JSC270: Assignment 4

Report

Mahathi Gandhamaneni

Google Colab notebook

Video Presentation

Student Number: 1007008140

UTORid: gandhama

Problem description and motivation

An area of data science that has always intrigued me is sentiment analysis, and using it to determine how the general public feels about a certain product or business. When I first heard about the viral Pink Sauce, which is created by TikTok creator @chef.pii, I saw a strong negative response on TikTok. However, as I started digging through more articles and tweets, I came across a largely mixed response. I wondered whether there are opinions that I wasn't seeing largely because the TikTok algorithm chose not to show me these. This led me to the question that I intend to solve in this report - "How does the general public on Twitter feel about the viral Pink Sauce?".

My answer to this question involves a straightforward sentiment analysis (involving positive, neutral, and negative sentiments), for which there is a lot of existing literature as it has been done many times for many different brands and companies. Although the method I am using will be similar to that performed in our labs and existing literature, what's different about the analysis is the product/brand I am targeting - Pink Sauce. I have not come across a sentiment analysis of this brand so far. Since it is a viral sensation, there is a lot of data available about it all over the internet, specifically on social media platforms like Twitter. I believe that using Twitter data on Pink Sauce will help answer this question as it is known for providing a rich source of a variety of opinions. I think this question may be difficult to answer because a lot of the tweets may be hard to classify, since a lot of the tweets use Gen-Z language, emojis, and memes in order to express emotions.

Describing the data

Only 1000 tweets were extracted due to time and computational constraints of Google Colab. The exact extraction parameters are as follows: The search word (or phrase rather) used was “pink sauce”. The tweets were extracted after July 17th since Google searches for “pink sauce” peaked on that day (based on complete data). Retweets were filtered out since they would create duplicate observations and thus make the dataset inadequate for training and testing on. Only English language tweets were considered. The most recent tweets were extracted in order to represent ever changing public opinions.

A new feature (‘Sentiment’ - the feature used in this sentiment analysis) based on the tweets was introduced. This was created by labeling the data for sentiment analysis (0, 1, 2 for negative, neutral, and positive respectively), using a method found here. There are 1000 observations and 1000 features (after preprocessing). Only a 1000 features were selected due to computational constraints. There have been many other sentimental analyses done before, but none of them on pink sauce tweets. Hence, I cannot compare my findings to any prior work. Some of the data being used relies on the use of emojis, Gen-Z language, memes, and tone in order to express emotion. This is a limitation of the data since it will be hard to classify and predict on these tweets accurately. Twitter sees 6,000 Tweets every second. This puts into perspective how actively people are using it as a conversation platform. The data we extracted can help us gain valuable customer insights as it represents varying ranges of current public opinion, which is its strength.

Exploratory data analysis

From computing the basic summary statistics of our labeled data, I noticed that the mean of the data is 1.003000, the standard deviation is 0.719381, and the 25th, 50th, and 75th percentiles are 0, 1, and 2 respectively. I created a pie plot and a bar graph in order to better understand the proportion and number of tweets of each sentiment. The data was then preprocessed - it was first tokenized, then URL tokens, punctuation, and special characters were removed. The tokens were all converted to lowercase and lemmatized, and then the first 100 stopwords were removed. Our list of words was then converted to TF-IDF vectors. I decided to use lemmatization and TF-IDF vectorization since they are stronger methods and usually have better accuracy (as exhibited in existing literature [6,7] as well Part I of this assignment).

Describe your machine learning model

Our question revolves around multiclass classification, since we need to assign each tweet to a positive, negative, or neutral sentiment. The model I chose to solve this problem is the Multinomial Naive Bayes model. It is a supervised learning model. I chose to use this Naive Bayes model since it is one of the most popular models for multiclass text classification. Using the Bayes theorem, the model estimates the sentiment of a tweet by assessing the likelihood of each sentiment for a given tweet and returns the sentiment with the highest probability. The strength of the model is that it's very fast at making predictions and easy to implement. Thus, it works well on classifying large data. The weakness of the Naive Bayes model is that it assumes each input variable (word) is independent of each other, and thus cannot learn the relationship between them. This may lead to incorrect predictions in some cases.

