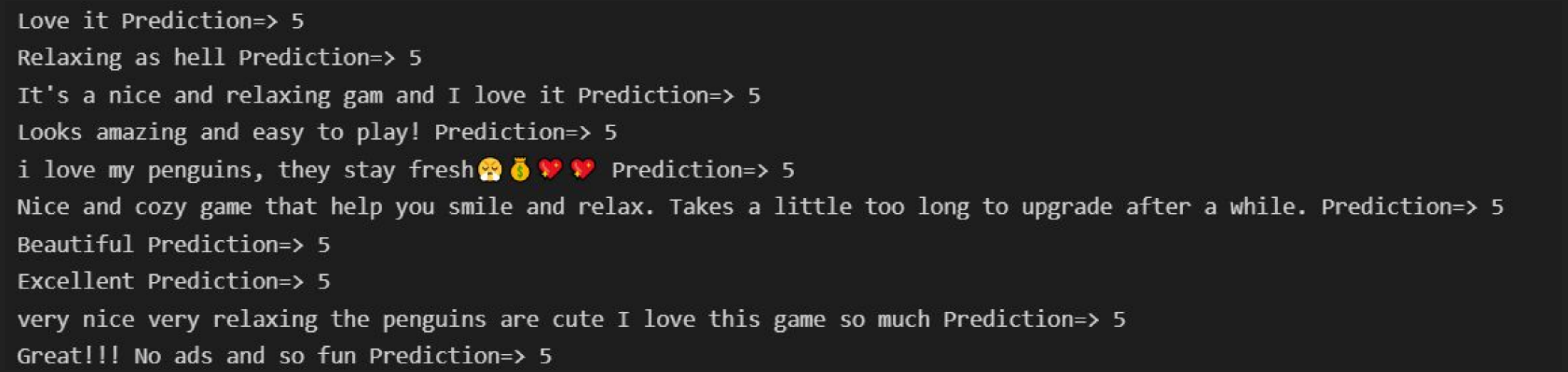
```
[ ]  len(list(X_train))
```

```
     16361
```

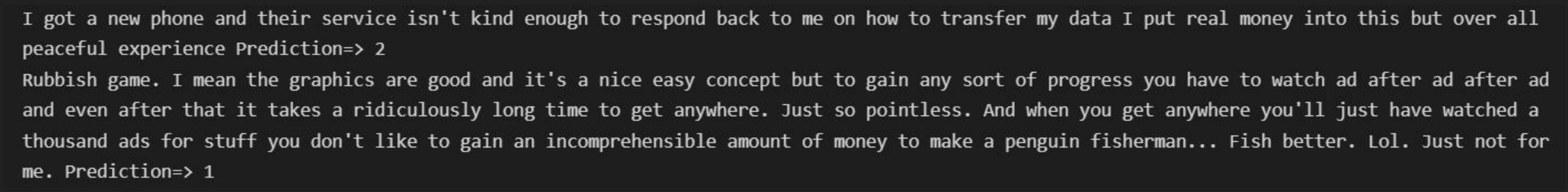
```
[ ]  y = pipe.fit(X_train, y_train)
```

```
     [Pipeline] ........... (step 1 of 3) Processing cleaner, total=   0.0s
     [Pipeline] ........ (step 2 of 3) Processing vectorizer, total= 2.7min
     Iteration 1, loss = 1.38773139
     Iteration 2, loss = 1.32633304
     Iteration 3, loss = 1.27048902
     Iteration 4, loss = 1.18363509
```

Prediction

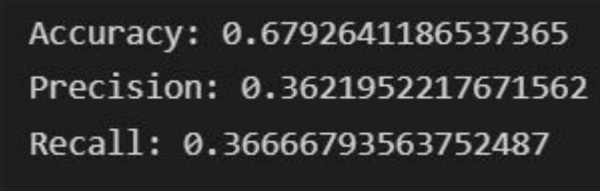


```
Love it Prediction=> 5
Relaxing as hell Prediction=> 5
It's a nice and relaxing gam and I love it Prediction=> 5
Looks amazing and easy to play! Prediction=> 5
i love my penguins, they stay fresh😤💰💖💖 Prediction=> 5
Nice and cozy game that help you smile and relax. Takes a little too long to upgrade after a while. Prediction=> 5
Beautiful Prediction=> 5
Excellent Prediction=> 5
very nice very relaxing the penguins are cute I love this game so much Prediction=> 5
Great!!! No ads and so fun Prediction=> 5
```

