*Gabrielle Beinars*
**https://github.com/gbeinars**

# *Drug Reviews*

## Executive Summary

### *Background*

The internet contains a variety of sources for information, thus providing an abundance of data available for analysis. This project involves sentiment analysis and evaluating a dataset that contains a list of drugs, with various reviews for each drug, the condition that the drug was prescribed for, and a rating. Analysis of this data provides a form of feedback, and an overall view of negative versus positive drug reviews. By understanding which drugs were rated more positively for a specific condition, allows for better treatment by doctors. Sentiment analysis is a supervised machine learning technique where text analysis and natural language processing to identify and study the information.[6]

### *Problem Statement*

Evaluating this data allows for a better idea of what drugs are preferred and may perform better for a given condition. This can provide guidance for doctors prescribing a drug and comfort for patients that may be new to the drug or trying to find a drug that works for them.

- Are there certain conditions that have more overall reviews?
- Using the reviews, how accurately can the sentiment of the review be predicted?

### *Scope*

This project includes a variety of techniques in cleaning the data and exploratory data analysis, sentiment analysis and modeling, evaluating the results obtained and drawing conclusions from these results. Sentiment analysis involves analyzing text data to determine the sentiment behind it, and uses a combination of natural language processing (NLP) and machine learning. It is a common approach for evaluating customer feedback and can provide a variety of benefits.[4]

# Method

## *Data Source*

The dataset can be found at UCI Machine Learning Repository. The raw used to create this dataset was obtained from online pharmaceutical review sites. The dataset includes patient reviews on specific drugs, with a text review, a rating on patient satisfaction of the drug, as well as how many users found the review helpful. The data was downloaded as two text files, split at a ratio of 75:25, train to test. [1] The variables include the drug name, the condition for which the drug is taken for, a text review from the patient, a numerical rating that corresponds to the review, and a useful count, that represents the number of online users that found the review useful.

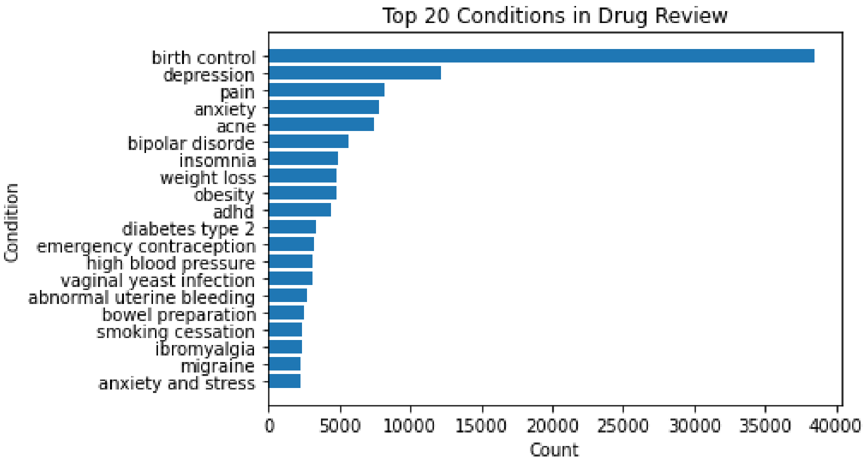
## *Data Import*

The data files were imported separately into Python as dataframes using pandas library where both dataframes were combined into one before exploring further. The next step was to check for missing values and duplicate observations, where 1,194 observations were removed because condition was blank, leaving the final shape of the dataset at 213,869 observations and six variables. Three variables, drug name, condition it was prescribed for, and review all required text cleaning, which involved lowercasing all words and removing any punctuation. Additionally, stop words were established and removed (Appendix). Stop words are English words that do not provide any value to the meaning of the text.[3]

