Text

Description automatically generated

**Course:** 2022-1005 IST 736 Text Mining

**Name:** Jacob N. Conard

**Email:** jconard@syr.edu

**Assignment:** Final Project

**Topic:** Final Project

**Date due:** 14 December 2022

**Date submitted:** 07 December 2022

**Number of Pages:** 10

**Final Project**

***Introduction:***

The Boeheim Public Relations firm has been retained to conduct analysis on sentiment classification (positive and negative determinations) of twitter post about popular video games. The goal of the firm’s analysis on the twitter post about popular video games is to determine how well computers can detect positive or negative sentiment of twitter post about popular video games by the words utilized in the statements. In this specific experiment two different machine learning models will be utilized to determine if computers can determine the sentiment of the twitter post about popular video games.

The Boeheim Public Relations firm conducted analysis on around 45,000 twitter post about popular video games, to make a determination on how well computers can detect positive and negative twitter post about popular video games.

Video game developers rely on social media posts to determine how well perceived the game is from the players around the world. When users and customers are commenting on social media about the game that they are playing the video game developers can leverage those social media posts to determine how well received the video game is being received by their customers. There are many social media posts about video games to include how well the game play and story line of the game is being perceived by the many customers and players around the world.

The amount of social media posts about video games can be overwhelming for video game developers. The Boeheim Public Relations firm has been retained to determine if computers can be leveraged to aggregate all of the twitter posts related to video games and formulate a positive or negative customer experience based on the reviews.

***Analysis:***

**About the Data**

The data is a consolidation of about 45,000 twitter post about popular video games around the world. Each review includes a sentiment label on whether the review is either positive or negative. The data is in a csv format, but requires tokenization and vectorization to get the data into a format where data analysis can be conducted on the data.

The Boeheim Public Relations Firm’s goal is to analyze the terms utilized in each social media post about video games to make a determination if there are specific terms (words / tokens) that are more common in social media posts that are labeled either positive or negative.

The data has an even split of twitter posts that are positive/negative as depicted in the visualization below:

Sentiment:

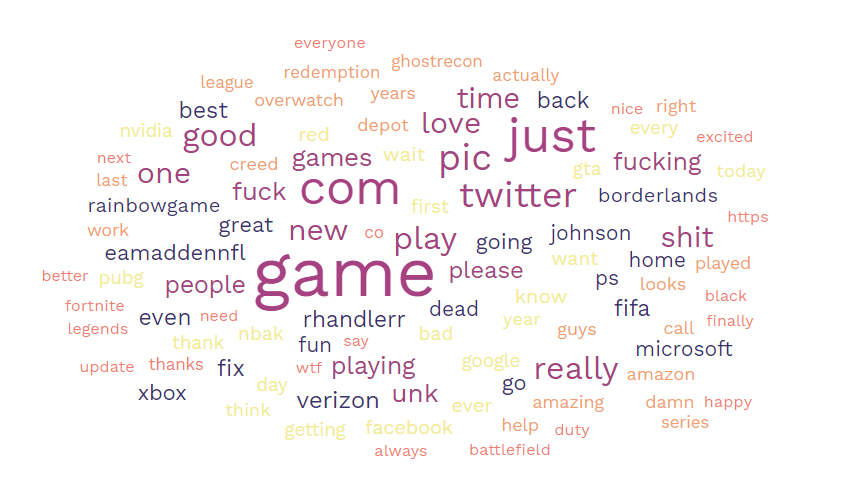
**Data Cleaning**

The data is a consolidation of about 45,000 twitter post about popular video games around the world. The original csv format required very little data cleaning. The data was in a csv format with the label of each twitter post the beginning of the content separated by a comma. The data needed to be read into a pandas dataframe to place the sentiment (positive/negative) in one column and the content of the tweet in the next column.

The twitter post about popular video games around the world also included many stop words, and those stop words were removed utilizing the built-in stop word removal from the sklearn CountVectorizer function.

