NLP-TWITTER SENTIMENT ANALYSIS

DSF-FT2 Phase 4 (Project)

Members

Medrine Waeni
Emmanuel Kipkorir
Betty Bett
Antony Brian
John Mungai
Samson Kamunyu

ISSUED BY:

Moringa School

1. BUSINESS UNDERSTANDING

1.1 Overview

Consumers will give their business to brands they trust, and emotions are at the heart of this trust. When people identify with a brand, it’s easy for them to support the company. Marketers work hard to build and strengthen emotional ties to help consumers feel they have a vested interest in the brand’s success.

There are 486 million users on twitter and a lot of brands have hopped onto the platform to reach these potential customers. With the ever increasing portfolio of products, there is also an increase in the sentiments held by people who have interacted with them. The results of the analysis will assist businesses in determining how consumers feel about their goods and services, providing information on where changes can be made and what can be promoted.

1.2 Problem Statement

Twitter has become a major platform for many organisations worldwide to interact with their clients and other businesses. Major brands such as Google and Apple have a large following on twitter hence this platform offers a direct point of contact to their large pool of users. The composition users range from those who like, hate or are neutral towards a particular product, marketing strategy or proposed changes.

This pool of diverse users helps a brand, such as Google and Apple, identify a consumer’s sentiment towards its products and services thus leading to insights on where to bring improvements and what to promote.

These major brands would like to get the attitudes of their customers to their products which are available in the market.

1.3 Business Objective

To accurately classify the polarity of tweets to predict customers' satisfaction when using either Apple or Google products.

1.4 Specific Objective

- To identify the relationship between tweet sentiment and brands.
- To identify how brands are associated with emotions.

1.5 Business Success Criteria

The project will be deemed successful if:

- The brand is matched to its tweet sentiments
- the consumer can identify which brand they should adopt depending on its sentiments

1.6 Assessing the Situation

The dataset for this project contains tweets expressing sentiments on two main brands: Apple and Google. The two brands offer various devices and services. For instance ipads, iphones and various applications from either brand. The data contains numerous missing values hence the approach of handling them will greatly influence the success of the model. To begin with, we shall use local jupyter notebooks and sync on a git repository.

1.7 Determining Project Goals

The goals of the project are to determine:

- Which features in each brand provide customer satisfaction
- Which brand has the highest positive review

1.8 Determining Project Success Criteria

Tentatively, the study will be judged a success if:

- The final model achieves an accuracy, precision and recall of 80% or more in classifying tweets as either positive, neutral or negative. How much confidence?
- The project can be finished within less than a week time for deployment so as to enable clients access to the online system.
- The final deployed model is able to guide the marketing team on how to improve their products and marketing strategy.
- The final model is able to guide the clients on which products to choose from.

1.9 Project Plan

Duration: The project has a duration of 2 days:

Day 1: Business Understanding, Data Understanding and Data Preparation

Day 2: Modeling, Evaluation and Deployment

Resources: There is enough available data for our analysis.

Personnels: The project has 6 young energetic dedicated personnel ready for the task.

2. DATA UNDERSTANDING

2.1 Overview

The data we have comprises tweets that indicate the attitudes or opinions of individuals who consume either Apple or Google products. The data contains three columns, which include the tweets, the brand and the sentiments of the tweet.

The tweets expressed positive, negative or no emotion towards the interactions.

The data was collected from CrowdFlower via data.world

2.2 Exploring data

The data set contains 36574 entries and 3 columns.

The sentiment column is our target variable with three categories: Negative, positive and neutral.

There is a lot of missing data in the sentiment column which is our target variable.

2.3 Describing data

Column	Description
Tweet	These are the tweets made by both the consumers of Apple and Google products and services.
Brand	The type of product manufactured by a particular company, either Google or Apple.
Sentiment	An opinion based on a consumers interaction with a product or service provided by the brand ie. a positive emotion, negative, or a neutral emotion.

3. Data Preparation

The objective of this step is to select data and clean noise that are less relevant to find the sentiment of tweets such as punctuation, special characters, numbers, and terms which don't carry much weightage in context to the text. We shall look at the following:

- **Validity:** We checked for duplicates and found 22 duplicates.We dealt with these duplicates by dropping them.

- **Consistency:** We verified that the values of various columns are consistent.
- **Completeness:** We checked for completeness of the data and found that there are two columns with missing data. The tweet column had only one missing value, and we dealt with it by dropping that row. The sentiment column had 5802 missing values and since that is more than half of our data, we kept it.
- **Uniformity:** The columns had very long names, so we renamed them, giving each column short and precise names. We them moved on to removing all the capitalization, punctuation and stop words within our data

4. Modeling

4.1 Baseline Model

The models considered are :

1. MultinomialNB which is our base model
2. AdaBoost
3. GradientBoosting
4. XGBoost

The following is the summary of our models:

	Basemodel	XGBoost	GradientBoosting	AdaBoost
Train F1 score	0.722738	0.924575	0.780253	0.663089
Test F1 score	0.589345	0.689042	0.681812	0.646121
Train accuracy	0.732226	0.924722	0.782920	0.670823
Test accuracy	0.610579	0.690834	0.686457	0.654264

4.2 Findings

The model XGBoost and GradientBoosting models performed very well on the training data but were relatively poor on the test data showing signs of overfiting.

The AdaBoost model though having a relatively low accuracy than the GradientBoosting and XGBoost models it generalized well on both training and test data hence was chosen as the final model for deployment.

On deployment using this AdaBoost model, the system did a good job in classifying semantics from a group of tweets and can be relied upon by our clients to give an accurate classification of tweets.

The link to the deployed site can be found here.

4.3 Recommendations

More balances data is needed to increase the accuracy of the model since the data used here was imbalanced and small for a Natural language processing Classifier to give a good classification.

The model accuracy can also be improved by using a neural network which also will require more data to improve the model accuracy.