## Dependencies

### Install

```
!pip install contractions
!pip install ipython-autotime
!pip install fastparquet
```

```
Requirement already satisfied: contractions in /usr/local/lib/python3.10/dist-packages (0.1.73)
Requirement already satisfied: textsearch>=0.0.21 in /usr/local/lib/python3.10/dist-packages (from contractions) (0.0.24)
Requirement already satisfied: anyascii in /usr/local/lib/python3.10/dist-packages (from textsearch>=0.0.21->contractions) (0.3.2)
Requirement already satisfied: pyahocorasick in /usr/local/lib/python3.10/dist-packages (from textsearch>=0.0.21->contractions) (2.0.0)
Requirement already satisfied: ipython-autotime in /usr/local/lib/python3.10/dist-packages (0.3.1)
Requirement already satisfied: ipython in /usr/local/lib/python3.10/dist-packages (from ipython-autotime) (7.34.0)
Requirement already satisfied: setuptools>=18.5 in /usr/local/lib/python3.10/dist-packages (from ipython->ipython-autotime) (67.7.2)
Requirement already satisfied: jedi>=0.16 in /usr/local/lib/python3.10/dist-packages (from ipython->ipython-autotime) (0.19.1)
Requirement already satisfied: decorator in /usr/local/lib/python3.10/dist-packages (from ipython->ipython-autotime) (4.4.2)
Requirement already satisfied: pickleshare in /usr/local/lib/python3.10/dist-packages (from ipython->ipython-autotime) (0.7.5)
Requirement already satisfied: traitlets>=4.2 in /usr/local/lib/python3.10/dist-packages (from ipython->ipython-autotime) (5.7.1)
Requirement already satisfied: prompt-toolkit!=3.0.0,!=3.0.1,<3.1.0,>=2.0.0 in /usr/local/lib/python3.10/dist-packages (from ipython->ipython-autotime) (3.0.39)
Requirement already satisfied: pygments in /usr/local/lib/python3.10/dist-packages (from ipython->ipython-autotime) (2.16.1)
Requirement already satisfied: backcall in /usr/local/lib/python3.10/dist-packages (from ipython->ipython-autotime) (0.2.0)
Requirement already satisfied: matplotlib-inline in /usr/local/lib/python3.10/dist-packages (from ipython->ipython-autotime) (0.1.6)
Requirement already satisfied: pexpect>4.3 in /usr/local/lib/python3.10/dist-packages (from ipython->ipython-autotime) (4.8.0)
Requirement already satisfied: parso<0.9.0,>=0.8.3 in /usr/local/lib/python3.10/dist-packages (from jedi>=0.16->ipython->ipython-autotime) (0.8.3)
Requirement already satisfied: ptyprocess>=0.5 in /usr/local/lib/python3.10/dist-packages (from pexpect>4.3->ipython->ipython-autotime) (0.7.0)
Requirement already satisfied: wcwidth in /usr/local/lib/python3.10/dist-packages (from prompt-toolkit!=3.0.0,!=3.0.1,<3.1.0,>=2.0.0->ipython->ipython-autotime) (0.2.8)
Requirement already satisfied: fastparquet in /usr/local/lib/python3.10/dist-packages (2023.8.0)
Requirement already satisfied: pandas>=1.5.0 in /usr/local/lib/python3.10/dist-packages (from fastparquet) (1.5.3)
Requirement already satisfied: numpy>=1.20.3 in /usr/local/lib/python3.10/dist-packages (from fastparquet) (1.23.5)
Requirement already satisfied: cramjam>=2.3 in /usr/local/lib/python3.10/dist-packages (from fastparquet) (2.7.0)
Requirement already satisfied: fsspec in /usr/local/lib/python3.10/dist-packages (from fastparquet) (2023.6.0)
Requirement already satisfied: packaging in /usr/local/lib/python3.10/dist-packages (from fastparquet) (23.2)
Requirement already satisfied: python-dateutil>=2.8.1 in /usr/local/lib/python3.10/dist-packages (from pandas>=1.5.0->fastparquet) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages (from pandas>=1.5.0->fastparquet) (2023.3.post1)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.8.1->pandas>=1.5.0->fastparquet) (1.16.0)
```

### Imports

```
import os
import re
import shutil
import unicodedata
import multiprocessing

import warnings
warnings.filterwarnings("ignore")

import numpy as np
import pandas as pd
import requests

import nltk
from nltk.corpus import stopwords, wordnet
from nltk.stem import WordNetLemmatizer
from nltk.tokenize import word_tokenize

nltk.download('punkt', quiet=True)
nltk.download('wordnet', quiet=True)
nltk.download('stopwords', quiet=True)
nltk.download('averaged_perceptron_tagger', quiet=True)

import contractions

import gensim
import gensim.downloader as api
from gensim.models import Word2Vec

from sklearn.model_selection import train_test_split
from sklearn.metrics import precision_score, recall_score, f1_score, accuracy_score

from sklearn.linear_model import Perceptron
from sklearn.svm import LinearSVC

import torch
import torch.nn as nn
import torch.optim as optim
from torch.utils.data.sampler import RandomSampler, BatchSampler
from torch.utils.data import Dataset, DataLoader

from tqdm.notebook import tqdm

%load_ext autotime
```

```
time: 467 µs (started: 2023-10-19 22:59:24 +00:00)
```

### Config

Set up important configuration parameters and file paths for the project, making it easy to manage various settings and paths from one centralized location

```
os.chdir("/content/drive/MyDrive/Colab Notebooks/CSCI544/HW3")
os.environ['CUDA_LAUNCH_BLOCKING'] = "1"

CURRENT_DIR = os.getcwd()


class DatasetConfig:
    RANDOM_STATE = 34
    TEST_SPLIT = 0.2
    N_SAMPLES_EACH_CLASS = 50000
    DATA_PATH = os.path.join(
        CURRENT_DIR, "amazon_reviews_us_Office_Products_v1_00.tsv.gz"
    )
    PROCESSED_DATA_PATH = os.path.join(
        CURRENT_DIR, "amazon_review_processed_sentiment_analysis.parquet"
    )
    PREPROCESSED_DATA_PATH = os.path.join(
        CURRENT_DIR, "amazon_review_preprocessed_sentiment_analysis.parquet"
    )
```

```
    BUILD_NEW = True
    if os.path.exists(PROCESSED_DATA_PATH) and os.path.exists(PREPROCESSED_DATA_PATH):
        BUILD_NEW = False


class Word2VecConfig:
    PRETRAINED_MODEL = "word2vec-google-news-300"
    PRETRAINED_DEFAULT_SAVE_PATH = os.path.join(
        gensim.downloader.BASE_DIR, PRETRAINED_MODEL, f"{PRETRAINED_MODEL}.gz"
    )
    PRETRAINED_MODEL_SAVE_PATH = os.path.join(
        CURRENT_DIR, PRETRAINED_MODEL, f"{PRETRAINED_MODEL}.gz"
    )
    WINDOW_SIZE = 13
    MAX_LENGTH = 300
    MIN_WORD_COUNT = 9
    CUSTOM_MODEL_PATH = os.path.join(CURRENT_DIR, "word2vec-custom.model")
```

```
time: 15.6 ms (started: 2023-10-19 22:59:24 +00:00)
```

# Helper Functions

## Download & Save Pretrained model

- Run the `api.load()` once and copied the model from temporary path to local drive for fast loading of model in memory.

