A Sentiment Analysis on Brands and Product Emotions

Business Understanding

With the constant improvement and increase in the dynamics of technology, Goku company is looking to stay in the loop of things, while adapting and remaining profitable. For this reason, the company plans to launch a new phone model. To ensure the success of their new product, the company would like to evaluate twitter comments on similar products from different brands to gain valuable insights on what customers think. This project entails building a sentiment analysis model that will evaluate the twitter comments and classify them as either positive, negative or neutral., and then provide actionable recommendations on what decisions the company should make concerning the nature of the phone model it intends to make and release into the market to increase chances of its acceptance and success.

Research Question

The main objective of this project is to build a sentiment analysis model that will establish whether twitter comments on the products are either positive, negative or neutral.

Objectives

To ensure that the model correctly classifies negative comments

To assess the performance of the model and determine possible areas of improvement.

Data Understanding

Data Source

The dataset was obtained from data.world.

Data Description

The dataset used for this project was sourced from CrowdFlower via data.world. The dataset contained over 9,000 tweets that human raters had rated as positive, negative, or neutral towards Apple and Google products.

The dataset was provided in CSV format, containing 3 columns - tweet text, sentiment rating, and tweet ID. The sentiment rating column was the target variable containing three possible values - positive, negative, or neutral.

Data Preparation

Loading the data

All the necessary libraries were imported in this step and the judge-1377884607_tweet_product_company.csv data was loaded using pandas.

Reading and Checking the data

The data was read and then briefly previewed, and checked for missing values and its shape which was (9093,3). Every column was also separately checked and the number of categories in each established.

Cleaning the data

The data cleaning process began with renaming of the columns from their originally too long names to shorter ones. Irrelevant rows (they did not provide any useful information) were deleted. These include rows with no tweet data, followed by tweets that were considered to not be in the English language, and finally rows with “I can’t tell” sentiments.

The data was found to have about 5802 missing values which was quite a large portion of the dataset and therefore dropping them would cause a lot of information to be lost. The Tweet column was analyzed and any mention of words like "Google", "ipad","Apple" etc. were then used to replace the missing values.

The user names on the twitter handles were made anonymous to protect their privacy. Finally, duplicates in the dataset were removed. The dataset was split into training and testing datasets, with an 80:20 split. The training dataset was used to train the NLP model, while the testing dataset was used to evaluate the performance of the model.

Exploratory Data Analysis

The dataset was analyzed using visualizations like word cloud to get a clearer picture of the nature of text and the most common words. This would provide a better understanding of the dataset as per tweet emotion.

Histograms were also plotted to establish the distribution of the three sentiment classes. Here, a massive imbalance was seen between the classes, with the neutral class being way higher than the other two, positive and negative.

Modeling

Pre-processing

Pre-processing is a crucial step in natural language processing (NLP) tasks that involves several steps. Firstly, unnecessary columns were dropped from the dataset for modeling purposes. Next, a corpus was created by collecting and organizing all the textual data. Then, tokenization was performed to break down the text into smaller units, such as words or subwords. Stop words, which are common words that add little meaning, were removed to reduce the dimensionality of the data and improve accuracy and efficiency when building the model.

To further reduce dimensionality in the dataset, lemmatization was carried out on the text data. The target variable (y) was then encoded.

For vectorizing the feature (X), the TfidfVectorizer is a popular tool that converts the text data into numerical representations using TF-IDF weighting, which was used to filter out common words to enhance performance. To deal with class imbalance, Synthetic Minority Over-sampling Technique (SMOTE) was used, which generated synthetic samples of the minority class to balance the dataset.

Model 1 Gaussian Naive Bayes

Gaussian Naive Bayes is an algorithm used for classification tasks. It assumes that features are independent and normally distributed, calculates the probability of a sample belonging to a class based on the probability of each feature value, and selects the class with the highest probability.

The model is a multi-class classifier with three classes: Negative, Neutral, and Positive. The model had an overall accuracy of 0.24, indicating that it correctly classified 24% of the samples in the test dataset.

