Jiazheng (Michael)'s Report for Assignment 1:

1. Problem Setup (i)

- **Background:** Developing a classifier to distinguish facts and fakes can minimize the negative impact of false information.
- **Key Questions:**
    - How well can we distinguish genuine facts from fabricated ones using linear classifiers?
    - Which linear classifier offer the best accuracy from our selected range?

2. Dataset Generation and Experimental Procedure (ii)

- **Dataset Generation:** Using GPT-4 to generate 150 facts and 150 fakes about cats. Prompt Example: "Generate a full list of 150 facts about cats. (no omission in the middle) Requirement: 1. Each fact should be one or two sentences long. 2. At least 30% facts are two sentences long. 3. One fact per line."
- **Common Data Preprocessing:**
    - Remove punctuation: Using Python String translate() and maketrans() functions to remove the punctuations () from the corpus.
    - Remove stop words: Using List Comprehension for looping through the tokenized words in the corpus to remove the stop words.
    - Lemmatization: Utilizing lemmatize() from WordNetLemmatizer class to convert the tokenized words to their initial lemma.
- **Special Data Preprocessing (Feature extraction using N-gram):** Utilizing CountVectorizer() from scikit-learn library to convert the corpus to the matrix with count of tokens as items. Setting up parameters such as (1,1), (2,2) and (1,2) to extract unigram, bigram and the combination of unigram and bigram.
- **Dataset Split:** Randomly selecting 80 percent data for training and remaining 20 percent data for testing.

3. Range of Parameter Settings Tried (iii)

- **Classifier Choice:** Utilizing Naive Bayes classifier, Logistic Regression classifier and SVM classifier from scikit-learn library. For each classifier, preprocessing data with three different settings for N-grams.
- **Hyperparameter Tuning:** Using 5 folds cross validation to choose the best parameter providing highest accuracy. The range of alpha in Naive Bayes classifier is [0.01, 0.1, 0.5, 1.0, 2.0, 5.0, 10.0] (Larger alpha helps the model handle the unseen data better), the range of C for l2 penalty in Logistic Regression classifier is [0.001, 0.01, 0.1, 1, 10, 100] (Smaller C helps the model

handle the unseen data better) and the range of C in SVM classifier is [0.1, 1, 10] (Smaller C helps the model handle the unseen data better).

4. Results and Conclusions (iv)

- **Results:**
  - 1-gram: Naive Bayes Accuracy: 0.78 (alpha=0.1) Logistic Regression Accuracy: 0.78 (C=1) SVM Accuracy: 0.8 (C=1)
  - 2-gram: Naive Bayes Accuracy: 0.65 (alpha=0.01) Logistic Regression Accuracy: 0.63 (C=1) SVM Accuracy: 0.63 (C=1)
  - 1-gram and 2-gram: Naive Bayes Accuracy: 0.8 (alpha=10) Logistic Regression Accuracy: 0.78 (C=1) SVM Accuracy: 0.77 (C=0.1)
- **Conclusions:**
  - With only 300 data points, linear classifiers achieve an accuracy of approximately 75 percent. For every 100 fakes or facts, they can correctly classify 25 more items than a random binary classifier would.
  - Both the Naive Bayes classifier (both 1-grams and 2-grams as features) and the SVM classifier (only 1-grams as features) offer the best accuracy, correctly classifying 80 percent of the data points in the test dataset.
  - All three models perform poorest on the dataset that contains only bigrams. One possible reason is that the majority of the data in my corpus consists of fewer than 10 words, so the count of distinct bigrams is relatively low. If we had more data and longer texts, the classifiers would likely perform better.

5. Limitations of Your Study (v)

- **Generalization Concerns:**
  - An accuracy of 80 percent is insufficient, especially when false information, such as in the news, appears more similar to authentic content than distinctions between fakes and facts about cats.
  - Although both the Naive Bayes and SVM classifiers deliver the same accuracy, Naive Bayes should be the preferred choice with more data points, given its relatively lower computational cost for training.
- **Assumptions:** In the Naive Bayes classifier, we assume that the occurrence of each word is independent of the occurrence of any other word, which is unreasonable in the fakes and facts. For instance, in the facts dataset, the words “cat” and “purr” are highly dependent, as most other animals do not purr.

Declaration

I acknowledge the use of GPT-4 to help me recall the function usage of NLTK and scikit-learn, generate the keywords of outline for this report at the initial stage and fix the grammar of the sentences I wrote at the final stage.