## *Exploratory Data Analysis (EDA)*

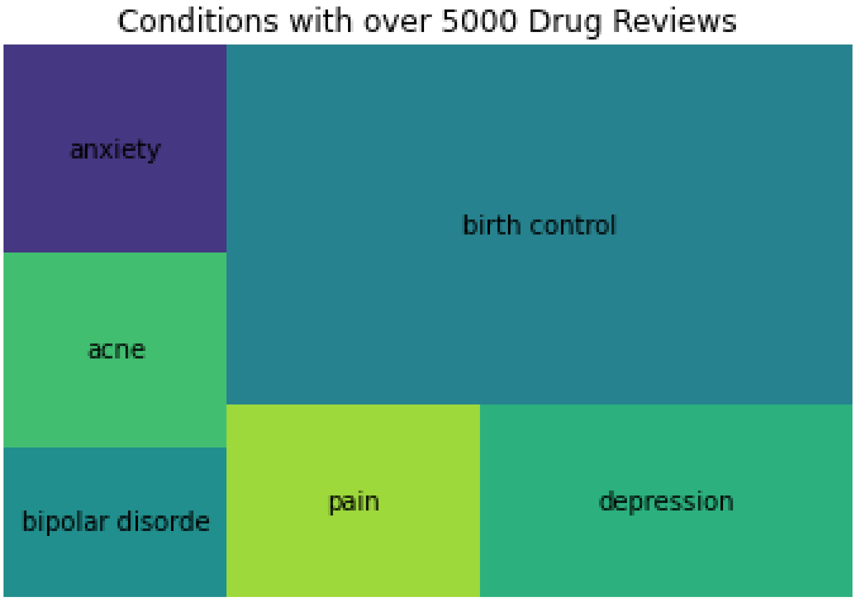
After the data cleaning was completed, exploratory data analysis began, where various visuals were created with the goal of identifying any patterns and gaining a better understanding of the data. First, the data was grouped by the condition for which each drug listed were prescribed to treat and sorted such that the top 20 conditions by drug review count (Figure 1).

*Figure 1: Horizontal Bar Chart of Top 20 Conditions in Drug Reviews*



A tree map was created to show conditions with over 5,000 drug reviews (Figure 2). The six conditions with over 5,000 drug reviews are birth control, depression, pain, anxiety, acne, and bipolar disorder.

*Figure 2: Tree Map of Conditions with Over 5,000 Drug Reviews*



The rating variable was explored by creating a bar chart that visualizes the count of review by rating from 1 to 10 (Figure 3). We can see that the largest count of reviews in one rating corresponded to a rating of 10. This variable will be the target variable used in modeling where ratings 8 through 10 were considered positive, and 1 through 7 considered negative. The final result after categorizing as positive or negative showed fairly balanced classes, with 56% of reviews positive, and 44% negative (Figure 4).

*Figure 3: Bar Chart of Count of Drug Review by Rating*

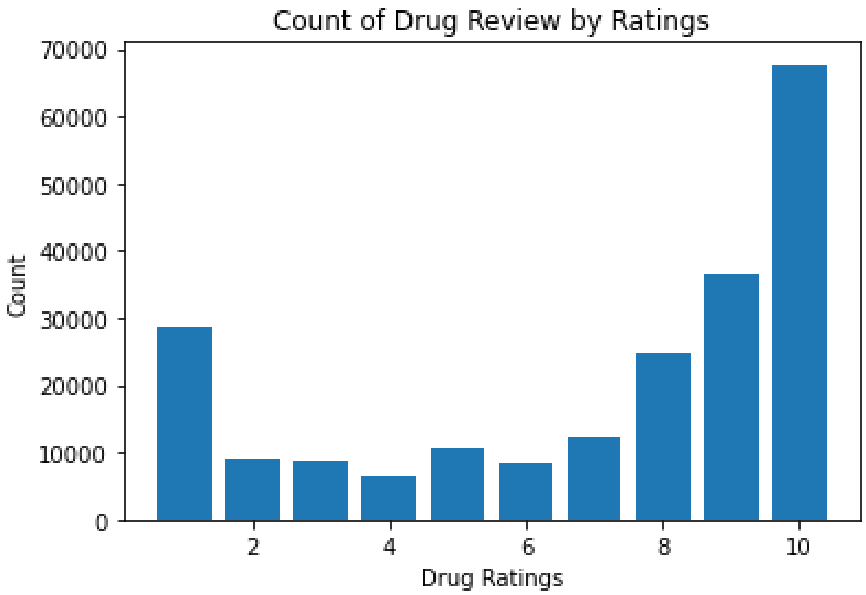
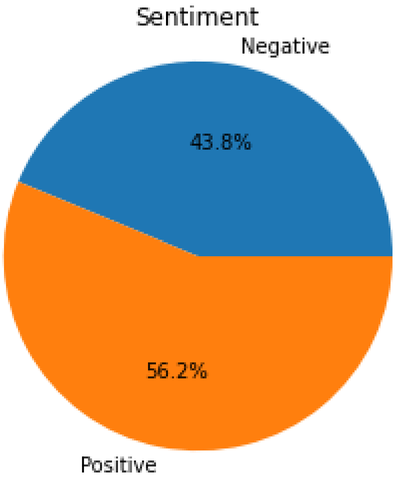


