Tweet Sentiment Analysis - Course 5, Task 4

Alert! Analytics

Primary Objective:

- Evaluate whether the contents of a tweet can be classified into a sentiment level.

Business Question:

- Can models to be built to predict sentiment of a tweet, for potential use by Helio in future commercial programs.

Data Set

For this analysis and modeling we are provided a data set with 1.6 million tweets. Variables contained within the data set include the tweet ID, date of tweet, query, username, tweet text, and the manually assigned sentiment "label" (for training purposes). The sentiment labels have been assigned as follows:

- 0 = Negative sentiment
- 2 = Neutral sentiment
- 4 = Positive sentiment

The modelling is performed in Microsoft Azure and Databricks due to the large size of the data set.

Data Preparation

The data was extremely tidy, with no missing rows or duplicate items. Before further processing, data type conversion was necessary, as the Databricks import did not assign the data types well.

- Label: Assigned to integer (0,2,4)
- Text: Assigned to string (str) for string manipulation functions
- Date: Assigned to datetime to enable sorting by day/hour/minute

Exploratory Data Analysis

Despite having three categories for sentiment, all of the observations within the provided data set had sentiments of either 0 or 4. Between those two, however, the training set is balanced with exactly 800,000 negative sentiment (0) tweets and 800,000 positive sentiment (4) tweets as shown in Figure 1.

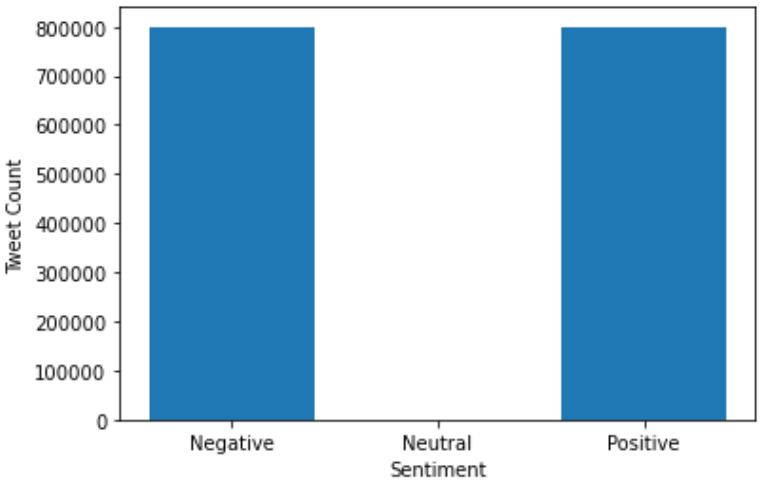


Figure 1: Data set sentiment composition

Business Questions

Is there any time dependency to sentiment?

