SOCI354

TERM PROJECT

Aslı DÖNMEZ

Defne ŞAHİN

**INTRODUCTION**

**Method**:

In our term project we examined our dataset in this study in two primary ways.In first part, with using hypothesis testing and the OLS model, we tried to explain the correlations between certain variables that were derived from the excel data in the first section. We developed our data to create two distinct machine learning systems in the second section. In order to achieve the highest possible results when creating these systems, we studied with three different algorithms: RandomForestClassifier, Logistic Regression, and K-NN.

**Data**:

For this study, we used an Excel file with customer reviews of particular brands' products that people had posted on Twitter. This dataset, consisting of about 7865 rows and 20 columns, includes information about the language and location of the consumers, the date and time that they tweeted about the product, the number of retweets and interactions, the content of the tweets, the likes and comments, the company group that the product is a member of, the age and gender of the consumers, and more. It also contains thorough details on the emotion theme, which is especially significant.

Relations that we examined in first part:

1. Gender & Negative/Positive Emotion
2. Brand Follower Count & Negative Comments / Positive Comments
3. Total Interaction of Brand & Follower Count

Machine Learning System that we create:

1. Emotion Prediction System
2. Negative Positive Tweet Prediction System

**RESULTS & DISCUSSION**

**1) Gender & Negative/Positive Emotion:**

First, we wanted to look into the possible relationship between gender and the comments received from this extensive dataset. To accomplish this, we first extracted the *'gender'* and *'emotions'* columns from the larger dataset and constructed a new table with two columns. To conduct such actions, we first used the **chi2\_contingency** function from the **scipy.stats** library. The chi2\_contingency function determines whether there is a statistical link between two categorical variables. Following that, we used the **seaborn library** to illustrate the following outputs and create a table content. The **matplotlib.pyplot** feature is another tool that helps us generate plots. **Statsmodels.api** must also be installed for statistical analysis that will be mentioned during the process of analysis. Finally, we can create a cross- tabulation between the two categorical variables which are gender and emotions with using the function of pd.crosstab()

As a result of these tasks, emotions can be broken down into subcategories including approval (“onaylanma”) , disgust ( “iğrenme”), desire ( “ arzu”) , anxiety ( “kaygı”) , fear ( “korku”) , and gratitude ( “minnet” ) etc. This table illustrates categorized statistics on how many times a certain emotion was mentioned by both men and women. We next classify these unique emotions into two categories as positive emotions and negative emotions. For this, we first classify the 49 detailed emotions as positive or negative. Then, using the label function, we add a third column titled 'label' to our table, which includes the 'gender' and 'emotion' columns. This new column is classified as either positive or negative.

**Hypothesis testing:**

***H0:****There is no difference in the proportion of negative Tweets between males and females*

***HA****:There is a difference in the proportion of negative Tweets between males and females.*

To evaluate the hypothesis' validity, we must first apply our statistical understanding to the data and obtain the p-value. If the estimated p-value is less than the significance level (usually set at 0.05), we reject the null hypothesis (H0). Before computing the p-value, I arrange the counts of labeled negative and positive comments in a contingency table. We then use chi-square statistics to determine the p-value. The p-value is found to be 0.00051, which is lower than the significance level (0.05), so we reject the null hypothesis. This suggests that "there is a statistically significant association between gender and emotion labels in the dataset."