References:

1. [Faster way to load word2vec model]
2. [Tutorial]

```
def load_pretrained_model():
    if not os.path.exists(Word2VecConfig.PRETRAINED_MODEL_SAVE_PATH):
        # Create a directory if it doesn't exist
        os.makedirs(Word2VecConfig.PRETRAINED_MODEL, exist_ok=True)
        # Download the model embeddings
        pretrained_model = api.load(Word2VecConfig.PRETRAINED_MODEL, return_path=True)
        # Copy & save the embeddings file
        shutil.copyfile(
            Word2VecConfig.PRETRAINED_DEFAULT_SAVE_PATH, Word2VecConfig.PRETRAINED_MODEL
        )
    else:
        pretrained_model = gensim.models.keyedvectors.KeyedVectors.load_word2vec_format(
            Word2VecConfig.PRETRAINED_MODEL_SAVE_PATH, binary=True
        )
    return pretrained_model


# Load the pretrained model
pretrained_model = load_pretrained_model()
```

```
time: 2min 1s (started: 2023-10-19 18:38:29 +00:00)
```

## Accelarator Configuration

```
def get_device():
    if torch.cuda.is_available():
        # Check if GPU is available
        return torch.device("cuda")
    else:
        # Use CPU if no GPU or TPU is available
        return torch.device("cpu")

device = get_device()
device
```

```
device(type='cpu')time: 11.8 ms (started: 2023-10-19 22:59:24 +00:00)
```

## Download Data

Checks if a file specified by `DatasetConfig.DATA_PATH` exists. If not, it downloads the file from a given URL and saves it with the same name. If the file already exists, it prints a message indicating so

```
if not os.path.exists(DatasetConfig.DATA_PATH):
    url = (
        "https://web.archive.org/web/20201127142707if_/https://s3.amazonaws.com/amazon-reviews-pds"
        "/tsv/amazon_reviews_us_Office_Products_v1_00.tsv.gz"
    )
    file_name = DatasetConfig.DATA_PATH

    with requests.get(url, stream=True) as response:
        with open(file_name, "wb") as file:
            for chunk in response.iter_content(chunk_size=8192):
                file.write(chunk)

    print(f"Downloaded file '{os.path.relpath(file_name)}' successfully.")
else:
    print(f"File '{DatasetConfig.DATA_PATH}' already exists.")
```

```
File '/content/drive/MyDrive/Colab Notebooks/CSCI544/HW3/amazon_reviews_us_Office_Products_v1_00.tsv.gz' already exists.
time: 4.21 ms (started: 2023-10-19 22:59:24 +00:00)
```

## Dataset Preparation

This code provides a pipeline for processing and preparing a dataset for sentiment analysis:

1. `LoadData` class loads a dataset from a specified path, keeping only relevant columns.
2. `ProcessData` class performs the following tasks:
    - Converts star ratings to numeric values.
    - Classifies sentiments based on star ratings (1 for negative, 2 for positive).
    - Balances the dataset by sampling an equal number of samples for both sentiments.
3. `CleanText` class defines various text cleaning operations:
    - Removing non-ASCII characters.

    - Expanding contractions.
    - Removing email addresses, URLs, and HTML tags.
    - Lowercasing and stripping spaces.
4. `clean_and_process_data` function executes the entire data processing pipeline:
    - Loads the data.
    - Applies basic processing.
    - Balances the dataset.
    - Cleans the text.
    - Tokenizes the reviews.
5. `preprocess_review_body` function generates word embeddings for each word in a review using a pre-trained Word2Vec model.
6. `get_reviews_dataset` function handles the entire data preprocessing and embedding generation process. It checks if the preprocessed data already exists, and if not, it performs the data preprocessing and saves the preprocessed data in Parquet format.

Overall, this pipeline ensures that the dataset is properly loaded, cleaned, processed, balanced, and transformed into embeddings suitable for sentiment analysis.

> Note:
>
> - Parquet format is efficient for storage.
> - Storing data to avoid running the pipeline and embedding generation process all over again.
> - Provides a ready-to-use dataset for sentiment analysis tasks, allowing for quicker experimentation and model training

## Read and Process

```
class LoadData:
    @staticmethod
    def load_data(path):
        df = pd.read_csv(
            path,
            sep="\t",
            usecols=["review_headline", "review_body", "star_rating"],
            on_bad_lines="skip",
            memory_map=True,
        )
        return df


class ProcessData:
    @staticmethod
    def filter_columns(df):
        return df.loc[:, ["review_body", "star_rating"]]

    @staticmethod
    def convert_star_rating(df):
        df["star_rating"] = pd.to_numeric(df["star_rating"], errors="coerce")
        df.dropna(subset=["star_rating"], inplace=True)
        return df

    @staticmethod
    def classify_sentiment(df):
        df["sentiment"] = df["star_rating"].apply(lambda x: 1 if x <= 3 else 2)
        return df

    @staticmethod
    def sample_data(df, n_samples, random_state):
        sampled_df = pd.concat(
            [
                df.query("sentiment==1").sample(n=n_samples, random_state=random_state),
                df.query("sentiment==2").sample(n=n_samples, random_state=random_state),
            ],
            ignore_index=True,
        ).sample(frac=1, random_state=random_state, ignore_index=True)

        sampled_df.drop(columns=["star_rating"], inplace=True)
        return sampled_df


class CleanText:
    @staticmethod
    def unicode_to_ascii(s):
        return "".join(
            c for c in unicodedata.normalize("NFD", s) if unicodedata.category(c) != "Mn"
        )

    @staticmethod
    def expand_contractions(text):
        """Expand contraction for eg., wouldn't => would not"""
        return contractions.fix(text)

    @staticmethod
    def remove_email_addresses(text):
        return re.sub(r"[a-zA-Z0-9_\-\.]+@[a-zA-Z0-9_\-\.]+\.[a-zA-Z]{2,5}", "", text)

    @staticmethod
    def remove_urls(text):
        return re.sub(r"\bhttps?:\/\/\S+|www\.\S+", "", text)

    @staticmethod
    def remove_html_tags(text):
        return re.sub(r"<.*?>", "", text)

    @staticmethod
    def clean_text(text):
        text = text.lower().strip()
        text = CleanText.unicode_to_ascii(text)
        # text = CleanText.remove_email_addresses(text)
        # text = CleanText.remove_urls(text)
        text = CleanText.remove_html_tags(text)
        text = CleanText.expand_contractions(text)

        # creating a space between a word and the punctuation following it
        # text = re.sub(r"([?.!,¿])", r" \1 ", text)
        # text = re.sub(r'[" "]+', " ", text)

        # removes all non-alphabetical characters
        # text = re.sub(r"[^a-zA-Z\s]+", "", text)

        # remove extra spaces
        # text = re.sub(" +", " ", text)
```

```
        return text


def clean_and_process_data(path):
    df = LoadData.load_data(path)

    # Basic processing
    df_filtered = ProcessData.filter_columns(df)
    df_filtered = ProcessData.convert_star_rating(df_filtered)
    df_filtered = ProcessData.classify_sentiment(df_filtered)

    balanced_df = ProcessData.sample_data(
        df_filtered, DatasetConfig.N_SAMPLES_EACH_CLASS, DatasetConfig.RANDOM_STATE
    )

    # Clean data
    balanced_df.dropna(inplace=True)
    balanced_df["review_body"] = balanced_df["review_body"].astype(str)
    balanced_df["review_body"] = balanced_df["review_body"].apply(CleanText.clean_text)
    # Drop reviews that are empty
    balanced_df = balanced_df.loc[balanced_df["review_body"].str.strip() != ""]

    # Tokenize Reviews
    balanced_df["review_body"] = balanced_df["review_body"].apply(word_tokenize)
    return balanced_df


def preprocess_review_body(text, word2vec_model, topn=None):
    embeddings = [word2vec_model[word] for word in text if word in word2vec_model]

    if topn is not None:
        embeddings = np.concatenate(embeddings[:topn], axis=0)
    else:
        embeddings = np.mean(embeddings, axis=0)
    return embeddings


def get_reviews_dataset(new=False):
    if new or not os.path.exists(DatasetConfig.DATA_PATH):
        balanced_df = clean_and_process_data(DatasetConfig.DATA_PATH)
        balanced_df.to_parquet(DatasetConfig.PROCESSED_DATA_PATH, index=False)

        # Preprocess data and generate word2vec embeddings Avg and top 10
        balanced_df["embeddings"] = balanced_df["review_body"].apply(
            lambda text: preprocess_review_body(text, pretrained_model, topn=None)
        )
        # Drop rows with NaN embeddings
        balanced_df.dropna(subset=["embeddings"], inplace=True)

        balanced_df["embeddings_top_10"] = balanced_df["review_body"].apply(
            lambda text: preprocess_review_body(text, pretrained_model, topn=10)
        )

        balanced_df.to_parquet(DatasetConfig.PREPROCESSED_DATA_PATH, index=False)
    else:
        balanced_df = pd.read_parquet(
            DatasetConfig.PREPROCESSED_DATA_PATH,
            # engine="fastparquet"
        )
    return balanced_df
```

time: 4.81 ms (started: 2023-10-19 22:59:25 +00:00)

```
balanced_df = get_reviews_dataset(
    new=DatasetConfig.BUILD_NEW
)
print("Total Records:", balanced_df.shape)
balanced_df.head(10)
```