The complete data set was aggregated by hour to investigate if

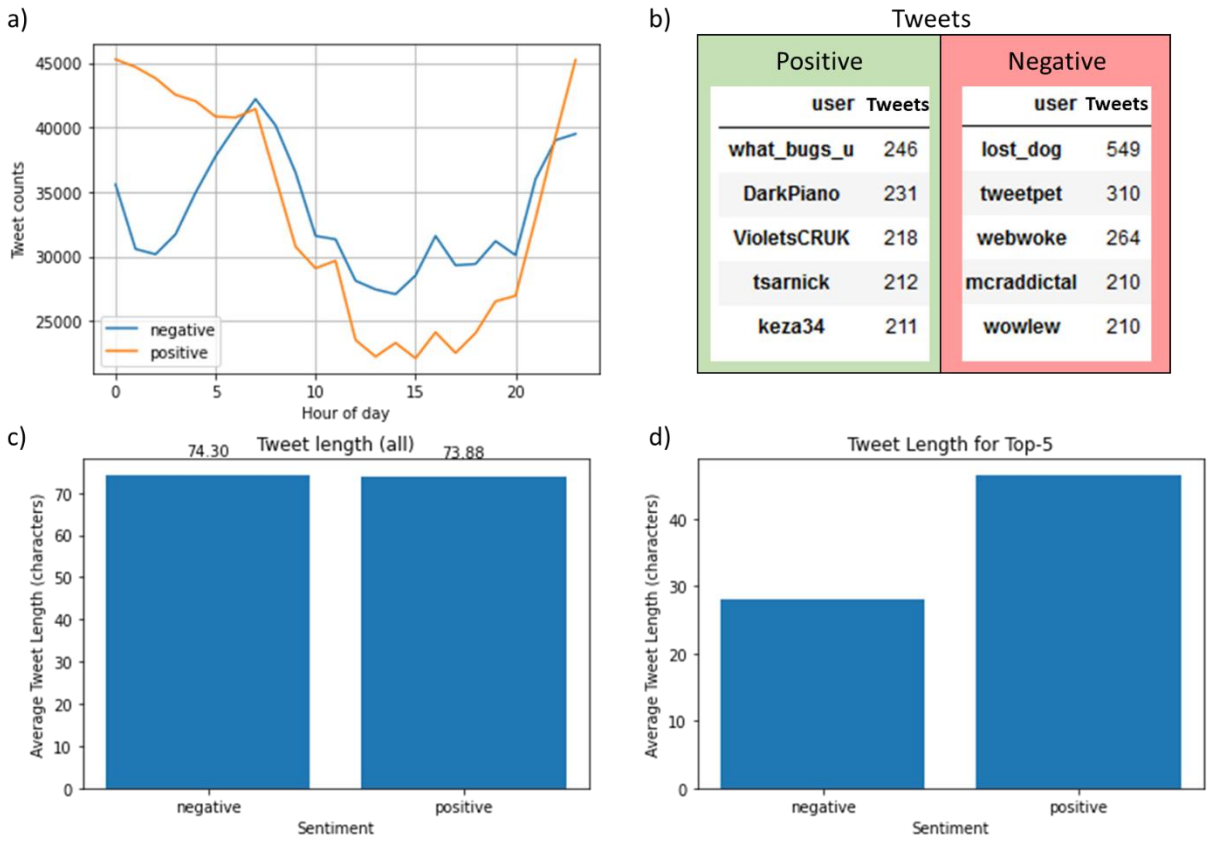


Figure 2: a) Tweet count versus hour of the day, b) Top-5 users for each sentiment, c) average length of all tweets, based on sentiment, d) Tweet length for the 5 most common positive and negative tweets,

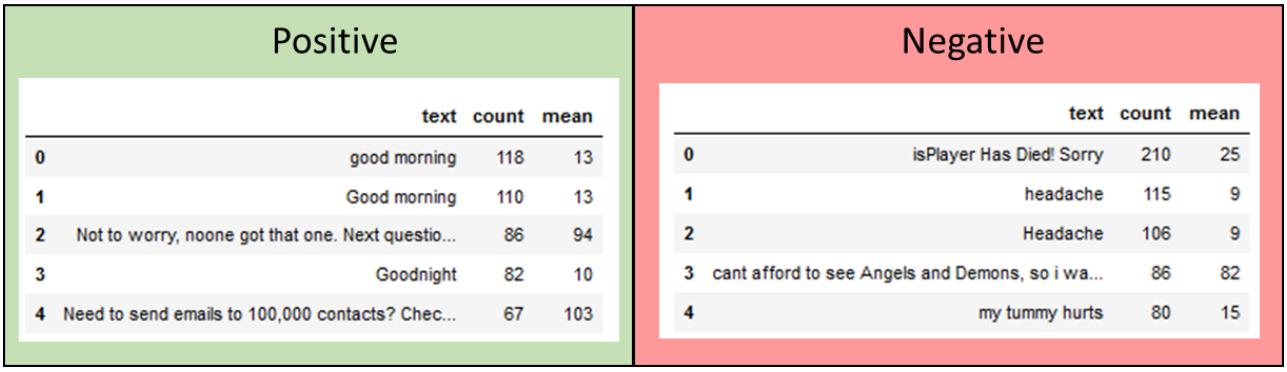
sentiment bias changed throughout the day. As shown in Figure 2a, for most of the day the relative frequency of positive and negative tweets tracked well with one-another, with highest usage outside of normal working hours (hours 8-17). However, the analysis showed a surprising result, with an overwhelming bias towards positive sentiment tweets during the midnight to 5AM range (hours 0-5). Negative tweets were also slightly more prevalent during the daylight hours 8AM-8PM.

Who are the top-5 users for each level of sentiment?

Figure 2b illustrates the users posting the highest number of positive sentiment tweets, as well as the users posting the highest number of negative sentiment tweets. Within the negative tweets, user "lost_dog" is prolific. This data set spans only 71 days, meaning that "lost_dog" averaged 7.7 negative sentiment tweets *per day*.

What is the average length of the top-5 tweets for each level of sentiment?