**2)Brand Follower Count & Comments**

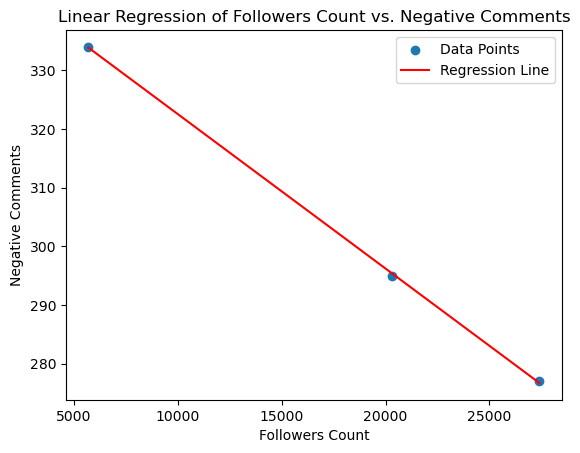
In our initial main data frame, we sorted the brand names referenced by customers in the keywords column. These keywords include hair care brands such as Clear, Elidor, Dove, and Elseve. This time, we developed a table to determine which emotions were associated with certain brand names. Then I repeated the data collection process, describing the emotions and displaying how many positive and negative responses a particular term (brand) received. For example, the keyword "Clear" can be related with 334 negative and 97 positive remarks.

**Hypothesis testing:**

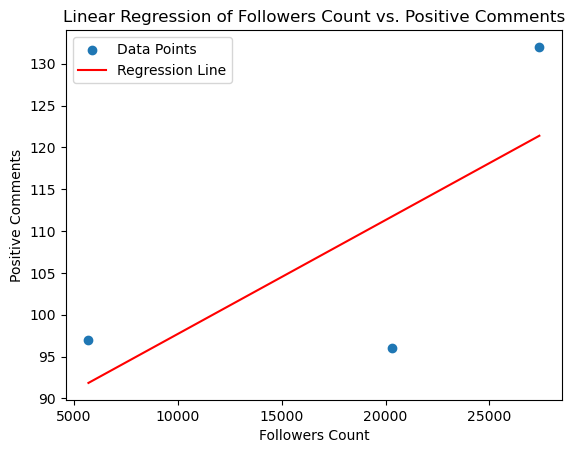
* **Negative Comments**

***H0:*** *There is no significant relationship between negative comments and followers count*

***HA:*** *There is a significant relationship between negative comments and followers count*



The link between the dependent variable, the quantity of unfavorable comments, and the independent variable, the number of followers, is depicted in this graph. It shows that the independent variables in this model account for 100% of the variance. Given that the P-value is less than 0.05, it can be concluded that the number of followers has a statistically significant role in explaining negative comments. The negative coefficients suggest that there is a strong correlation between the quantity of followers and the explanation of negative comments.

* ***Positive comments :*** 

To begin with, the R-squared number shows how much of the variance in the dependent variable can be attributed to the independent variable. The amount of brand followers is the independent variable in our model. The model indicates that this percentage is 54%. The P-value reflects the statistical significance of the independent variables in the model if it is employed, as it was in the previous hypothesis. It is determined that there is not a substantial statistical correlation between follower numbers and favorable comments because our P-value is bigger than 0.05 (significant point). In other words, the null hypothesis is true, and that is the conclusion drawn. The coefficients show the relationship between the number of followers and the independent factors. The statistical significance of the relationship between follower count and the number of positive comments is not supported by this low coefficient. Reiterating that the number of followers has no discernible impact on favorable comments, the prior 54% rate suggests that just 54% of the data can be explained.

| **description** | **followers\_count** | **Total Interaction** |
| --- | --- | --- |
| Clear Türkiye Resmi Twitter Sayfası | 5688.0 | 30652.0 |
|  |  |  |
| Dove Türkiye resmi Twitter hesabıdır - Dove il... | 20317.0 | 13672.0 |
| Elidor Türkiye resmi Twitter hesabı - Kampanya... | 27406.0 | 39366.0 |