Total Records: (99862, 4)

| | review_body | sentiment | embeddings | embeddings_top_10 |
|---|---|---|---|---|
| **0** | [i, set, up, a, photo, booth, at, my, sister, ... | 2 | [0.016994974, 0.024544675, -0.010975713, 0.093... | [-0.22558594, -0.01953125, 0.09082031, 0.23730... |
| **1** | [like, everyone, else, ,, i, like, saving, mon... | 1 | [0.044110615, 0.036876563, 0.0371785, 0.113560... | [0.103515625, 0.13769531, -0.0029754639, 0.181... |
| **2** | [the, pen, is, perfect, what, i, want, !, howe... | 2 | [0.026102701, 0.029064532, 0.010800962, 0.0622... | [0.080078125, 0.10498047, 0.049804688, 0.05346... |
| **3** | [i, think, they, are, too, expensive, for, the... | 1 | [-0.0039075767, 0.032967318, 0.02339106, 0.113... | [-0.22558594, -0.01953125, 0.09082031, 0.23730... |
| **4** | [black, is, working, wonderfully, ,, and, both... | 1 | [0.034285888, 0.013478661, 0.041618653, 0.1132... | [0.10498047, 0.018432617, 0.008972168, -0.0128... |
| **5** | [i, have, problems, with, the, moveable, tab, ... | 1 | [0.010405041, 0.026173819, 0.03433373, 0.09698... | [-0.22558594, -0.01953125, 0.09082031, 0.23730... |
| **6** | [this, printer, sucks, !, it, started, out, wo... | 1 | [0.05581854, 0.035414256, 0.047512088, 0.09278... | [0.109375, 0.140625, -0.03173828, 0.16601562, ... |
| **7** | [the, ink, on, these, cartridges, leak, ., i, ... | 1 | [0.0037488434, 0.053543895, 0.038638465, 0.134... | [0.080078125, 0.10498047, 0.049804688, 0.05346... |
| **8** | [it, gets, points, for, working, as, designed,... | 2 | [0.046220347, 0.029853666, 0.058699824, 0.0745... | [0.084472656, -0.0003528595, 0.053222656, 0.09... |
| **9** | [i, ordered, these, and, they, work, just, fin... | 1 | [-0.0013514927, 0.016482098, 0.031290326, 0.07... | [-0.22558594, -0.01953125, 0.09082031, 0.23730... |

time: 29.1 s (started: 2023-10-19 22:59:25 +00:00)

## Review Body stats

Mean number of words = 66

Median number of words = 37

Limiting sequence length for RNN based embeddings = 45

```
balanced_df["review_body"].apply(len).describe()
```

```
count    99862.000000
mean        65.937384
std        100.170130
min          1.000000
25%         20.000000
50%         37.000000
75%         76.000000
max       4847.000000
Name: review_body, dtype: float64time: 47.1 ms (started: 2023-10-19 23:00:40 +00:00)
```

## Train and Test Spilts

```
train_df, test_df = train_test_split(
    balanced_df,
    test_size=DatasetConfig.TEST_SPLIT,
    random_state=DatasetConfig.RANDOM_STATE,
```

```
    stratify=balanced_df["sentiment"]
)
```

```
time: 50.7 ms (started: 2023-10-19 23:00:43 +00:00)
```

## Word Embedding

### Semantic similarity examples with pretrained embeddings

```
# Example 1: King - Man + Woman = Queen
result = pretrained_model.most_similar(positive=['woman', 'king'], negative=['man'])
print(f"Semantic Similarity: {result[0][0]}")

# Example 2: excellent ~ outstanding
result = pretrained_model.similarity('excellent', 'outstanding')
print(f"Semantic Similarity: {result}")

# Example 3: Paris - France + Italy = Milan
result = pretrained_model.most_similar(positive=['Italy', 'Paris'], negative=['France'])
print(f"Semantic Similarity: {result[0][0]}")

# Example 4: Car - Wheel + Boat = Yacht
result = pretrained_model.most_similar(positive=['Boat', 'Car'], negative=['Wheel'])
print(f"Semantic Similarity: {result[0][0]}")

# Example 5: Delicious ~ Tasty
result = pretrained_model.similarity('Delicious', 'Tasty')
print(f"Semantic Similarity: {result}")

# Example 6: Computer ~ Plant
result = pretrained_model.similarity('Computer', 'Plant')
print(f"Semantic Similarity: {result}")

# Example 7: Cat ~ Dog
result = pretrained_model.similarity('Cat', 'Dog')
print(f"Semantic Similarity: {result}")
```

```
Semantic Similarity: queen
Semantic Similarity: 0.5567485690116882
Semantic Similarity: Milan
Semantic Similarity: Yacht
Semantic Similarity: 0.5718502402305603
Semantic Similarity: 0.04445184767246246
Semantic Similarity: 0.6061107516288757
time: 9.78 s (started: 2023-10-18 21:26:10 +00:00)
```

```
del pretrained_model
```

```
time: 776 µs (started: 2023-10-19 11:13:25 +00:00)
```

### Custom Word2Vec Embeddings Generation

```
sentences=train_df["review_body"].apply(lambda x: x.tolist()).tolist()

# Train Word2Vec model
w2v_model_custom = Word2Vec(
    sentences=sentences,
    vector_size=Word2VecConfig.MAX_LENGTH,
    window=Word2VecConfig.WINDOW_SIZE,
    min_count=Word2VecConfig.MIN_WORD_COUNT,
    workers=multiprocessing.cpu_count()
)

# Save the model
w2v_model_custom.save(Word2VecConfig.CUSTOM_MODEL_PATH)
```

```
time: 1min 30s (started: 2023-10-18 21:36:09 +00:00)
```

### Test Custom Embeddings

```
# Load the custom model
w2v_model_custom = Word2Vec.load(Word2VecConfig.CUSTOM_MODEL_PATH)

# Example 1: King - Man + Woman = Queen
res = w2v_model_custom.wv.most_similar(positive=['woman', 'king'], negative=['man'])
print(f"Semantic Similarity (Custom Model): {res[0]}")

# Example 2: excellent ~ outstanding
res = w2v_model_custom.wv.similarity('excellent', 'outstanding')
print(f"Semantic Similarity (Custom Model): {res}")
```

```
Semantic Similarity (Custom Model): ('queen', 0.5723455548286438)
Semantic Similarity (Custom Model): 0.7957370281219482
time: 241 ms (started: 2023-10-18 21:37:47 +00:00)
```

## Conclusion

**What do you conclude from comparing vectors generated by yourself and the pretrained model? Which of the Word2Vec models seems to encode semantic similarities between words better?**

1. **Custom-trained Word2Vec Model:**
    - **Strengths:**
        - Captures domain-specific relationships and nuances as it trained on very specific dataset.
    - **Weaknesses:**
        - It may not perform as well on tasks outside of its training domain.
        - The quality of embeddings heavily depends on the dataset used for training.
        - For example, if the dataset is small or not representative of the overall language, the embeddings may be less reliable.
2. **Pretrained "word2vec-google-news-300" Model:**
    - **Strengths:**
        - This model has been pretrained on a massive corpus of text from various domains, making it highly versatile and capable of capturing a wide range of semantic relationships.
        - It can generalize well to different tasks and domains.

- **Weaknesses:**
  - While it provides strong generalization, it may not capture domain-specific relationships as effectively as a model trained on domain-specific data.