*Figure 4: Pie Chart of Positive vs Negative Reviews*



## *Modeling*

To begin with modeling, the text data must be converted to numerical. Positive and negative sentiment were replaced with 1 and 0, respectively. As the purpose of this project is sentiment analysis, a subset of the original data frame containing only the clean review and sentiment were carried forward. Python's sklearn library was used to vectorize the words, where the words are converted to numbers in a way that a machine can understand. This involves establishing the vectorizer, passing a list so that all the words in the text can be found and counted.[5] This process is referred to as creating a bag of words. A function was created to run all of this in one command, and return the most commonly used words which can be seen in the Appendix. The data then needs to be split to a training and test set so that the training data can be used to train the model and then the test data is used to test the model.

The first model was created using logistic regression. Logistic regression is a supervised machine learning technique specifically for classification problems such as this, where there is a binary target variable of positive or negative. Labeled data is provided by use of a training set, where it is then tested for accuracy with the test set. It is considered supervised because the model continuously predicts where each prediction is evaluated and corrected until it reaches an acceptable performance.[7]

Another modeling approach utilized is Naïve Bayes classifier which is considered a "probabilistic classifier" and contains a subset of more specific classifiers. For this project, multinomial naïve bayes and Bernoulli naïve bayes are used. Bernoulli naïve bayes classifier is used as the feature vectors consist of binary data.[8] Both the Bernoulli and multinomial Naïve Bayes models take the same approach of count vectorization where a bag of words is created, as discussed above.[9]

# Results

Out of the three models explored, logistic regression performed the best with an accuracy of 69.3%. The results of all models are shown in Table 1. Additional metrics used for evaluating the models are precision, recall and f1-score, which is included in the appendix as a classification report, and summarized in Table 2. These are metrics that evaluate the accuracy of the model and shows the number of times a model predicts correctly or incorrectly, by categorizing as true positive, true negative, false positive or false negative. Precision looks at the ratio of true positives compared to all positives, while recall measures the accuracy of predicting true positives. The F1 score looks at both precision and recall and represents the balance of the two metrics. The precision, recall and F1 score are broken down by the target variable of 0 and 1, negative and positive, respectively, and the results indicate highest precision and recall for positive sentiment throughout all models.[11]

Table 1: Modeling Results

| Model | Accuracy |
|---|---|
| Logistic Regression | 69.3% |
| Multinomial Naïve Bayes | 66.7% |
| Bernoulli Naïve Bayes | 66.4% |

Table 2: Classification Report Summary

| Model | Sentiment | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Logistic Regression | Negative | 0.60 | 0.66 | 0.63 |
| | Positive | 0.76 | 0.71 | 0.74 |
| Multinomial Naïve Bayes | Negative | 0.58 | 0.63 | 0.60 |
| | Positive | 0.74 | 0.69 | 0.71 |
| Bernoulli Naïve Bayes | Negative | 0.56 | 0.63 | 0.59 |
| | Positive | 0.75 | 0.69 | 0.71 |

# Discussion

## *Assumptions*

With logistic regression, the first assumption is that the target variable is binary, which is the primary reason that this model was selected, as our target variable is the sentiment, positive or negative.  A second assumption is that the observations are independent of one another.  This assumption is correct here as each observation is a separate review.  The third assumption is that there is no multicollinearity among the explanatory variables.  Multicollinearity is present when there are two or more variables that show high correlation to one another.  A fourth assumption is that there are no extreme outliers or observations that are strongly influencing the data.  A fifth assumption is that a linear relationship should be observed between the explanatory variables and the logit of the response.  The logit of the response is defined mathematically, where,

$Logit\ (p) = \log\left(\frac{p}{(1-p)}\right)$, where p is the probability of a positive outcome.