## **3) Total Interaction of Brand & Follower Count**

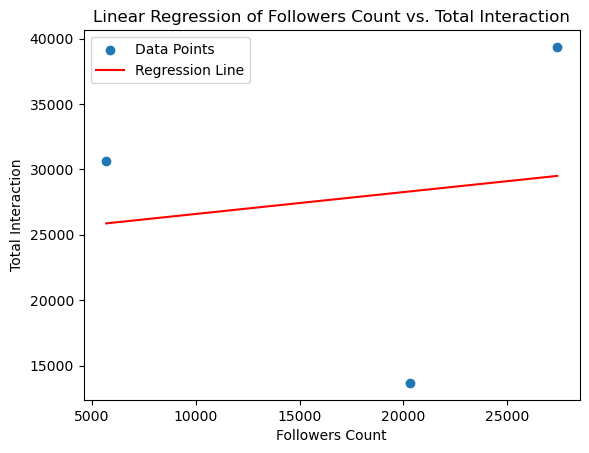
Another hypothesis we developed based on the same data is if brands with more Twitter followers receive more positive feedback. To examine these reactions, we isolated the brand's follower counts, total interactions, and the "description" part where consumers talked about the brand from the main dataset. We integrated them to make a table.

**Hypothesis testing:**

***H0:*** *There is no significant relationship between followers\_count and Total Interaction*

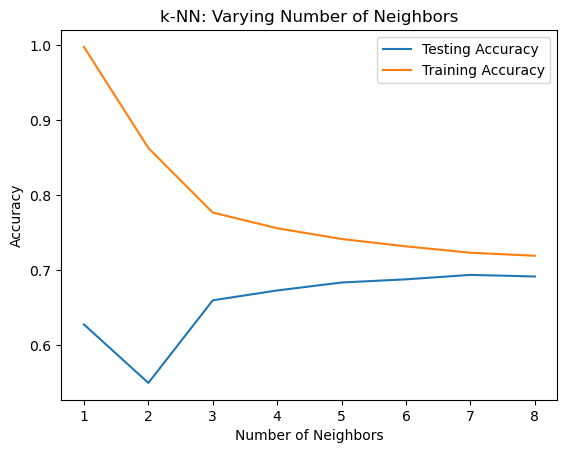
***HA:*** *There is a significant relationship between followers\_count and Total Interaction*

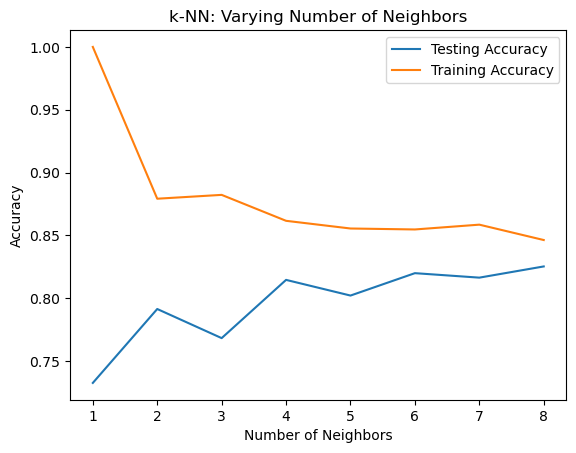
* **Total interaction**

When we look at the grafic related to total interaction and followers count, the p-value is 0.9010 and the F-statistic is 0.02048. This suggests that your model lacks statistical significance, meaning that followers\_count has no apparent effect on total interaction. Additionally, the reason for the horizontal regression line can be explained if changes in the independent variable (followers count) do not significantly affect changes in the dependent variable ( total interaction). The slope is nearly zero here.

**Machine Learning:**

We aimed to establish a system that would facilitate variable analyses, which would speed up investigations like the gender and emotion study we carried out before. We created a method to predict the feelings of the people who submit tweets for this reason, using the emotions and tweet\_text columns from our Twitter dataset. In order to create a more successful model and utilize our data more effectively, we then classified the emotions in our dataset as either positive or negative. This allowed us to create a more effective model that can identify which group each comment belongs to. We used three distinct strategies in the development of these two models in order to determine which system had the best level of performance.