- The semantic similarity score is higher for the pretrained model compared to the custom model. This indicates that the pretrained model is better at encoding semantic similarities between words.
- The custom Word2Vec model, which was trained on the provided dataset, may not have had access to as diverse and extensive a corpus as the pretrained model. This can lead to limitations in its ability to generalize and capture nuanced semantic relationships.

```
del w2v_model_custom, res, sentences
```

```
time: 319 µs (started: 2023-10-19 01:33:10 +00:00)
```

## Simple Models

```
def evaluate_model(model, X_test, y_test):
    # Predict on the test set
    y_pred = model.predict(X_test)

    # Calculate evaluation metrics
    precision = precision_score(y_test, y_pred, average="binary")
    recall = recall_score(y_test, y_pred, average="binary")
    f1 = f1_score(y_test, y_pred, average="binary")
    accuracy = accuracy_score(y_test, y_pred)

    return precision, recall, f1, accuracy


def train_and_evaluate_model(model_class, X_train, y_train, X_test, y_test, **model_params):
    # Initialize model
    model = model_class(**model_params)

    # Train the model
    model.fit(X_train, y_train)

    # Evaluate model
    precision, recall, f1, accuracy = evaluate_model(model, X_test, y_test)
    return model, precision, recall, f1, accuracy
```

```
time: 1.08 ms (started: 2023-10-19 23:00:53 +00:00)
```

```
X_train = np.vstack(train_df["embeddings"])
y_train = train_df["sentiment"]
X_test = np.vstack(test_df["embeddings"])
y_test = test_df["sentiment"]
```

```
time: 270 ms (started: 2023-10-19 23:00:54 +00:00)
```

## SVM

| Params | Precision | Recall | F1 | Accuracy | Features Used |
|---|---|---|---|---|---|
| LinearSVC(C=0.1, max_iter=10000) | 0.7997 | 0.8671 | 0.8320 | 0.8321 | Word2Vec |
| LinearSVC(max_iter=10000) | 0.8045 | 0.8623 | 0.8324 | 0.8262 | Word2Vec |
| LinearSVC(C=0.01, max_iter=15000) | 0.7836 | 0.8835 | 0.8305 | 0.8281 | Word2Vec |

```
# Train and evaluate LinearSVC model
(
    _,
    precision_svc,
    recall_svc,
    f1_svc,
    acc_svc
) = train_and_evaluate_model(
    LinearSVC,
    X_train, y_train, X_test, y_test,
    max_iter=10000,
    # C=0.1
)

print(f'Precision Recall F1 Accuracy (LinearSVC): {precision_svc:.4f} {recall_svc:.4f} {f1_svc:.4f} {acc_svc:.4f}')
```

```
Precision Recall F1 Accuracy (LinearSVC): 0.8045 0.8623 0.8324 0.8262
time: 28.2 s (started: 2023-10-19 01:31:55 +00:00)
```

## Perceptron

| Params | Precision | Recall | F1 | Accuracy | Features Used |
|---|---|---|---|---|---|
| Perceptron(eta0=0.01, max_iter=5000, penalty='elasticnet', warm_start=True) | 0.7693 | 0.8778 | 0.8200 | 0.8071 | Word2Vec |
| Perceptron(max_iter=5000) | 0.7786 | 0.8613 | 0.8179 | 0.8110 | Word2Vec |
| Perceptron() | 0.7786 | 0.8613 | 0.8179 | 0.8110 | Word2Vec |
| Perceptron(eta0=0.1, max_iter=5000, penalty='elasticnet', warm_start=True) | 0.5977 | 0.9844 | 0.7438 | 0.6655 | Word2Vec |
| Perceptron(eta0=0.001, max_iter=10000, penalty='l2') | 0.7367 | 0.9114 | 0.8148 | 0.7849 | Word2Vec |
| Perceptron(eta0=0.01, max_iter=10000, penalty='l2', warm_start=True) | 0.7653 | 0.8789 | 0.8181 | 0.8002 | Word2Vec |
| Perceptron(eta0=0.01, penalty='l1', warm_start=True) | 0.6133 | 0.9813 | 0.7548 | 0.6832 | Word2Vec |

```
# Train and evaluate Perceptron model using BoW features
(
    _,
    precision_perceptron,
    recall_perceptron,
    f1_perceptron,
    acc_perceptron
) = train_and_evaluate_model(
    Perceptron,
    X_train, y_train, X_test, y_test,
    max_iter=5000,
    eta0=0.01,
    warm_start=True,
    penalty="elasticnet"
)

print(f'Precision Recall F1 (Perceptron): {precision_perceptron:.4f} {recall_perceptron:.4f} {f1_perceptron:.4f} {acc_perceptron:.4f}')
```

```
Precision Recall F1 (Perceptron): 0.7693 0.8778 0.8200 0.8071
time: 1.1 s (started: 2023-10-19 01:32:24 +00:00)
```