Lastly, it is assumed that the sample size is large enough such that conclusions can be drawn from the fitted model.[10]

There are two assumptions made when using the Naïve Bayes classifier.  First, is the assumption of independence where all observations are assumed to be independent of one another.  In this project, this would be the assumption that the words are not related.  The second assumption is that the frequencies are relative, meaning that it is assumed that each feature is balanced providing the same weight.[6],[9] It is important to note that these assumptions are often violated in the real world as it is not always practical, but models tend to still perform well.[9]

## *Limitations/Challenges*

One of the challenges frequently encountered with text cleaning, is whether the text is cleaned sufficiently for modeling.  There could be spelling errors or incorrect syntax, that could interpreted incorrectly, thus presenting issues in concluding the correct information.  It is important to be mindful that a computer cannot understand the meaning of text, but must learn it through the model.

### *Next Steps*

There are additional ways in which this dataset could be used for modeling. For example, classification could be used to classify the condition based on the review and the drug, or clustering could be used, as well. These are great options for in the future and to further understand drug reviews and how they relate to the drug and condition.

## References

[1] UCI Machine Learning Repository: Drug Review Dataset (drugs.com) data set. (n.d.). Retrieved November 6, 2021, from https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29.

[2] Felix Gräßer, Surya Kallumadi, Hagen Malberg, and Sebastian Zaunseder. 2018. Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning. In Proceedings of the 2018 International Conference on Digital Health (DH '18). ACM, New York, NY, USA, 121-125.

[3] Python - remove Stopwords. (n.d.). Retrieved November 14, 2021, from https://www.tutorialspoint.com/python_text_processing/python_remove_stopwords.htm.

[4] Shivanandhan, M. (2020, September 30). *What is sentiment analysis? A Complete Guide for Beginners*. freeCodeCamp.org. Retrieved November 14, 2021, from https://www.freecodecamp.org/news/what-is-sentiment-analysis-a-complete-guide-to-for-beginners/.

[5] Béjaoui, M. (2021, April 9). *Counting words in python with scikit-learn's countvectorizer*. Meher Béjaoui's Blog. Retrieved November 14, 2021, from https://www.meherbejaoui.com/python/counting-words-in-python-with-scikit-learn's-countvectorizer/.

[6] Mashalkar, A. (2021, March 7). *Sentiment analysis using logistic regression and naive Bayes*. Medium. Retrieved November 14, 2021, from https://towardsdatascience.com/sentiment-analysis-using-logistic-regression-and-naive-bayes-16b806eb4c4b.

[7] Kolamanvitha. (2021, May 9). *Twitter sentiment analysis using logistic regression*. Medium. Retrieved November 14, 2021, from https://medium.com/nerd-for-tech/twitter-sentiment-analysis-using-logistic-regression-ff9944982c67.

[8] Oliveira, L. (2017, August 1). *Sentiment Analysis Analysis Part 1 - naive Bayes classifier*. Medium. Retrieved November 14, 2021, from https://medium.com/nlpython/sentiment-analysis-analysis-ee5da4448e37#:~:text=Sentiment%20analysis%20is%20an%20area%20of%20research%20that,is%20positive%20or%20negative.%20The%20Naive%20Bayes%20classifier.

[9] *DSC 2 22 06 naive Bayes assumptions - learn.co*. Learn - A platform for education. (n.d.). Retrieved November 14, 2021, from https://portal.flatironschool.com/lessons/dsc-2-22-06-naive-bayes-assumptions.

