Sentiment Analysis Assignment using k-nearest neighbors.

STUDENT ID | SA 01D01

Assignment Objective

Evaluate different comments on different posts on social media: Instagram and TikTok by targeting a specific brand. You are asked to classify the sentiment of these comments for those posts based on emotion: Positive, Negative and Neural and determine which platform has the most negative or positive comments. Also, predict if new unforeseen comments will be the same.

Methodology of Data Collection

For the purposes of the assignment, the researcher contemplated using Pepsi Global as the specific brand for conducting the objective of this challenge. The researcher started with scraping the comments on TikTok using a JavaScript language web scraper to extract all the comments made on several posts on the Pepsi TikTok platform. More specifically, the code was pasted and ran in the web console and the comments were extracted. This included the first order comments as well as all the replies to the first order comments. Afterwards, they were committed to a CSV file. Additionally, the comments for the most recent posts made on Instagram were also extracted and were placed in a separate CSV file as well. It was discovered by the researcher that the comments made throughout the Instagram posts were found to be sparser than the comments found throughout the TikTok posts. This highlights an interesting feature where on average, there are more user engagement on TikTok than Instagram in the comments section. This phenomenon is based on the design of the social media application and how its contents reach its users as well as the brand.

Feature Engineering using Excel.

The comments undergo manual feature engineering. This was done so that the training dataset was suitable for increasing the accuracy of the KNN model. This involved converting abbreviated texts to its English root word, converting Spanish and Arabic languages to English using the Excel translate tool and inputting and rewording some of the comments as some words were missing which could decrease the accuracy of the machine learning predictions.

Overview of the Model

The model involved performing sentiment analysis using a Pre-trained BERT model and subsequently applying a K-Nearest Neighbors (KNN) classifier on the predicted sentiment labels. As it relates to the type of BERT model used, it was a fine-tuned BERT Model for sentiment analysis called '[nlptown/bert-base-multilingual-uncased-sentiment]' from the Hugging Face platform. BERT was used because provides a utility for contextualized word embeddings, which captures the meaning of a word based on the surrounding context in a sentence. Additionally, it was trained on a massive amount of text data and can capture semantic relationships, nuances, and subtleties in language. The hope was that it could accurately classify the sentiment of the social media comments based on its context.

Intricates of the BERT Model and how it was used.

Interestingly, it was discovered by the researcher that the BERT model was designed to predict the sentiment of product reviews giving them a range of stars between 1 and 5 ranging from 1 star being a Negative review to Neural which is 3 stars to Positive which is 5 stars. Each of these 5 stars corresponds to a column in a 1 row by 5 column matrix (tensor). Each column contains a logit value which is a weighted predicted score for the star class a review is predicted to be in. In the context of this project, the BERT Model takes a social media comment as an input and then it gets a value based on the predicted class a comment might belong to. The logit values are then logistically scaled (normalized) in the sense that that they all should add up to 1 and the class which the highest probability represents the star that it is given. This correlates to the sentiment of that comment. For example, the comment **'I Love you'** has a predicted value (y_pred) of

tensor([[-2.1243, -2.5862, -0.8446, 1.2741, 3.4680]] where each of these commas-separated logit values are in 5 separate columns all adding up to 1. As you can see, the 5th column has the highest value (3.4680) and so the comment is assigned the highest positive sentiment which is 5 stars.

The researcher thought that if the classes could be condensed from 5 stars to three stars: Negative, Neutral and Positive by way of mapping, then it could solve the objective of the assignment. This was applied in the model.

KNN Model using a Pretrained BERT model.

In view of this, the comments were preprocessed into a pandas data frame, the special characters were removed from the comments, and the emojis were turned into words. Additionally, the comments were tokenized and preprocessed for sentiment analysis using the BERT tokenizer. The BERT embeddings for the comments were extracted using the fine-tuned BERT model and importantly, the model applied predicted sentiment labels by mapping numeric labels to sentiment categories (Negative, Neutral, Positive). For the train/ test split feature, the Logits values for each of the comments in the dataset as in the example shown above became the value for X and Y became the predicted sentiment class (Negative, Neutral, Positive) that a comment might be predicted to belong to. Grid Search was applied to tune the KNN hyperparameters and it was discovered that the best value for k for the Instagram Sentiments were 3 whilst the TikTok Sentiments were 9. The social media application with the most Positive sentiments was found to be Instagram and TikTok was found to be the social media application with the most Negative sentiment. The classifications report for both social media applications are shown below on the next page.

Instagram Sentiment Classification Report

The model initially struggled to accurately capture a few of the actual sentiment labels for the training data comments due to the complexities of social media language. To improve this, additional feature engineering was conducted on the comments, and this enhanced the model's performance. Then after, the KNN model successfully captured most of the actual sentiments, which were predominantly positive. However, the negative and neutral sentiments were in the minority classes and their neighborhoods were oversampled by the positive class. This explains why the classification report only showed the positive class. Despite this problem, the test data still correctly classified most data points as positive for the Instagram comments, demonstrating an accuracy score of 100%. This suggests that Instagram is the platform with the most positive sentiments. This speaks to its effectiveness in classifying unforeseen comments.

Best value for k = 3

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Positive	1.00	1.00	1.00	10
Accuracy			1.00	10
Macro Avg	1.00	1.00	1.00	10
Weighted Avg	1.00	1.00	1.00	10

TikTok Sentiment Classification Report

Likewise for the Instagram model, the model initially struggled to accurately capture just a few of the actual sentiment labels for the training data comments due to the complexities of social media comments. To improve this, additional feature engineering was conducted on the comments such as making the comments clearer of the sentiment that they were trying to evoke, and this enhanced the model's performance. Then after, the KNN model successfully captured most of the actual sentiments, which were predominantly Negative as 20 out of the 38 comments in the test data were predicted to be Negative comments. The classification report located to the right shows the breakdown of the values in the classes. The model demonstrated an accuracy of 89% in determining the sentiment classes for the comments in the test data. This speaks to its effectiveness in classifying unforeseen comments.

Best value for k = 9

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Negative	0.94	0.85	0.89	20
Neutral	0.67	0.80	0.73	5
Positive	0.93	1.00	0.96	13
Accuracy			0.89	38
Macro Avg	0.85	0.88	0.86	38
Weighted	0.90	0.89	0.90	38