## With TFIDF Features

## Homework 1 Script Edited

```
# @title Homework 1 Script Edited

%%writefile HW1-CSCI544-wo-neg-sw.py
# Python Version: 3.10.12

import re
import unicodedata

import warnings

warnings.filterwarnings("ignore")

import numpy as np
import pandas as pd

import nltk
from nltk.corpus import stopwords, wordnet
from nltk.stem import WordNetLemmatizer
from nltk.tokenize import word_tokenize

nltk.download("punkt", quiet=True)
nltk.download("wordnet", quiet=True)
nltk.download("stopwords", quiet=True)
nltk.download("averaged_perceptron_tagger", quiet=True)

import contractions

from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics import precision_score, recall_score, f1_score, accuracy_score

from sklearn.linear_model import Perceptron
from sklearn.svm import LinearSVC


class Config:
    RANDOM_STATE = 56
    DATA_PATH = "amazon_reviews_us_Office_Products_v1_00.tsv.gz"
    TEST_SPLIT = 0.2
    N_SAMPLES_EACH_CLASS = 50000
    NUM_TFIDF_FEATURES = 5000
    NUM_BOW_FEATURES = 5000


class DataLoader:
    @staticmethod
    def load_data(path):
        df = pd.read_csv(
            path,
            sep="\t",
            usecols=["review_headline", "review_body", "star_rating"],
            on_bad_lines="skip",
            memory_map=True,
        )
        return df


class DataProcessor:
    @staticmethod
    def filter_columns(df):
        return df.loc[:, ["review_body", "star_rating"]]

    @staticmethod
    def convert_star_rating(df):
        df["star_rating"] = pd.to_numeric(df["star_rating"], errors="coerce")
        df.dropna(subset=["star_rating"], inplace=True)
        return df

    @staticmethod
    def classify_sentiment(df):
        df["sentiment"] = df["star_rating"].apply(lambda x: 1 if x <= 3 else 2)
        return df

    @staticmethod
    def sample_data(df, n_samples, random_state):
        sampled_df = pd.concat(
            [
                df.query("sentiment==1").sample(n=n_samples, random_state=random_state),
                df.query("sentiment==2").sample(n=n_samples, random_state=random_state),
            ],
            ignore_index=True,
        ).sample(frac=1, random_state=random_state)

        sampled_df.drop(columns=["star_rating"], inplace=True)
        return sampled_df


class TextCleaner:
    @staticmethod
    def unicode_to_ascii(s):
        return "".join(
            c for c in unicodedata.normalize("NFD", s) if unicodedata.category(c) != "Mn"
        )

    @staticmethod
    def expand_contractions(text):
        return contractions.fix(text)

    @staticmethod
    def remove_email_addresses(text):
        return re.sub(r"[a-zA-Z0-9_\-\.]+@[a-zA-Z0-9_\-\.]+\.[a-zA-Z]{2,5}", " ", text)

    @staticmethod
    def remove_urls(text):
        return re.sub(r"\bhttps?:\/\/\S+|www\.\S+", " ", text)

    @staticmethod
    def remove_html_tags(text):
        return re.sub(r"<.*?>", "", text)

    @staticmethod
    def clean_text(text):
```

```
        text = TextCleaner.unicode_to_ascii(text.lower().strip())
        # replacing email addresses with empty string
        text = TextCleaner.remove_email_addresses(text)
        # replacing urls with empty string
        text = TextCleaner.remove_urls(text)
        # Remove HTML tags
        text = TextCleaner.remove_html_tags(text)
        # Expand contraction for eg., wouldn't => would not
        text = TextCleaner.expand_contractions(text)
        # creating a space between a word and the punctuation following it
        text = re.sub(r"([?.!,¿])", r" \1 ", text)
        text = re.sub(r'[" "]+', " ", text)
        # removes all non-alphabetical characters
        text = re.sub(r"[^a-zA-Z\s]+", "", text)
        # remove extra spaces
        text = re.sub(" +", " ", text)
        text = text.strip()
        return text


class TextPreprocessor:
    lemmatizer = WordNetLemmatizer()

    @staticmethod
    def get_stopwords_pattern():
        # Stopword list
        og_stopwords = set(stopwords.words("english"))

        # Define a list of negative words to remove
        neg_words = ["no", "not", "nor", "neither", "none", "never", "nobody", "nowhere"]
        custom_stopwords = [word for word in og_stopwords if word not in neg_words]
        pattern = re.compile(r"\b(" + r"|".join(custom_stopwords) + r")\b\s*")
        return pattern

    @staticmethod
    def pos_tagger(tag):
        if tag.startswith("J"):
            return wordnet.ADJ
        elif tag.startswith("V"):
            return wordnet.VERB
        elif tag.startswith("N"):
            return wordnet.NOUN
        elif tag.startswith("R"):
            return wordnet.ADV
        else:
            return None

    @staticmethod
    def lemmatize_text_using_pos_tags(text):
        words = nltk.pos_tag(word_tokenize(text))
        words = map(lambda x: (x[0], TextPreprocessor.pos_tagger(x[1])), words)
        lemmatized_words = [
            TextPreprocessor.lemmatizer.lemmatize(word, tag) if tag else word for word, tag in words
        ]
        return " ".join(lemmatized_words)

    @staticmethod
    def lemmatize_text(text):
        words = word_tokenize(text)
        lemmatized_words = [TextPreprocessor.lemmatizer.lemmatize(word) for word in words]
        return " ".join(lemmatized_words)

    pattern = get_stopwords_pattern()

    @staticmethod
    def preprocess_text(text):
        # replacing all the stopwords
        text = TextPreprocessor.pattern.sub("", text)
        text = TextPreprocessor.lemmatize_text(text)
        return text


clean_text_vect = np.vectorize(TextCleaner.clean_text)
preprocess_text_vect = np.vectorize(TextPreprocessor.preprocess_text)


def clean_and_process_data(path):
    df = DataLoader.load_data(path)
    df_filtered = DataProcessor.filter_columns(df)
    df_filtered = DataProcessor.convert_star_rating(df_filtered)
    df_filtered = DataProcessor.classify_sentiment(df_filtered)

    balanced_df = DataProcessor.sample_data(
        df_filtered, Config.N_SAMPLES_EACH_CLASS, Config.RANDOM_STATE
    )

    balanced_df["review_body"] = balanced_df["review_body"].astype(str)

    # Clean data
    # avg_len_before_clean = balanced_df["review_body"].apply(len).mean()
    balanced_df["review_body"] = balanced_df["review_body"].apply(clean_text_vect)
    # Drop reviews that are empty
    balanced_df = balanced_df.loc[balanced_df["review_body"].str.strip() != ""]
    # avg_len_after_clean = balanced_df["review_body"].apply(len).mean()

    # Preprocess data
    # avg_len_before_preprocess = avg_len_after_clean
    balanced_df["review_body"] = balanced_df["review_body"].apply(preprocess_text_vect)
    # avg_len_after_preprocess = balanced_df["review_body"].apply(len).mean()

    # Print Results
    # print(f"{avg_len_before_clean:.2f}, {avg_len_after_clean:.2f}")
    # print(f"{avg_len_before_preprocess:.2f}, {avg_len_after_preprocess:.2f}")

    return balanced_df


def evaluate_model(model, X_test, y_test):
    # Predict on the test set
    y_pred = model.predict(X_test)

    # Calculate evaluation metrics
    precision = precision_score(y_test, y_pred, average="binary")
    recall = recall_score(y_test, y_pred, average="binary")
    f1 = f1_score(y_test, y_pred, average="binary")
    accuracy = accuracy_score(y_test, y_pred)

    return precision, recall, f1, accuracy
```

```
def train_and_evaluate_model(model_class, X_train, y_train, X_test, y_test, **model_params):
    # Initialize model
    model = model_class(**model_params)

    # Train the model
    model.fit(X_train, y_train)

    # Evaluate model
    precision, recall, f1, accuracy = evaluate_model(model, X_test, y_test)
    return model, precision, recall, f1, accuracy


def main():
    balanced_df = clean_and_process_data(Config.DATA_PATH)

    # Splitting the reviews dataset
    X_train, X_test, y_train, y_test = train_test_split(
        balanced_df["review_body"],
        balanced_df["sentiment"],
        test_size=Config.TEST_SPLIT,
        random_state=Config.RANDOM_STATE,
    )

    # Feature Extraction
    tfidf_vectorizer = TfidfVectorizer(max_features=Config.NUM_TFIDF_FEATURES)
    X_train_tfidf = tfidf_vectorizer.fit_transform(X_train)
    X_test_tfidf = tfidf_vectorizer.transform(X_test)

    # Train and evaluate Perceptron model using TF-IDF features
    (
        _,
        precision_perceptron_tfidf,
        recall_perceptron_tfidf,
        f1_perceptron_tfidf,
        acc_perceptron_tfidf
    ) = train_and_evaluate_model(
        Perceptron, X_train_tfidf, y_train, X_test_tfidf, y_test, max_iter=4000
    )

    # Train and evaluate SVM model using TF-IDF features
    (
        _,
        precision_svm_tfidf,
        recall_svm_tfidf,
        f1_svm_tfidf,
        acc_svm_tfidf
    ) = train_and_evaluate_model(
        LinearSVC, X_train_tfidf, y_train, X_test_tfidf, y_test, max_iter=2500
    )

    # Print the results
    print("Precision Recall F1-Score Accuracy")
    print("Perceptron")
    print(
        f"{precision_perceptron_tfidf:.4f} {recall_perceptron_tfidf:.4f} {f1_perceptron_tfidf:.4f} {acc_perceptron_tfidf:.4f}"
    )

    print("SVM: LinearSVC")
    print(f"{precision_svm_tfidf:.4f} {recall_svm_tfidf:.4f} {f1_svm_tfidf:.4f} {acc_svm_tfidf:.4f}")


if __name__ == "__main__":
    main()
```

```
Overwriting HW1-CSCI544-wo-neg-sw.py
time: 50.9 ms (started: 2023-10-19 22:02:47 +00:00)
```

```
!python HW1-CSCI544-wo-neg-sw.py
```

```
Precision Recall F1-Score Accuracy
Perceptron
0.7637 0.8702 0.8135 0.7998
SVM: LinearSVC
0.8573 0.8602 0.8588 0.8581
time: 4min 1s (started: 2023-10-19 22:02:52 +00:00)
```

## Conclusion

### Best Accuracies

| Model | Accuracy | Features Used |
|---|---|---|
| Perceptron | 0.8110 | Word2Vec |
| LinearSVC | 0.8321 | Word2Vec |
| Perceptron | 0.7998 | TF-IDF |
| LinearSVC | 0.8581 | TF-IDF |

1. LinearSVC outperforms Perceptron for both feature types (Word2Vec and TF-IDF).
   - LinearSVC is better suited for this classification task compared to Perceptron.
2. When using Word2Vec features, both Perceptron and LinearSVC achieve lower accuracy compared to when using TF-IDF features.
   - Word2Vec embeddings might not be as effective for this specific sentiment classification task as compared to TF-IDF vectors.
3. The LinearSVC model performs particularly well with TF-IDF features, achieving an accuracy of 85.81%.
   - TF-IDF vectors are highly effective in capturing important information for sentiment classification in this dataset.