**K-NN:**

First, we apply the K-Nearest Neighbors (K-NN) algorithm. Using the nearest neighbors of a data point in the training set, the K-NN algorithm predicts or classes the data point. The data is split into two categories using this algorithm: test and training. This data's division ratio is indicated by the test score. To get the best performance out of the model, this ratio must be used. This ratio was adjusted to 0.3 in our investigation, where 30% of the data were used to analyze and 70% for algorithm training.The models with the highest accuracy were the emotion prediction model (k=7) and the negative-positive comment model (k=6).

Determining the value of k, which indicates the number of nearest neighbors to take into account, is another crucial element. We produced a graph that contrasted the test and training accuracies in order to determine the best k value. The emotion prediction model's and the negative-positive comment prediction's k values at which the models' test and training accuracies were highest were found to be 7 and 6, respectively.

**Logistic Regression:**

Based on the combination of features in the dataset, this method predicts the likelihood of belonging to a specific category using a classification algorithm. This method learns the model with training data and applies it to the classification of new data, whereas the other method employs neighbors to make predictions. This method is split using the test score similarly to K-NN, producing a test score and a Classification Report.

**Random Forest Classifier:**

It is made up of several decision trees that have been trained using various random selections of data and feature subsets. Combining these trees' predictions yields a result when classifying new data. With this approach, every tree is trained separately, and the results of these trainings are used to make a group conclusion. The test score is also used to split this algorithm, which yields a Classification Report and accuracy.

**Cross-Validation:**

We used this approach to assess the models' capacity for generalization after completing the computations for all three models. This approach evaluates the model's performance on several subsets of data in an effort to lower the hazards of both underfitting and overfitting. Using distinct subsets of the training data for both training and testing the model, cross-validation yields a more reliable performance evaluation.

**Emotion Prediction Results:**

**KNN:**

Accuracy: 0.6936440677966101

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification Report** | precision | recall | f1-score | support |
| ['ofke'] | 0.37 | 0.08 | 0.13 | 167 |
| ['igrenme', 'onaylamama'] | 0.00 | 0.00 | 0.00 | 1 |
| … | … | … | … | … |
| ['notr'] | 0.62 | 0.25 | 0.36 | 171 |

**Random Forest Classifier:**

Accuracy: 0.713135593220339

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification Report** | precision | recall | f1-score | support |
| ['ofke'] | 0.44 | 0.07 | 0.11 | 167 |
| ['igrenme', 'onaylamama'] | 0.00 | 0.00 | 0.00 | 1 |
| … | … | … | … | … |
| ['notr'] | 0.90 | 0.21 | 0.34 | 171 |

**Logistic Regression:**

Accuracy: 0.701271186440678

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification Report** | precision | recall | f1-score | support |
| ['ofke'] | 0.44 | 0.07 | 0.11 | 167 |
| ['igrenme', 'onaylamama'] | 0.00 | 0.00 | 0.00 | 1 |
| … | … | … | … | … |
| ['notr'] | 0.90 | 0.21 | 0.34 | 171 |

**Cross-Validation Scores:**

|  |  |
| --- | --- |
| **KNN Cross-Validation Scores:** | **[0.70390554 0.68029064 0.69754768 0.70572207 0.69573115]** |
| KNN Mean CV Score: | **0.6966394187102634** |
| Logistic Regression Cross-Validation Scores: | **[0.70572207 0.70299728 0.70118074 0.70844687 0.70118074]** |
| Logistic Regression Mean CV Score: | **0.7039055404178021** |
| Random Forest Cross-Validation Scores | **[0.70844687 0.70935513 0.71298819 0.72116258 0.72207084]** |
| Random Forest Mean CV Score: | **0.7148047229791099** |

**Results:**

We determined that the Random Forest Classifier model (0.7131) had the highest accuracy among the systems when we assessed their overall accuracy. The K-Nearest Neighbors (KNN) model (0.6936) and Logistic Regression (0.7013) surpassed this success. But a closer look at the models showed that they performed poorly in predicting some emotions, even with their high accuracy rates. In particular, the model did not do well in predicting some emotions, as evidenced by precision, recall, and F1-score values that were around zero for several emotions.

