Task 2

The task of Natural Language Processing is possible to complete using NLTK library, which is designed specifically to this kind of projects.
The general approach is to extract the most important words and assign values to them that will reflect their importance. Having the matrix of values assigned to every word, we can implent the data in machine learning model.

The course of Natural Language Processing in this task:

1. Removing punctuation
2. Removing stopwords
3. Indentifying and counting the unique words
4. Counting the Tf – idf values
5. Choosing and implenting the machine learning model

- *TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).*
- *IDF(t) = log_e(Total number of documents / Number of documents with term t in it).*

I decided to try 2 different models often used in binary classification cases with :
The MultinomialNB and LogisticRegression.
Those models are frequently used and proved to be well suited to this kind of features and classification tasks.

- Bayesian classification, is finding the probability of a label having some observed features.
- Logistic Regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable.

The Logistic Regresion model had overall better metrics therefore should be considered as the prefered one in this particular case.