Overall, based on the provided performance metrics, it seems that TF-IDF features are more effective for this sentiment classification task compared to the Word2Vec embeddings. However, it's important to note that the effectiveness of features can vary depending on the specific dataset and task.

```
del balanced_df
del X_train, y_train, X_test, y_test
```

```
time: 557 µs (started: 2023-10-19 23:01:03 +00:00)
```

## Create Pytorch Dataset

- Custom pytorch dataset for on-the-fly processing an d efficient resource utilization
- Each sample in this dataset includes embeddings and their corresponding target label. The label is adjusted by subtracting 1 from the label value in the DataFrame

- Using `DataLoader`'s
    - Used to load and manage batches of data during the training process.
    - Handle tasks like shuffling, batching, and parallel data loading, making it easier to feed data to the model.

```
class AmazonReviewsSentimentDataset(Dataset):
    def __init__(self, df, embeddings_col_name, label_col_name):
        self.data = df
        self.embeddings_col_name = embeddings_col_name
        self.label_col_name = label_col_name

    def __len__(self):
        return len(self.data)

    def __getitem__(self, idx):
        if idx >= self.__len__():
            raise IndexError

        label = self.data.iloc[idx][self.label_col_name] - 1
        embeddings = self.data.iloc[idx][self.embeddings_col_name]

        return {
            "embeddings": torch.tensor(embeddings, dtype=torch.float32),
            "target":  torch.tensor(label, dtype=torch.long)
        }
```

```
time: 1.19 ms (started: 2023-10-19 23:01:06 +00:00)
```

```
train_dataset = AmazonReviewsSentimentDataset(
    train_df, embeddings_col_name="embeddings", label_col_name="sentiment"
)
valid_dataset = AmazonReviewsSentimentDataset(
    test_df, embeddings_col_name="embeddings", label_col_name="sentiment"
)
```

```
time: 702 µs (started: 2023-10-19 23:01:08 +00:00)
```

## Loaders & Samplers

```
TRAIN_BATCH_SIZE = 128
VALID_BATCH_SIZE = 64
NUM_PARALLEL_WORKERS = multiprocessing.cpu_count()
EPOCHS = 10
```

```
time: 769 µs (started: 2023-10-19 23:01:12 +00:00)
```

```
# train_sampler = RandomSampler(train_dataset)

train_data_loader = DataLoader(
    train_dataset,
    batch_size=TRAIN_BATCH_SIZE,
    # sampler=train_sampler,
    drop_last=True,
    shuffle=True,
    # num_workers=NUM_PARALLEL_WORKERS
)

# valid_sampler = RandomSampler(valid_dataset)

valid_data_loader = DataLoader(
    valid_dataset,
    batch_size=VALID_BATCH_SIZE,
    # sampler=valid_sampler,
    drop_last=False,
    shuffle=False,
    # num_workers=NUM_PARALLEL_WORKERS
)

test_data_loader = DataLoader(
    valid_dataset,
    batch_size=valid_dataset.data.__len__(),
    # sampler=valid_sampler,
    drop_last=False,
    shuffle=False,
    # num_workers=NUM_PARALLEL_WORKERS
)
```

```
time: 7.5 ms (started: 2023-10-19 18:41:36 +00:00)
```

## Training & Evaluation Functions

- `compute_accuracy` calculates the accuracy of model predictions given true labels.
- `train_loop_fn` handles one training epoch, updating the model's weights based on computed gradients.
- `eval_loop_fn` handles one validation epoch, computing the model's performance on the validation set.
- `train_and_evaluate` orchestrates the training process, saving checkpoints if specified. It reports metrics after each epoch. If a final model path is provided, it saves the model at the end.

```
def compute_accuracy(outputs, labels):
    predicted = torch.argmax(outputs.data, dim=1)

    predicted = predicted.detach().cpu().numpy()
    labels = labels.detach().cpu().numpy()

    acc = accuracy_score(labels, predicted)
    return acc

def train_loop_fn(data_loader, model, optimizer, loss_fn, device):
    model.train()
    train_loss = 0.0
    acc = []

    for batch in tqdm(data_loader):
        embeddings = batch['embeddings'].to(device, dtype=torch.float32, non_blocking=True)
        labels = batch['target'].to(device, dtype=torch.long, non_blocking=True)

        optimizer.zero_grad()

        outputs = model(embeddings.float())
        loss = loss_fn(outputs, labels)

```

```
        loss.backward()
        optimizer.step()

        train_loss += loss.item()*len(labels)
        acc.append(compute_accuracy(outputs, labels))

    acc = sum(acc)/len(acc)
    return train_loss, acc

def eval_loop_fn(data_loader, model, device):
    valid_loss = 0.0
    acc = []
    model.eval()

    for batch in data_loader:
        embeddings = batch['embeddings'].to(device, dtype=torch.float32, non_blocking=True)
        labels = batch['target'].to(device, dtype=torch.long, non_blocking=True)

        outputs = model(embeddings.float())

        loss = criterion(outputs, labels)
        valid_loss += loss.item()*len(labels)

        acc.append(compute_accuracy(outputs, labels))

    acc = sum(acc)/len(acc)

    return valid_loss, acc


def train_and_evaluate(
    model,
    train_data_loader, valid_data_loader,
    optimizer, loss_fn,
    device,
    num_epochs,
    checkpoint=False,
    path="model.pt"
):
    if checkpoint:
        dirname = path.split(".")[0]
        checkpoint_path = os.path.join(dirname)
        if os.path.exists(checkpoint_path):
            shutil.rmtree(checkpoint_path)
        os.makedirs(dirname)

    for epoch in range(num_epochs):
        # Train Step
        train_loss, train_acc = train_loop_fn(
            train_data_loader, model, optimizer, loss_fn, device
        )

        # Validation Step
        valid_loss, valid_acc = eval_loop_fn(valid_data_loader, model, device)

        train_loss /= len(train_data_loader.dataset)
        valid_loss /= len(valid_data_loader.dataset)

        if checkpoint:
            cp = os.path.join(checkpoint_path, f"{path}_{epoch}.pt")
            torch.save(model.state_dict(), cp)
            print(f"Saved Checkpoint to '{cp}'")

        epoch_log = (
            f"Epoch {epoch+1}/{num_epochs},"
            f" Train Accuracy={train_acc:.4f}, Validation Accuracy={valid_acc:.4f},"
            f" Train Loss={train_loss:.4f}, Validation Loss={valid_loss:.4f}"
        )
        print(epoch_log)

    torch.save(model.state_dict(), path)
    return model
```

time: 2.94 ms (started: 2023-10-19 23:01:15 +00:00)

## Feedforward Neural Networks

```
class MLP(nn.Module):
    def __init__(self, num_input_features, num_classes):
        super(MLP, self).__init__()
        # Input size is 300 (Word2Vec dimensions)
        self.fc1 = nn.Linear(num_input_features, 50)
        self.fc2 = nn.Linear(50, 5)
        # Output size is 2 for binary classification
        self.fc3 = nn.Linear(5, num_classes)

    def forward(self, x):
        x = torch.relu(self.fc1(x))
        x = torch.relu(self.fc2(x))
        x = self.fc3(x)
        return x


net = MLP(num_input_features=Word2VecConfig.MAX_LENGTH, num_classes=2)
criterion = nn.CrossEntropyLoss()
optimizer = optim.SGD(net.parameters(), lr=0.01)
net = net.to(device)
```

time: 4.36 ms (started: 2023-10-19 10:33:35 +00:00)