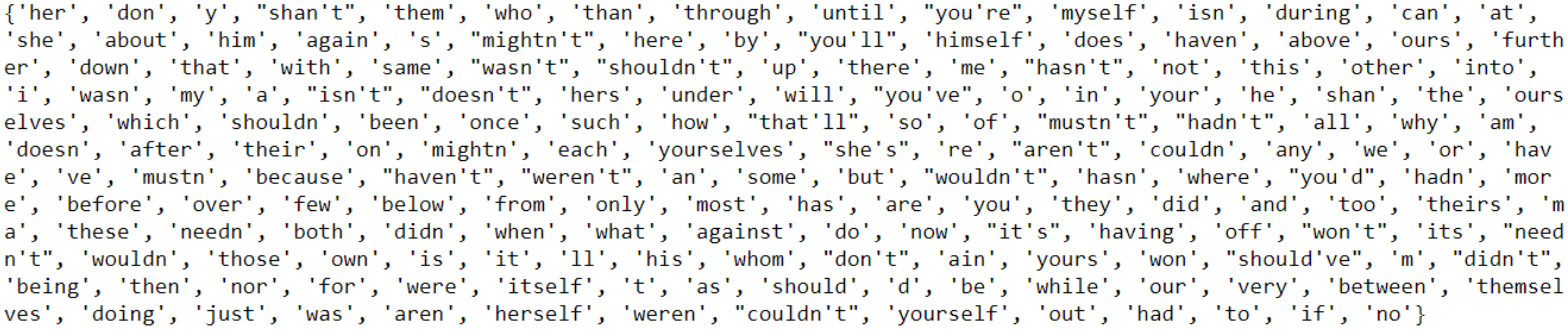
[10] Zach. (2020, October 13). *The 6 assumptions of logistic regression (with examples)*. Statology. Retrieved November 14, 2021, from https://www.statology.org/assumptions-of-logistic-regression/.

[11] *Precision, recall and F1 explained (in plain English)*. DataGroomr.com. (2021, May 24). Retrieved October 23, 2021, from https://datagroomr.com/precision-recall-and-f1-explained-in-plain-english/#:~:text=Precision%20and%20recall%20%28and%20F1%20score%20as%20well%29,where%20the%20model%20correctly%20predicts%20the%20positive%20class.

*Gabrielle Beinars*
https://github.com/gbeinars

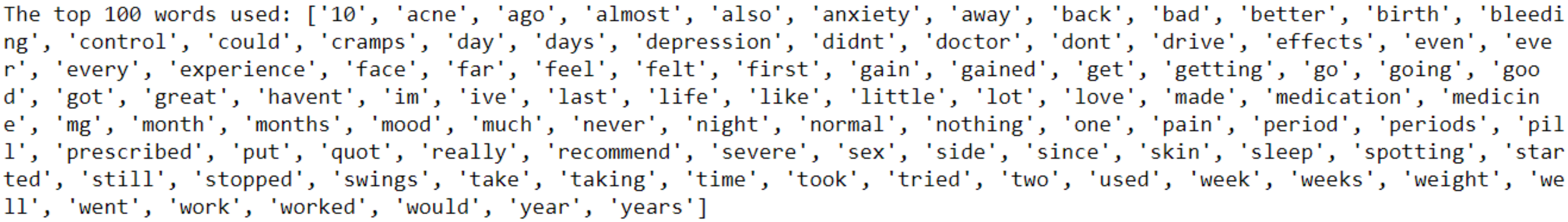
# Appendix

Below are the stop words removed from the drug reviews during text data cleaning.