As shown in Figure 2c, across the entire data set there was no significant difference in length between positive sentiment tweets and negative sentiment tweets. However, upon investigating the 5 most common tweets of each sentiment type, we see a clear difference. As shown in Figure 2d, the top-5 negative tweets were on average much shorter in length than the top-5 positive tweets. The content of the top-5 tweets from each sentiment are shown in Figure 3, which illustrate there is a large amount of variation in the individual tweet lengths. Unfortunately for Twitter, the 5th most frequent "positive" tweet appears to be advertising spam.



Positive

	text	count	mean
0	good morning	118	13
1	Good morning	110	13
2	Not to worry, noone got that one. Next questio...	86	94
3	Goodnight	82	10
4	Need to send emails to 100,000 contacts? Chec...	67	103

Negative

	text	count	mean
0	isPlayer Has Died! Sorry	210	25
1	headache	115	9
2	Headache	106	9
3	cant afford to see Angels and Demons, so i wa...	86	82
4	my tummy hurts	80	15

Figure 3: Top 5 most common tweets for each sentiment

Modelling

Three algorithms were investigated to build a text-based sentiment model. Support Vector Machines (SVM) and Naïve Bayes (NB) are both well suited for textual data. Meanwhile, Decision Trees (DT) are highly effective in the training set is sufficiently large (as it is in this case). In evaluating the various models, this technical case does not have importance bias – that is, a false negative prediction is no worse to our business application than a false positive. Also, since our training data is well balanced between positive and negative sentiment observations (Figure 1), F1 follows accuracy very closely. As a result, for this case accuracy is the main metric for model evaluation.

SVM				
	Precision	Recall	F1-score	Accuracy
Negative	0.79	0.79	0.79	0.79
Positive	0.79	0.79	0.79	
Decision Tree				
	Precision	Recall	F1-score	Accuracy
Negative	0.72	0.72	0.72	0.72
Positive	0.72	0.71	0.72	
Naïve Bayes				
	Precision	Recall	F1-score	Accuracy
Negative	0.82	0.75	0.79	0.78
Positive	0.73	0.8	0.76	

Figure 4: Model generation results

As shown in Figure 4, SVM achieves the greatest accuracy. The individual Precision, Recall, and F1-score metrics also demonstrate consistent performance across all metrics. Naïve Bayes shows very similar performance to SVM, with certain situations and metrics showing improved performance and others showing degraded performance. Depending on the specific needs of follow-on projects using tweet sentiment analysis, Naïve Bayes could be a more attractive option. Further tuning of these models could potentially yield improved accuracy results.

Business Implications

By being able to accurately classify tweet content sentiment, this technology could be used to help Helio use tweets to further compare sentiments of the handset models investigated previously. By filtering the tweets for those containing the handset name ("iPhone" or "Galaxy"), the sentiment of those tweets could be predicted using the same model.

Lessons Learned

Microsoft Azure and Databricks are powerful cloud compute frameworks and are well suited to resource-intensive data science problems. Beyond data sets such as this one with a high number of

observations, Databricks would be useful for data sets with large numbers of variables or in situations where a large number of machine learning algorithms are to be evaluated. Between the models selected, SVM scales much worse with the data size, as the compute times were orders of magnitude longer than Naïve Bayes and Decision Trees. Even with the power of Azure and Databricks, SVM took excessively long to model on a standard cluster.

The internal graphing and dashboarding abilities of Databricks are useful for rapid exploratory data analysis, and can be much faster to set up than graphing with pandas or matplotlib. However, the customizability of the plots is more limited, so are more useful for internal usage and understanding instead of customer-facing presentations.

When comparing Amazon Web Services (AWS) to the Microsoft Azure framework. AWS was first launched in 2002, while Azure was not introduced until 2010, giving AWS a significant head start to establish itself as the leader in the cloud compute landscape. However, Azure quickly caught up and now holds 29% of the cloud market to Amazon's 41%. AWS also has a reputation of being trusted by high-profile customers such as Netflix, Facebook, LinkedIn and many more.

In a technical comparison, both support hybrid cloud, but Azure executes the hybrid cloud much more efficiently. Azure is also considered to have greater awareness of enterprise needs. Not surprisingly, Azure gives high levels of support for Microsoft legacy apps and has support for mixed Linux/Windows environments. AWS, on the other hand, excels in having more data centers, improving availability and offering lower latency.

Both platforms are criticized for their poor customer support, and AWS in particular makes the situation worse by having too many choices and products available – resulting in overcomplication of new deployments.