Model 2 Decision Tree Model

Decision Tree is a machine learning algorithm for classification and regression tasks. It recursively splits data based on features that provide the most information gain or reduce impurity. It creates a tree structure where leaves represent class labels or regression values. The model has an overall accuracy of 0.54, indicating that it correctly classified 54% of the samples in the test dataset.

Model 3 Support Vector Machine

Support Vector Machine (SVM) is a machine learning algorithm for classification and regression tasks. It finds the best hyperplane that separates the data into different classes while maximizing the margin. SVM can handle both linear and non-linearly separable data using kernel functions. It is effective in handling high-dimensional data and class imbalance

As per this model report, the precision for Negative sentiment is 0.35, which means that out of all the samples predicted as Negative, 35% were actually Negative. The recall for Negative sentiment is 0.42. The f1-score for Negative sentiment is 0.38. The accuracy of the model is 0.62, which means that it correctly predicted 62% of the samples.

Model 4 Random Forest Classifier

Random Forest Classifier is an ensemble learning algorithm used for classification tasks in machine learning. It is a collection of decision trees, where each tree is built using a random subset of the features and a random subset of the training samples. The predictions from all the trees are combined to make a final prediction.

The precision for Negative sentiment is 0.33. The recall for Negative sentiment is 0.23, which means that out of all the actual Negative samples, the model was able to correctly predict 23% of them. The f1-score for Negative sentiment is 0.27. The accuracy of the model is 0.64, which means that it correctly predicted 64% of the samples

Conclusion

- Neutral emotions had the highest frequency at 5,500 words, positive at 2,900 and negative at 600 words.
- Apple products had the highest mention followed by google. The least mentions were from the Android app.
- The best performing model in this analysis was the Support Vector Machine and was tuned with C=1000, gamma=0.01 and kernel='rbf'. The SVM model had a recall and an accuracy score of 65% which is a balanced score compared to the other models. The recall score is important as it ensures that the model correctly identifies the true positives. This model ensures that the 65% of negative scores are placed in their right class as the

magnitude of classifying a negative as any other emotion is much greater than any other misclassification.

Even though it may not have the highest level of accuracy, implementing automated Twitter sentiment analysis would represent a positive move towards effectively keeping track of Twitter users' attitudes towards Company Goku's latest mobile phone.

Recommendations

- Company Goku should utilize the model to keep track of the general sentiment towards the mobile phone industry and also to observe the attitudes of people towards competing products.
- Company Goku to utilize Twitter's API to screen and select tweets containing relevant hashtags and text related to their mobile phone. These chosen tweets can then be evaluated by the model to determine their sentiment, providing a means to

monitor and keep up-to-date with the current attitudes of Twitter users towards their product.

- The model would be useful to the company as they can use it to identify users sentiments about thier products and act upon this. They could use the positive tweets to build on their strengths and use the negative ones to identify potential growth areas.
- The company can consider building and incorporating features similar to Apple phones as they have the most positive sentiments among all brands and products.
- Establish a notification system that can keep a check on any alterations in sentiment, allowing for swift action to be taken.
- The company should continuously update and improve the model as new data becomes available to ensure the most accurate and effective analysis possible. This could lead to an improvement of the model in the long run.

Future Improvement Ideas

1. Improve the granularity of emotional analysis by incorporating a more detailed scale. Not all text data will express the same level of negativity or positivity. To address this, using a scale that ranges from very negative to somewhat negative, neutral, somewhat positive, and very positive, can help to identify the subtleties in the sentiment analysis. This approach can enable taking appropriate actions according to the severity of the situation.
2. Broaden the range of the sentiment analysis monitoring by including additional publicly accessible text data sources. There are several sources like public forums or other social media platforms, as well as product reviews, which can provide valuable insights into the overall sentiment towards a product. Though product reviews typically include a rating, the overall sentiment may not always be accurately represented by the rating. A new model is necessary to classify this type of data, as it has a different structure than tweets.
3. Obtain additional labeled Tweets to enhance the model's accuracy. The current dataset utilized for training the model is

comparatively limited, comprising approximately 9000 tweets. Rebuilding the model with a more extensive dataset is expected to boost its efficacy.