We can evaluate the model’s performance using a confusion matrix for each class. We can calculate metrics from the confusion matrix and compare those to that of the baseline model (A DummyClassifier from the scikit-learn library that ignores input features and makes predictions based on the most frequent class label).

Results and Conclusions

- **Accuracy**: The accuracy of the Naive Bayes model is 61.5%, which is higher than that of the baseline (51.5%).
- **Precision (Positive Predicted Value)**: For the Naive Bayes model, we see that out of all the tweets that the model predicted would be negative, 90% were. Similarly for the predicted neutral and positive tweets, only 59% of them were neutral and 79% of them were positive. For the baseline (DummyClassifier), out of all the tweets that the model predicted would be neutral, only 52% were. For the predicted neutral and positive tweets, 0% of them were neutral and positive. This behavior is expected based on the functioning of the DummyClassifier. Thus we can conclude that the Naive Bayes model is more precise than the baseline.
- **Recall (Sensitivity)**: For the Naive Bayes model, we see that out of all tweets that were negative, the model predicted this correctly for 18% of these tweets. Similarly, for all the tweets that were neutral and positive, 100% and 23% of them respectively were predicted correctly. For the baseline, out of all the tweets that were neutral, the model predicted 100% of them correctly. For all tweets that were positive and neutral, 0% of them were predicted correctly, which is in accordance with the behavior of the Dummy Classifier. Overall, the sensitivity of the Naive Bayes model is better than the baseline.

- **F1**: For the Naive Bayes model, the F1 scores for each sentiment are 0.31, 0.74, 0.35 respectively. Based on these scores we can conclude that the model does the best job at predicting whether or not a tweet is neutral, but does a poor job at predicting whether a tweet is positive or not, and neutral or not. For the baseline, the F1 scores for each sentiment are 0, 0.68, 0 respectively. Based on these scores we can conclude that the model does the best job at predicting whether a tweet is neutral or not, but does a very poor job at predicting whether or not a tweet is positive or neutral. Based on these scores, we can conclude that the Naive Bayes model has better F1 scores for each sentiment, and thus does a better job at predicting than the baseline.

Our comparison of the Naive Bayes model to the baseline model on the above metrics tells us that it performs better than the baseline. Thus, we can use the Naive Bayes model to make predictions on Pink Sauce tweets with 61.5% accuracy. Since the accuracy of the model is not very high, it is hard to determine how valuable the customer insights we gain from this sentiment analysis will be. The model parameters used for the Multinomial Naive Bayes model were the default parameters. Laplace smoothing was used to prevent zero probabilities. Prior probabilities were adjusted according to the data as required. If there were no computational constraints, I would employ the use of more data and features, so that I would have more data to train and test the model on, and theoretically, the model would make better predictions.

REFERENCES

1. https://fortune.com/2022/07/22/what-is-pink-sauce-tiktok-production-paused-chef-pii/

2. https://www.researchgate.net/post/What-situation-we-have-to-use-ROC-curve-analysis
3. https://thecleverprogrammer.com/2021/11/24/add-labels-to-a-dataset-for-sentiment-analysis/
4. https://sproutsocial.com/insights/benefits-of-twitter/
5. https://highdemandskills.com/naive-bayes-generative/#:~:text=What%20kind%20of%20classification%20model,of%20generating%20new%20data%20points
6. https://www.researchgate.net/publication/271293133_Stemming_and_Lemmatization_A_Comparison_of_Retrieval_Performances
7. https://link.springer.com/chapter/10.1007/978-3-030-85521-5_19
8. https://www.statology.org/sklearn-classification-report/