Nowadays, social media plays an essential role in people's daily life. People are willing to share their lives and opinions on social media. Sometimes, their opinions might impact a lot in some areas. Therefore, as a global social media network company--Bacefook International, they want to develop an automatic system for tracking the opinions of their users on "bweets". For this automatic system, the direct stakeholder will be Bacefook company and people using the "bweets". This is because the Bacefook company can retrieve the information from "bweets" that user post and use machine learning techniques to learn each individual's preference and apply it in the appropriate field, such as push relevant notifications to users (like TikTok). Meanwhile, users can also be the stakeholder since they can review others' posts, which indicates that others' positive and negative opinions might influence users' decisions. In addition, there are also some indirect stakeholders, such as product companies, restaurants, hotels, etc. The reason these companies are indirect stakeholders is that users might share their opinions on "bweets" for those companies, which might impact the performance of these companies. In this particular project, we will review to train up the text classification tool for sentiment analysis. Therefore, in this project, the stakeholder will be hotel customers and hotels. The review of the hotels will impact the decision of other customers whether they want to live in this hotel or not. Furthermore, the review of the hotels will also influence hotels' reputation and might affect hotels' performance.

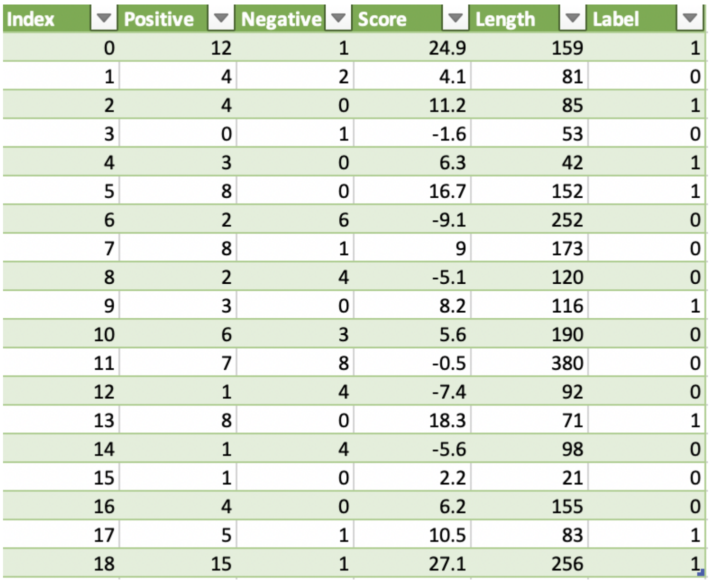
To better develop the automatic system for tracking the opinions of Bacefook users, I’ve used the small corpus of short hotel reviews as the dataset to train the model. The dataset contains 170 hotel customers reviews, and each data has its id, review context, and label. The label is described as 0 or 1 in this dataset, in which 0 represents negative reviews (customers complain about the hotel) and 1 illustrates the positive review (customers satisfied about the hotel). Two models are being trained on for this project: Naive Bayes and Logistic Regression.

By training these two models, especially the Logistic Regression, the system can better understand the positive and negative sentiment in human language. In this case, it will be helpful when we apply it to Bacefook International’s data to retrieve users’ opinions.

To improve the result, I used Logistic Regression and added four features to it. Vader sentiment dictionary is used in this model to help us allocate features. The four features that I selected are:

- Positive lexicon words(count).
- Negative lexicon words(count).
- The score of each word in the sentence(number).
- The length of the sentence(number).

The result shows that the precision for this model by using these features is 0.875. As we can see from the table below, the sentences with higher positive lexicon words numbers tend to be labeled as positive reviews, and higher negative lexicon words numbers tend to be labeled as negative reviews. Besides this, the sentence score also plays an essential role in predicting. Sentences labeled as positive reviews tend to have higher and positive scores, and negative reviews tend to have lower and negative scores.



Index	Positive	Negative	Score	Length	Label
0	12	1	24.9	159	1
1	4	2	4.1	81	0
2	4	0	11.2	85	1
3	0	1	-1.6	53	0
4	3	0	6.3	42	1
5	8	0	16.7	152	1
6	2	6	-9.1	252	0
7	8	1	9	173	0
8	2	4	-5.1	120	0
9	3	0	8.2	116	1
10	6	3	5.6	190	0
11	7	8	-0.5	380	0
12	1	4	-7.4	92	0
13	8	0	18.3	71	1
14	1	4	-5.6	98	0
15	1	0	2.2	21	0
16	4	0	6.2	155	0
17	5	1	10.5	83	1
18	15	1	27.1	256	1

Although the model performance is great on this dataset, we still need larger datasets to train our model and improve it. The dataset, such as Amazon products review, can be applied in our model. As we know, amazon customers will share their own experiences and opinions on the products they purchased and give the product a review from one star to five stars. Therefore, we can use this dataset in our model to predict product popularity based on the customers' opinions. Because of the high quantity of datasets from Amazon, our model can be learned with sufficient data, and it might help improve the precision.

However, some biases may be learned by the model from using this dataset. For example, our improved model is logistic regression, which can not be used for classification tasks with more than two class labels by default. We need to revise the model to Multinomial Logistic Regression to apply in this data set. Furthermore, some of the negative reviews for products might not be the product itself problem. For instance, the product might be damaged during the shipment or haven’t been picked up by customers for a long time. These all might cause a negative review of the product.