Sirut Buasai
CS 525
Prof. Xiaozhong Liu
Assignment 2

Task 1, 2, 3, 4
Implementation of each method is in the attached file `sbuasai2_hw2_ipynb`.

Task 5
1. Result Analysis

Method	Precision	Recall	Accuracy	F1
TF-IDF - Logistic Regression	0.875	0.868	0.872	0.872
TF-IDF - Random Forest	0.891	0.890	0.891	0.890
TF-IDF - Support Vector Machine	0.884	0.881	0.878	0.877
Word2Vec - Logistic Regression	0.821	0.802	0.813	0.811
Word2Vec - Random Forest	0.851	0.858	0.854	0.854
Word2Vec - Support Vector Machine	0.820	0.821	0.813	0.811
BERT w/o fine tune	0.890	0.709	0.811	0.789
BERT fine tune	0.947	0.902	0.925	0.923

2. Result Interpretation
    a. According to the table above, the BERT model with fine-tuning outperformed all other models. This may be attributed to the efficacy of the BERT model to embed different words based on the contextual data from the documents. As such, when fine-tuned to the sentiment analysis task, the model was able to outperform others.

    Although I would have expected the vanilla BERT model (w/o fine-tuning) to outperform TF-IDF and Word2Vec, the results suggest that it did as well as the TF-IDF feature set while outperforming the Word2Vec feature set. One reason for the similar performance between TF-IDF and the vanilla BERT is that given the context of the dataset being sentiment analysis of amazon products, positive and negative reviews could be distinguishable by the significance of certain positive words – such as “like”, “great”, and “good” – and negative words – such as “dislike”, “bad”, “horrible”. As such, TF-IDF is able to map these words across documents to create a separable decision boundary as well as the BERT model. On the other hand, it is expected that BERT would outperform Word2Vec due to the advantage of BERT being able to create different vectors for the same word with different meanings depending on the context around the word. Thus, the vanilla BERT would simply be a more advanced and detailed version of the Word2Vec feature set.

Lastly, TF-IDF outperforms the Word2Vec feature set. This may be attributed to TF-IDF's ability to find the significance between positive and negative words in reviews. Although Word2Vec also can establish positive and negative words into their latent space, TF-IDF would yield a much larger feature space given that every word is mapped to its own feature. As such, in this dataset, the TF-IDF approach may yield better result purely from the fact that it has more feature space than Word2Vec.