**Data Exploration**

Once the stop words, numerical digits, special characters, and new lines were removed, the most important tokens in all of the reviews remained. A word cloud was created to visualize the most common tokens included in the twitter post about popular video games around the world:



**Models Utilized**

**Naïve Bayes Multinomial**

The Naïve Bayes Multinomial classification model was utilized during the analysis of the twitter post about popular video games dataset. Multinomial Naive Bayes algorithm is a probabilistic learning method that is mostly used in Natural Language Processing (NLP). The algorithm is based on the Bayes theorem and predicts the tag of a text such as a piece of email or newspaper article. It calculates the probability of each tag for a given sample and then gives the tag with the highest probability as output. Naive Bayes classifier is a collection of many algorithms where all the algorithms share one common principle, and that is each feature being classified is not related to any other feature. The presence or absence of a feature does not affect the presence or absence of the other feature. (<https://www.upgrad.com/blog/multinomial-naive-bayes-explained/>)

The data was split into a training and a test dataset and the labels were removed for proper predictions. The Naïve Bayes Multinomial classification model was utilized with default parameters and when the data was split into test and training datasets the test size parameter was set at 0.4.

**Support Vector Machine**

In machine learning, support vector machines (SVMs) are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis. SVMs are one of the most robust prediction methods, being based on statistical learning frameworks or VC theory proposed by Vapnik (1982, 1995) and Chervonenkis (1974). Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. SVM maps training examples to points in space so as to maximize the width of the gap between the two categories. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall (<https://en.wikipedia.org/wiki/Support_vector_machine>).

The data was split into a training and a test dataset and the labels were removed for proper predictions. The SVM classification model was utilized with default parameters and when the data was split into test and training datasets the test size parameter was set at 0.4.

**Support Vector Kernels**

Three different SVM Kernels were utilized to determine which one performed the best against the twitter post about popular video games dataset. The SVC or linear kernel was utilized, the Radial Basis Function or RBF kernel was utilized, and finally the SVM polynomial kernel was utilized and tested. The polynomial kernel is a kernel function commonly used with support vector machines (SVMs) and other kernelized models, that represents the similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models. The RBF kernel is a popular kernel function that is utilized in various kernelized learning algorithms and commonly used in support vector machine classification.

**Model Comparison**

The Naïve Bayes Multinomial and SVM model (with different kernel applications) will both be utilized to determine which model performs better. Those details will be expanded on in the Results section of the report as well as in the Conclusion at a high level.

***Results:***

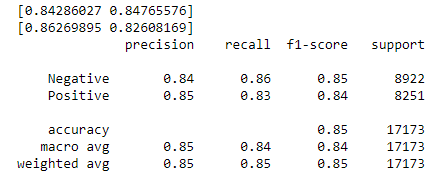
The Naïve Bayes Multinomial classification model was utilized to determine how well the model performed at predicting positive and negative sentiment in twitter post about popular video games. The training data was a fairly even slit of positive and negative views and can be visualized below:



The unigram CountVectorizer was utilized for all models with ‘latin-1’ encoding, a minimum document frequency of 5 and the removal of stop words.

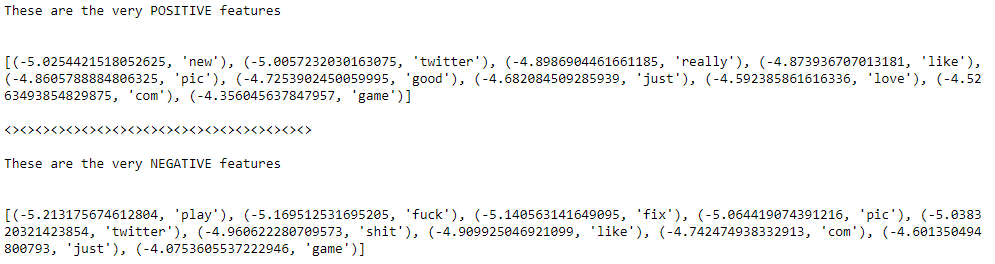
**Multinomial Naïve Bayes**

The Naïve Bayes multinomial classification model performed very well, 85% overall, in predicting the sentiment of the twitter post about popular video games given the tweet text. Below is the classification report for the Naïve Bayes multinomial classification model in predicting the sentiment of a twitter post about popular video games:



The classification report above shows that the Naïve Bayes multinomial classification model had about an 85% accuracy in predicting negative sentiment and about an 84% accuracy in predicting a positive sentiment. That is much better than a ‘random guess’ accuracy.