## With average Word2Vec features

```
model = train_and_evaluate(
    model=net,
    train_data_loader=train_data_loader,
    valid_data_loader=valid_data_loader,
    optimizer=optimizer,
    loss_fn=criterion,
    device=device,
    num_epochs=20,
    checkpoint=True,
    path="mlp_w_avg_w2v_feat.pt"
)
```

```
100%                                   624/624 [00:21<00:00, 35.15it/s]
Saved Checkpoint to 'mlp_w_avg_w2v_feat/mlp_w_avg_w2v_feat.pt_0.pt'
Epoch 1/20, Train Accuracy=0.7113, Validation Accuracy=0.7318, Train Loss=0.6679, Validation Loss=0.6583
100%                                   624/624 [00:21<00:00, 20.21it/s]
Saved Checkpoint to 'mlp_w_avg_w2v_feat/mlp_w_avg_w2v_feat.pt_1.pt'
Epoch 2/20, Train Accuracy=0.7298, Validation Accuracy=0.7375, Train Loss=0.6438, Validation Loss=0.6256
100%                                   624/624 [00:19<00:00, 35.05it/s]
Saved Checkpoint to 'mlp_w_avg_w2v_feat/mlp_w_avg_w2v_feat.pt_2.pt'
Epoch 3/20, Train Accuracy=0.7470, Validation Accuracy=0.7608, Train Loss=0.6036, Validation Loss=0.5779
100%                                   624/624 [00:19<00:00, 34.62it/s]
Saved Checkpoint to 'mlp_w_avg_w2v_feat/mlp_w_avg_w2v_feat.pt_3.pt'
Epoch 4/20, Train Accuracy=0.7637, Validation Accuracy=0.7773, Train Loss=0.5553, Validation Loss=0.5302
100%                                   624/624 [00:21<00:00, 13.51it/s]
Saved Checkpoint to 'mlp_w_avg_w2v_feat/mlp_w_avg_w2v_feat.pt_4.pt'
Epoch 5/20, Train Accuracy=0.7766, Validation Accuracy=0.7893, Train Loss=0.5137, Validation Loss=0.4946
100%                                   624/624 [00:20<00:00, 33.96it/s]
Saved Checkpoint to 'mlp_w_avg_w2v_feat/mlp_w_avg_w2v_feat.pt_5.pt'
Epoch 6/20, Train Accuracy=0.7862, Validation Accuracy=0.7958, Train Loss=0.4848, Validation Loss=0.4710
100%                                   624/624 [00:21<00:00, 34.93it/s]
Saved Checkpoint to 'mlp_w_avg_w2v_feat/mlp_w_avg_w2v_feat.pt_6.pt'
Epoch 7/20, Train Accuracy=0.7932, Validation Accuracy=0.7976, Train Loss=0.4662, Validation Loss=0.4568
100%                                   624/624 [00:27<00:00, 34.54it/s]
Saved Checkpoint to 'mlp_w_avg_w2v_feat/mlp_w_avg_w2v_feat.pt_7.pt'
Epoch 8/20, Train Accuracy=0.7987, Validation Accuracy=0.8046, Train Loss=0.4536, Validation Loss=0.4458
100%                                   624/624 [00:35<00:00, 28.22it/s]
Saved Checkpoint to 'mlp_w_avg_w2v_feat/mlp_w_avg_w2v_feat.pt_8.pt'
Epoch 9/20, Train Accuracy=0.8035, Validation Accuracy=0.8087, Train Loss=0.4446, Validation Loss=0.4378
100%                                   624/624 [00:23<00:00, 15.18it/s]
Saved Checkpoint to 'mlp_w_avg_w2v_feat/mlp_w_avg_w2v_feat.pt_9.pt'
Epoch 10/20, Train Accuracy=0.8066, Validation Accuracy=0.8103, Train Loss=0.4374, Validation Loss=0.4322
100%                                   624/624 [00:29<00:00, 7.02it/s]
Saved Checkpoint to 'mlp_w_avg_w2v_feat/mlp_w_avg_w2v_feat.pt_10.pt'
Epoch 11/20, Train Accuracy=0.8097, Validation Accuracy=0.8128, Train Loss=0.4314, Validation Loss=0.4266
100%                                   624/624 [00:21<00:00, 22.66it/s]
Saved Checkpoint to 'mlp_w_avg_w2v_feat/mlp_w_avg_w2v_feat.pt_11.pt'
Epoch 12/20, Train Accuracy=0.8116, Validation Accuracy=0.8178, Train Loss=0.4263, Validation Loss=0.4221
100%                                   624/624 [00:25<00:00, 13.12it/s]
Saved Checkpoint to 'mlp_w_avg_w2v_feat/mlp_w_avg_w2v_feat.pt_12.pt'
Epoch 13/20, Train Accuracy=0.8137, Validation Accuracy=0.8198, Train Loss=0.4220, Validation Loss=0.4186
100%                                   624/624 [00:32<00:00, 34.89it/s]
Saved Checkpoint to 'mlp_w_avg_w2v_feat/mlp_w_avg_w2v_feat.pt_13.pt'
Epoch 14/20, Train Accuracy=0.8153, Validation Accuracy=0.8206, Train Loss=0.4178, Validation Loss=0.4147
100%                                   624/624 [00:19<00:00, 33.72it/s]
Saved Checkpoint to 'mlp_w_avg_w2v_feat/mlp_w_avg_w2v_feat.pt_14.pt'
Epoch 15/20, Train Accuracy=0.8171, Validation Accuracy=0.8211, Train Loss=0.4143, Validation Loss=0.4115
100%                                   624/624 [00:21<00:00, 30.96it/s]
Saved Checkpoint to 'mlp_w_avg_w2v_feat/mlp_w_avg_w2v_feat.pt_15.pt'
Epoch 16/20, Train Accuracy=0.8182, Validation Accuracy=0.8241, Train Loss=0.4112, Validation Loss=0.4102
100%                                   624/624 [00:21<00:00, 30.42it/s]
Saved Checkpoint to 'mlp_w_avg_w2v_feat/mlp_w_avg_w2v_feat.pt_16.pt'
Epoch 17/20, Train Accuracy=0.8190, Validation Accuracy=0.8240, Train Loss=0.4083, Validation Loss=0.4066
100%                                   624/624 [00:21<00:00, 21.98it/s]
Saved Checkpoint to 'mlp_w_avg_w2v_feat/mlp_w_avg_w2v_feat.pt_17.pt'
Epoch 18/20, Train Accuracy=0.8191, Validation Accuracy=0.8245, Train Loss=0.4058, Validation Loss=0.4044
100%                                   624/624 [00:19<00:00, 33.59it/s]
Saved Checkpoint to 'mlp_w_avg_w2v_feat/mlp_w_avg_w2v_feat.pt_18.pt'
Epoch 19/20, Train Accuracy=0.8220, Validation Accuracy=0.8261, Train Loss=0.4035, Validation Loss=0.4025
100%                                   624/624 [00:19<00:00, 33.92it/s]
Saved Checkpoint to 'mlp_w_avg_w2v_feat/mlp_w_avg_w2v_feat.pt_19.pt'
Epoch 20/20, Train Accuracy=0.8228, Validation Accuracy=0.8267, Train Loss=0.4014, Validation Loss=0.4009
time: 9min 29s (started: 2023-10-19 07:40:53 +00:00)
```

Overall Accuracy on Test Set

```
path_to_saved_model = 'mlp_w_avg_w2v_feat.pt'
model = MLP(num_input_features=Word2VecConfig.MAX_LENGTH, num_classes=2)
model.load_state_dict(torch.load(path_to_saved_model, map_location=device))

for batch in test_data_loader:
    embeddings = batch['embeddings'].to(device, dtype=torch.float32, non_blocking=True)
    y_pred = model(embeddings)
    y_true = batch["target"].to(device, dtype=torch.long, non_blocking=True)

acc = compute_accuracy(y_pred, y_true)
print("Accuracy (Test Dataset):", round(acc,4))
```

```
Accuracy (Test Dataset): 0.8262
time: 8.79 s (started: 2023-10-19 10:33:58 +00:00)
```

## With top 10 Word2Vec features

- Embeddings are padded for maintaining consistent input dimensions across different samples in a batch.

```
class ARDatasetWithTop10Embeddings(Dataset):
    def __init__(self, df, embeddings_col_name, label_col_name, max_length):
        self.data = df
        self.embeddings_col_name = embeddings_col_name
        self.label_col_name = label_col_name
        self.max_length = max_length

    def __len__(self):
        return len(self.data)

    def __getitem__(self, idx):
        if idx >= self.__len__():
            raise IndexError

        label = self.data.iloc[idx][self.label_col_name] - 1
        embeddings = self.data.iloc[idx][self.embeddings_col_name]

        # Pad embeddings to max_length
        if len(embeddings) < self.max_length:
            padding = np.zeros(self.max_length - len(embeddings))
            embeddings = np.concatenate((embeddings, padding))

```

```
        return {
            "embeddings": torch.tensor(embeddings, dtype=torch.float32),
            "target":  torch.tensor(label, dtype=torch.long)
        }

train_dataset = ARDatasetWithTop10Embeddings(
    train_df, embeddings_col_name="embeddings_top_10", label_col_name="sentiment", max_length=3000
)
valid_dataset = ARDatasetWithTop10Embeddings(
    test_df, embeddings_col_name="embeddings_top_10", label_col_name="sentiment", max_length=3000
)

train_data_loader = DataLoader(
    train_dataset,
    batch_size=TRAIN_BATCH_SIZE,
    drop_last=True,
    shuffle=True,
)

valid_data_loader = DataLoader(
    valid_dataset,
    batch_size=VALID_BATCH_SIZE,
    drop_last=False,
    shuffle=False,
)

test_data_loader = DataLoader(
    valid_dataset,
    batch_size=valid_dataset.__len__(),
    drop_last=False,
    shuffle=False,
)
```

```
time: 12 ms (started: 2023-10-19 10:36:08 +00:00)
```

```
net2 = MLP(num_input_features=3000, num_classes=2)
criterion = nn.CrossEntropyLoss()
optimizer = optim.SGD(net2.parameters(), lr=0.01)
net2 = net2.to(device)


model2 = train_and_evaluate(
    model=net2,
    train_data_loader=train_data_loader,
    valid_data_loader=valid_data_loader,
    optimizer=optimizer,
    loss_fn=criterion,
    device=device,
    num_epochs=20,
    checkpoint=True,
    path="mlp_w_top10_w2v_feat.pt"
)
```

```
100%                                   624/624 [00:30<00:00, 25.53it/s]
Saved Checkpoint to 'mlp_w_top10_w2v_feat/mlp_w_top10_w2v_feat.pt_0.pt'
Epoch 1/20, Train Accuracy=0.4960, Validation Accuracy=0.4998, Train Loss=0.6930, Validation Loss=0.6932
100%                                   624/624 [00:31<00:00, 17.36it/s]
Saved Checkpoint to 'mlp_w_top10_w2v_feat/mlp_w_top10_w2v_feat.pt_1.pt'
```