The main cause of this poor performance is the small number of data points for specific emotions in our dataset. It should be mentioned that as the support count for various emotions rises, so does the model's capacity to predict those feelings. Emotions with more examples, like "neutral," for instance, can improve the predictive power of the model.

The Random Forest Classifier model has the highest score when we look at the models' cross-validation results. With the highest overall accuracy and cross-validation scores, this model is the strongest and has the greatest potential for accurate predictions. Therefore, using the Random Forest Classifier model would be the best course of action, barring the poor emotion representations in our dataset.

**Negative-Positive Prediction:**

Rather than developing distinct systems to recognize emotions such as desire, disapproval, and happiness, we tried to group these feelings into two primary categories: positive and negative, in order to overcome the low data problem that the first system was experiencing.

Our system was supposed to be able to choose between these two groups. We employed the previously developed "label\_emotions" function to divide our data into two sections. Negative emotions were determined as ['disapproval', 'anger', 'sadness', 'regret', 'surprise', 'disgust', 'anxiety', 'despair'], while positive emotions were identified as ['desire', 'happiness', 'gratitude', 'hope', 'love', 'approval'].

**KNN:**

Accuracy: 0.8199643493761141

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| negative | 0.83 | 0.94 | 0.88 | 406 |
| positive | 0.76 | 0.50 | 0.61 | 155 |

**Logistic Regression:**

Accuracy:0.8324420677361853

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| negative | 0.82 | 0.99 | 0.90 | 406 |
| positive | 0.93 | 0.43 | 0.58 | 155 |

**Random Forest Classifier:** **Cross-Validation Scores:**

|  |  |
| --- | --- |
| KNN Cross-Validation Scores: | **[0.79770992 0.82061069 0.79770992 0.79310345 0.79310345]** |
| KNN Mean CV Score: | **0.8004474861805739** |
| Logistic Regression Cross-Validation Scores: | **[0.7519084 0.77099237 0.77099237 0.77394636 0.75095785]** |
| Logistic Regression Mean CV Score: | **0.7637594688660758** |
| Random Forest Cross-Validation Scores | **[0.75572519 0.77862595 0.76717557 0.80076628 0.77777778]** |
| Random Forest Mean CV Score: | **0.7760141557719867** |

Accuracy:0.8342245989304813

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| negative | 0.82 | 0.99 | 0.90 | 406 |
| positive | 0.93 | 0.43 | 0.58 | 155 |

**Results:**

We found that the Random Forest (0.834) and Logistic Regression (0.832) models have marginally greater accuracy rates than the KNN (0.820) model when evaluating the overall accuracy rates of the systems. This suggests that the Logistic Regression and Random Forest models outperform the others overall.

When we compared the positive and negative accuracy rates in a detailed examination of the models, we evaluated the accuracy rates for the positive and negative states and discovered that all three models had a negative F1-score more than 90%. This shows that the models are good at identifying unfavorable tweets. In contrast to the negative F1-score, the positive F1-score was considerably lower. The primary cause of this discrepancy is that, in contrast to the positive class (155), the negative class (406) contains more examples of support. Therefore, the models have proved more successful in recognizing negatives since during model training, there are more negative examples than positive examples.

We found that the KNN model had the highest cross-validation score when comparing the models' scores. Overall though, despite not showing much success in the positive class, the Random Forest Classifier model stands out as the most appropriate model to be selected because of its highest accuracy rate and successful F1-score in the negative class.

Choosing the Random Forest Classifier model would therefore be the best course of action. With its successful F1-score in the negative class, good cross-validation score, and high accuracy rate, this model stands out. This option would be the best and most dependable one given the imbalances and class distributions in our dataset.

**REFERENCE**

* Data: shampoo\_tweet\_filtered\_merged\_23042024.xlsx