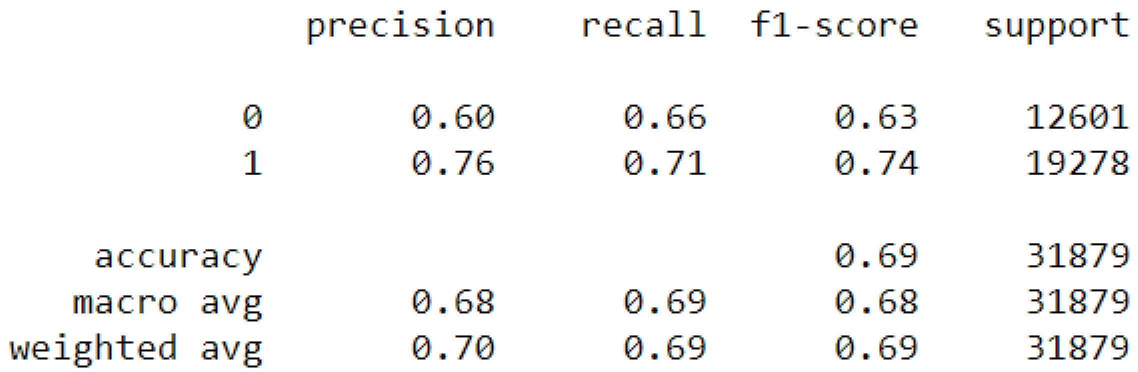
```
{'her', 'don', 'y', "shan't", 'them', 'who', 'than', 'through', 'until', "you're", 'myself', 'isn', 'during', 'can', 'at',
'she', 'about', 'him', 'again', 's', "mightn't", 'here', 'by', "you'll", 'himself', 'does', 'haven', 'above', 'ours', 'furth
er', 'down', 'that', 'with', 'same', "wasn't", "shouldn't", 'up', 'there', 'me', "hasn't", 'not', 'this', 'other', 'into',
'i', 'wasn', 'my', 'a', "isn't", "doesn't", 'hers', 'under', 'will', "you've", 'o', 'in', 'your', 'he', 'shan', 'the', 'ours
elves', 'which', 'shouldn', 'been', 'once', 'such', 'how', "that'll", 'so', 'of', "mustn't", "hadn't", 'all', 'why', 'am',
'doesn', 'after', 'their', 'on', 'mightn', 'each', 'yourselves', "she's", 're', "aren't", 'couldn', 'any', 'we', 'or', 'hav
e', 've', 'mustn', 'because', "haven't", "weren't", 'an', 'some', 'but', "wouldn't", 'hasn', 'where', "you'd", 'hadn', 'mor
e', 'before', 'over', 'few', 'below', 'from', 'only', 'most', 'has', 'are', 'you', 'they', 'did', 'and', 'too', 'theirs', 'm
a', 'these', 'needn', 'both', 'didn', 'when', 'what', 'against', 'do', 'now', "it's", 'having', 'off', "won't", 'its', "need
n't", 'wouldn', 'those', 'own', 'is', 'it', 'll', 'his', 'whom', "don't", 'ain', 'yours', 'won', "should've", 'm', "didn't",
'being', 'then', 'nor', 'for', 'were', 'itself', 't', 'as', 'should', 'd', 'be', 'while', 'our', 'very', 'between', 'themsel
ves', 'doing', 'just', 'was', 'aren', 'herself', 'weren', "couldn't", 'yourself', 'out', 'had', 'to', 'if', 'no'}
```

Below are the top 100 words in the reviews.



```
The top 100 words used: ['10', 'acne', 'ago', 'almost', 'also', 'anxiety', 'away', 'back', 'bad', 'better', 'birth', 'bleedi
ng', 'control', 'could', 'cramps', 'day', 'days', 'depression', 'didnt', 'doctor', 'dont', 'drive', 'effects', 'even', 'eve
r', 'every', 'experience', 'face', 'far', 'feel', 'felt', 'first', 'gain', 'gained', 'get', 'getting', 'go', 'going', 'goo
d', 'got', 'great', 'havent', 'im', 'ive', 'last', 'life', 'like', 'little', 'lot', 'love', 'made', 'medication', 'medicin
e', 'mg', 'month', 'months', 'mood', 'much', 'never', 'night', 'normal', 'nothing', 'one', 'pain', 'period', 'periods', 'pil
l', 'prescribed', 'put', 'quot', 'really', 'recommend', 'severe', 'sex', 'side', 'since', 'skin', 'sleep', 'spotting', 'star
ted', 'still', 'stopped', 'swings', 'take', 'taking', 'time', 'took', 'tried', 'two', 'used', 'week', 'weeks', 'weight', 'we
ll', 'went', 'work', 'worked', 'would', 'year', 'years']
```