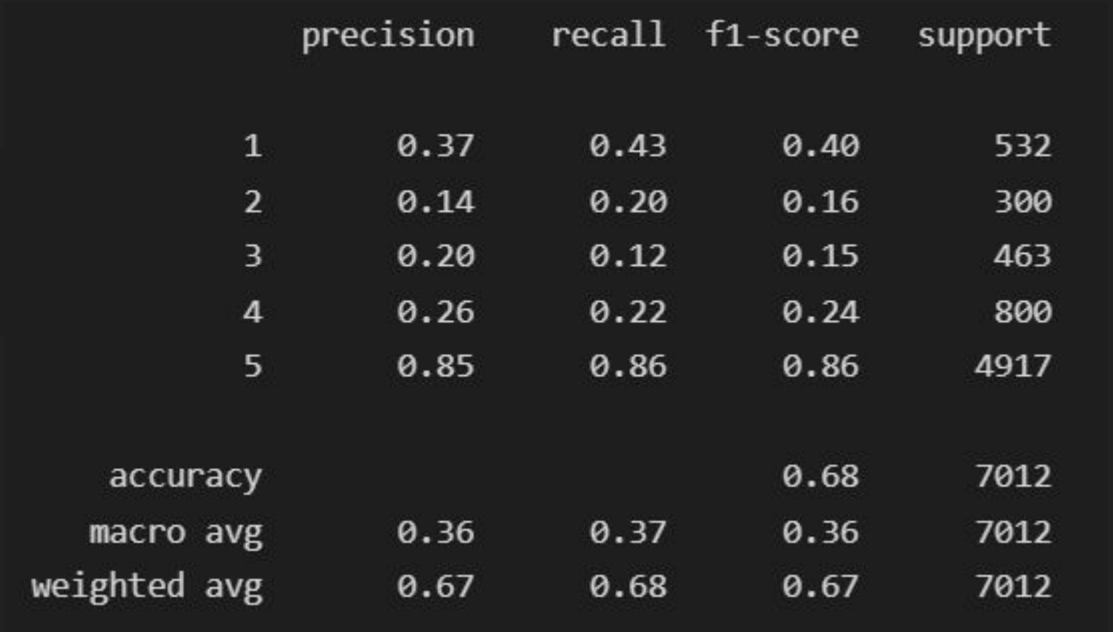


```
I got a new phone and their service isn't kind enough to respond back to me on how to transfer my data I put real money into this but over all
peaceful experience Prediction=> 2
Rubbish game. I mean the graphics are good and it's a nice easy concept but to gain any sort of progress you have to watch ad after ad after ad
and even after that it takes a ridiculously long time to get anywhere. Just so pointless. And when you get anywhere you'll just have watched a
thousand ads for stuff you don't like to gain an incomprehensible amount of money to make a penguin fisherman... Fish better. Lol. Just not for
me. Prediction=> 1
```

Key metrics



```
Accuracy: 0.6792641186537365
Precision: 0.3621952217671562
Recall: 0.3666793563752487
```

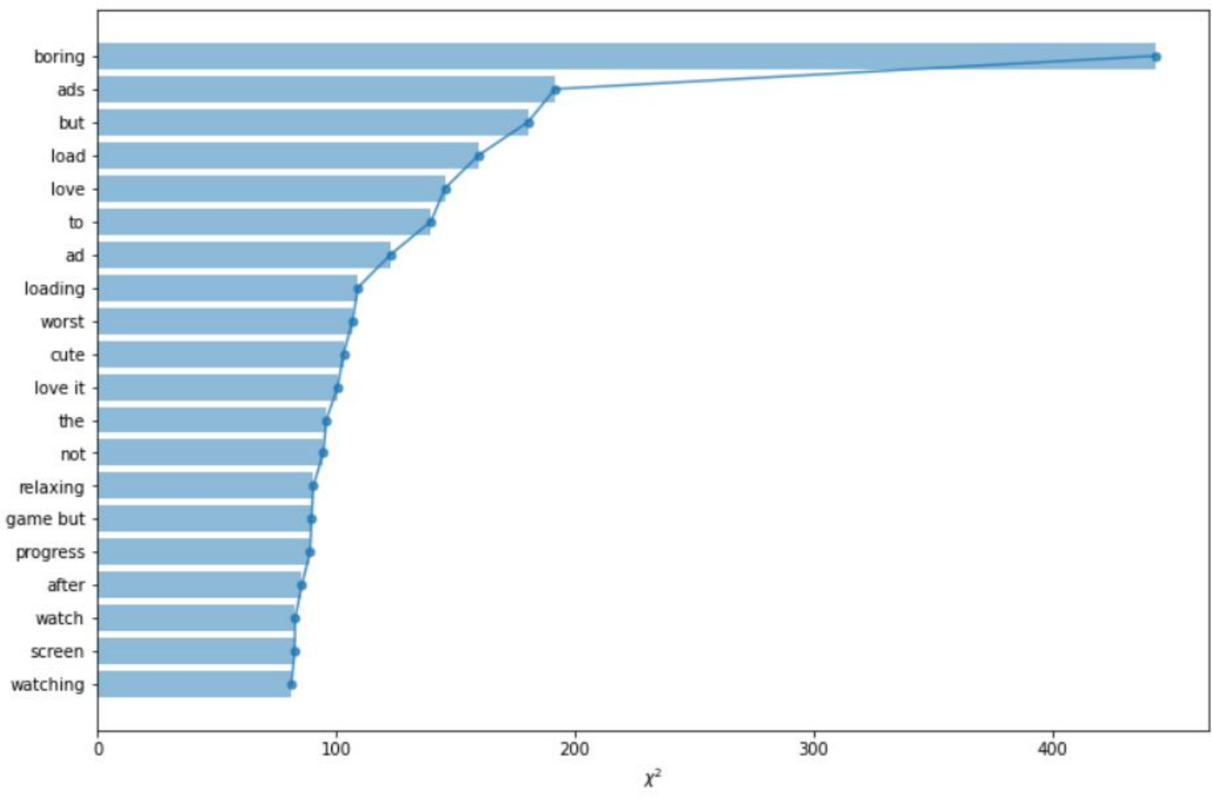


```
              precision    recall  f1-score   support

           1       0.37      0.43      0.40       532
           2       0.14      0.20      0.16       300
           3       0.20      0.12      0.15       463
           4       0.26      0.22      0.24       800
           5       0.85      0.86      0.86      4917

    accuracy                           0.68      7012
   macro avg       0.36      0.37      0.36      7012
weighted avg       0.67      0.68      0.67      7012
```

Chi-square scores to identify buzz words



Thank you