There were popular tokens that were weighted to determine whether the model would classify the tweet as either positive or negative. Those tokens and their weight can be visualized below:



The aforementioned visualization shows that in the determination of a positive tweet by the model the common terms identified were ‘love’, ‘good’, and ‘new’. This is just what would be expected and terms that a human would use to determine if a tweet was positive. Additionally, some of the common tokens that were weighted very heavy in the determination of a negative tweet tended to be derogatory words that a human would see and associate with negative or poor sentiment for the tweet. This model is performing perfectly which led to the overall 85% accuracy of the model overall in the positive and negative classification of twitter post.

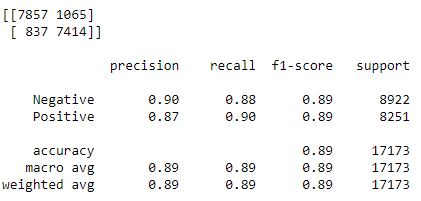
In the confusion matrix below, the model had the most errors incorrectly predicting a ‘Positive’ tweet as ‘Negative’. This is likely due to the use of negation words and the model could be better tuned to handle negation in a better manner.



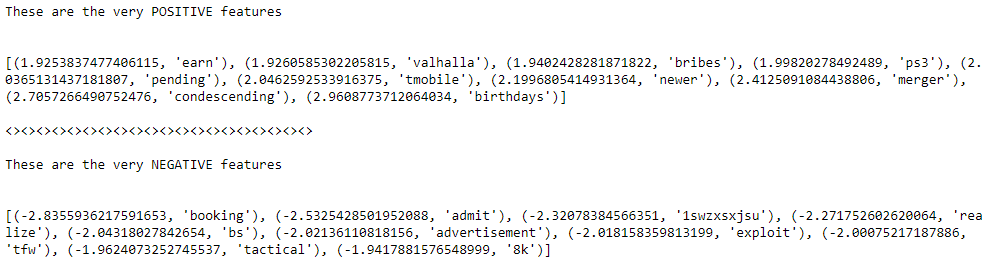


**Linear Support Vector Machine**

The Linear Support Vector Machine classification model performed very well in predicting the sentiment of the twitter post about popular video games given the tweet text. Below is the classification report for the Linear Support Vector Machine classification model in predicting the sentiment of a twitter post about popular video games:



The classification report above shows that the Linear Support Vector Machine classification model had about an 89% accuracy in predicting negative sentiment and about an 89% accuracy in predicting a positive sentiment. That is much better than a ‘random guess’ accuracy. This model performed better than the Multinomial Naïve Bayes model. Additionally, the Linear Support Vector Machine classification model identified some ‘positive tokens’ and some ‘negative tokens’. Those tokens can be seen below:



Finally, the Linear Support Vector Machine model was tested utilizing cross validation. In K Fold cross validation, the data is divided into k subsets. Now the holdout method is repeated k times, such that each time, one of the k subsets is used as the test set/ validation set and the other k-1 subsets are put together to form a training set. The error estimation is averaged over all k trials to get total effectiveness of our model. As can be seen, every data point gets to be in a validation set exactly once, and gets to be in a training set k-1 times. This significantly reduces bias as we are using most of the data for fitting, and also significantly reduces variance as most of the data is also being used in validation set. Interchanging the training and test sets also adds to the effectiveness of this method. As a general rule and empirical evidence, K = 5 or 10 is generally preferred, but nothing’s fixed and it can take any value (<https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f>). The overall cross validation score for the Linear Support Vector Machine model in the twitter post about popular video games was 88.7%, and the cross validation was tested using 5 folds. The accuracy of the five folds can be seen below:

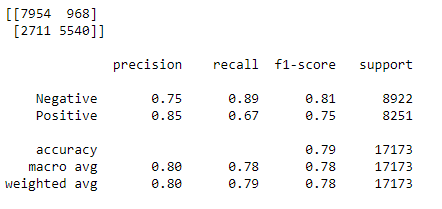


The overall accuracy of the testing with the 5-folds can be identified below:



**Support Vector Machine RBF Kernel**

The Support Vector Machine RBF classification model performed above average overall in predicting the sentiment of the twitter post about popular video games given the twitter post text. Below is the classification report for the Support Vector Machine RBF classification model in predicting the sentiment of a twitter post about popular video games:



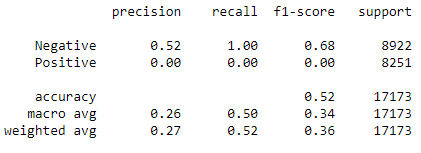
The classification report above shows that the Support Vector Machine RBF classification model had about an 81% accuracy in predicting negative sentiment and about an 75% accuracy in predicting a positive sentiment. That is much better than a ‘random guess’ accuracy. This model performed worse than the Multinomial Naïve Bayes model and also performed worse than the Linear Support Vector Machine classification model. The SVM model with the RBF kernel struggled the most assigning positive twitter post as negative. This is likely the cause of the model struggling with negation; however, even with the negation challenge the model still maintained an overall accuracy of 78%. Below is a visualization of the amount of errors where the actual sentiment was ‘Positive’ but the model predicted “Negative’.



**Support Vector Machine Poly Kernel**

The Support Vector Machine Poly classification model performed the worst overall in predicting the sentiment of the twitter post about popular video games given the twitter post text. Below is the classification report for the Support Vector Machine Poly classification model in predicting the sentiment of a twitter post about popular video games:





The classification report above shows that the Support Vector Machine Poly classification model had about a 0% accuracy in predicting negative sentiment and about an 68% accuracy in predicting a positive sentiment. That is much worse than a ‘random guess’ accuracy. The model overall accuracy was 51% and is about the same as the ‘random guess’ accuracy. This model was by far the worst performer overall and had a very difficult time predicating the accuracy of twitter post about popular video games with a negative sentiment and did not get one prediction correct. This is likely due to the fact that the SVM model with the Poly kernel was not the correct implementation for the twitter post about popular video games data.

**Model Comparison**

Overall, the best performance of a model was the Support Vector Machine linear kernel model in determining the positive or negative sentiment of twitter post about popular video games with a 89% accuracy in predicting negative sentiment and about a 89% accuracy in predicting a positive sentiment. The model with the worst overall performance was the Support Vector Machine Poly kernel model in determining the sentiment of a twitter post about popular video games with a 0% accuracy in predicting negative twitter post about popular video games and about a 68% accuracy in predicting a positive twitter post about popular video games.

The Boeheim Public Relations firm recommendation is for the Support Vector Machine linear kernel model to be utilized to determine the positive or negative sentiment of a twitter post about popular video games, and for the Support Vector Machine Poly kernel model not to be utilized in the prediction of sentiment for twitter post about popular video games.

Additionally, the Boeheim Public Relations firm also recommends that utilizing machine learning models to predict the sentiment of a twitter post about popular video games, given the sample data, is good practice and most models perform much better than ‘random guess’ accuracy, with the exception of the Support Vector Machine Poly kernel.

***Conclusion:***

The Boeheim Public Relations firm has been retained to conduct analysis on sentiment classification (positive and negative determinations) of twitter post about popular video games. The goal of the firm’s analysis on the twitter post about popular video games was to determine how well computers can detect positive or negative sentiment of twitter post about popular video games by the words utilized in the statements. In this specific experiment two different machine learning models will be utilized to determine if computers can determine the sentiment of the twitter post about popular video games.

Based on the firm’s analysis it is entirely possible to predict the sentiment (positive or negative) of twitter post about popular video games utilizing computers. The firm is able to predict the sentiment (positive or negative) of a twitter post about popular video games at a level of accuracy around 89%. Furthermore, based on the firm’s analysis the accuracy of the sentiment predictions of the twitter post about popular video games data is very dependent on selecting the correct machine learning algorithm. Some machine learning models perform much better than others, and the firm recommends that when doing sentiment analysis on twitter post about popular video games that close attention is paid to the models that are being utilized.

The firm’s recommendation is to utilize computers to predict the sentiment (positive or negative) of twitter post about popular video games. After the viability study of using computers to predict the sentiment of twitter post about popular video games, the firm is available for future contracts to utilize the prediction methods discovered to provide sentiment analysis of twitter post about popular video games.