Logistic Regression Classification Report

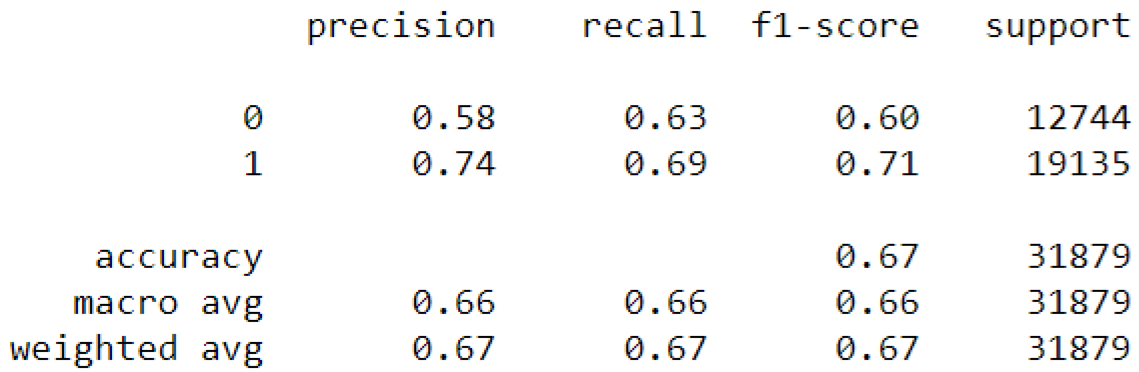


```
              precision    recall  f1-score   support

           0       0.60      0.66      0.63     12601
           1       0.76      0.71      0.74     19278

    accuracy                           0.69     31879
   macro avg       0.68      0.69      0.68     31879
weighted avg       0.70      0.69      0.69     31879
```

Multinomial Naïve Bayes Classification Report

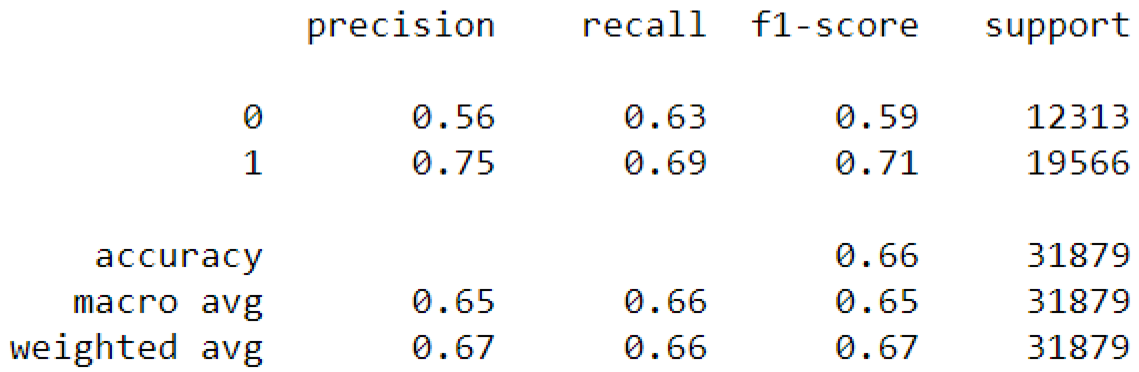


```
              precision    recall  f1-score   support

           0       0.58      0.63      0.60     12744
           1       0.74      0.69      0.71     19135

    accuracy                           0.67     31879
   macro avg       0.66      0.66      0.66     31879
weighted avg       0.67      0.67      0.67     31879
```

Bernoulli Naïve Bayes Classification Report



```
              precision    recall  f1-score   support

           0       0.56      0.63      0.59     12313
           1       0.75      0.69      0.71     19566

    accuracy                           0.66     31879
   macro avg       0.65      0.66      0.65     31879
weighted avg       0.67      0.66      0.67     31879
```