Overall Accracy on Test Set

```
path_to_saved_model = 'mlp_w_top10_w2v_feat.pt'
model = MLP(num_input_features=3000, num_classes=2)
model.load_state_dict(torch.load(path_to_saved_model, map_location=device))

for batch in test_data_loader:
    embeddings = batch['embeddings'].to(device, dtype=torch.float32, non_blocking=True)
    y_pred = model(embeddings)
    y_true = batch["target"].to(device, dtype=torch.long, non_blocking=True)

acc = compute_accuracy(y_pred, y_true)
print("Accuracy (Test Dataset):", round(acc,4))
```

```
Accuracy (Test Dataset): 0.7589
time: 5.48 s (started: 2023-10-19 10:36:17 +00:00)
Epoch 7/20, Train Accuracy=0.5746, Validation Accuracy=0.5004, Train Loss=0.6922, Validation Loss=0.6919
```

### Comparision with Simple Model

The LinearSVC model trained on TF-IDF features was the most effective in this scenario, outperforming both simple models and MLP models trained with Word2Vec embeddings.

### Conclusion

1. **Feature Importance**:
    - The choice of features significantly impacts model performance.
    - In this case, TF-IDF features proved to be the most informative for sentiment analysis, as evidenced by the high accuracy achieved by LinearSVC with TF-IDF.
2. **Complexity vs. Performance**:
    - Simple models like Perceptron and LinearSVC can sometimes outperform more complex models.
    - This is evident in the case where LinearSVC with TF-IDF outperformed the MLP models.
3. **Embedding Selection**:
    - Not all embeddings are equally effective. The choice of Word2Vec embeddings, particularly using the average vectors, yielded competitive results, showcasing the importance of using quality word embeddings.
4. **Dimensionality Matters**:
    - Using only the top 10 Word2Vec embeddings didn't capture enough information for sentiment analysis.
    - It's important to consider the dimensionality of the embeddings and how well they represent the underlying semantics.

## Recurrent Neural Networks

```
class ARDatasetFullEmb(Dataset):
    def __init__(self, df, embeddings_col_name, label_col_name, max_length):
        self.data = df
        self.embeddings_col_name = embeddings_col_name
        self.label_col_name = label_col_name
        self.max_length = max_length

    def __len__(self):
        return len(self.data)

    def __getitem__(self, idx):
        if idx >= self.__len__():
            raise IndexError

        label = self.data.iloc[idx][self.label_col_name] - 1
        embeddings = self.data.iloc[idx][self.embeddings_col_name]

        # Pad embeddings to max_length
        if len(embeddings) < self.max_length:
            padding = np.zeros(self.max_length - len(embeddings))
            embeddings = np.concatenate((embeddings, padding))

        embeddings = embeddings.reshape(10, 300)

        return {
            "embeddings": torch.tensor(embeddings, dtype=torch.float32),
            "target":  torch.tensor(label, dtype=torch.long)
        }

train_dataset = ARDatasetFullEmb(
    train_df, embeddings_col_name="embeddings_top_10", label_col_name="sentiment", max_length=3000,
)
valid_dataset = ARDatasetFullEmb(
    test_df, embeddings_col_name="embeddings_top_10", label_col_name="sentiment", max_length=3000,
)

train_data_loader = DataLoader(
    train_dataset,
    batch_size=TRAIN_BATCH_SIZE,
    drop_last=True,
    shuffle=True,
)

valid_data_loader = DataLoader(
    valid_dataset,
    batch_size=VALID_BATCH_SIZE,
    drop_last=False,
    shuffle=False,
)

test_data_loader = DataLoader(
    valid_dataset,
    batch_size=valid_dataset.__len__(),
    drop_last=False,
    shuffle=False,
)
```

```
time: 2.38 ms (started: 2023-10-19 23:51:17 +00:00)
```

```python
def compute_accuracy(outputs, labels):
    predicted = torch.argmax(outputs.data, dim=1)

    predicted = predicted.detach().cpu().numpy()
    labels = labels.detach().cpu().numpy()

    acc = accuracy_score(labels, predicted)
    return acc

def train_loop_fn(data_loader, model, optimizer, loss_fn, device):
    model.train()
    train_loss = 0.0
    acc = []

    for batch in tqdm(data_loader):
        optimizer.zero_grad()

        embeddings = batch['embeddings'].detach()
        labels = batch['target'].detach()

        all_emb = torch.stack(embeddings).to(device, dtype=torch.float32, non_blocking=True)
        all_lb = torch.stack(labels).to(device, dtype=torch.long, non_blocking=True)

        outputs, _ = model(all_emb.float())
        loss = loss_fn(outputs, all_lb)

        loss.backward()
        optimizer.step()

        train_loss += loss.item()*len(all_lb)
        acc.append(compute_accuracy(outputs, all_lb))

    acc = sum(acc)/len(acc)
    return train_loss, acc

def eval_loop_fn(data_loader, model, device):
    valid_loss = 0.0
    acc = []
    model.eval()

    for batch in data_loader:
        embeddings = batch['embeddings'].detach()
        labels = batch['target'].detach()

        all_emb = torch.stack(embeddings).to(device, dtype=torch.float32, non_blocking=True)
        all_lb = torch.stack(labels).to(device, dtype=torch.long, non_blocking=True)

        outputs, _ = model(all_emb.float())

        loss = criterion(outputs, all_lb)
        valid_loss += loss.item()*len(all_lb)

        acc.append(compute_accuracy(outputs, all_lb))

    acc = sum(acc)/len(acc)

    return valid_loss, acc


def train_and_evaluate(
    model,
    train_data_loader, valid_data_loader,
    optimizer, loss_fn,
    device,
    num_epochs,
    checkpoint=False,
    path="model.pt"
):
    if checkpoint:
        dirname = path.split(".")[0]
        checkpoint_path = os.path.join(dirname)
        if os.path.exists(checkpoint_path):
            shutil.rmtree(checkpoint_path)
        os.makedirs(dirname)

    for epoch in range(num_epochs):
        # Train Step
        train_loss, train_acc = train_loop_fn(
            train_data_loader, model, optimizer, loss_fn, device
        )

        # Validation Step
        valid_loss, valid_acc = eval_loop_fn(valid_data_loader, model, device)

        train_loss /= len(train_data_loader.dataset)
        valid_loss /= len(valid_data_loader.dataset)

        if checkpoint:
            cp = os.path.join(checkpoint_path, f"{path}_{epoch}.pt")
            torch.save(model.state_dict(), cp)
            print(f"Saved Checkpoint to '{cp}'")

        epoch_log = (
            f"Epoch {epoch+1}/{num_epochs},"
            f" Train Accuracy={train_acc:.4f}, Validation Accuracy={valid_acc:.4f},"
            f" Train Loss={train_loss:.4f}, Validation Loss={valid_loss:.4f}"
        )
        print(epoch_log)

    torch.save(model.state_dict(), path)
    return model
```

```
time: 3.38 ms (started: 2023-10-19 23:56:33 +00:00)
```

## Simple RNN

```python
class RNNModel(nn.Module):
    def __init__(
        self, input_size, hidden_size, num_layers, output_size, model_type="rnn"
    ):
        super(RNNModel, self).__init__()

        self.hidden_size = hidden_size
        self.num_layers = num_layers
        self.model_type = model_type

```

```
        if model_type == "gru":
            self.layer = nn.GRU(input_size, hidden_size, num_layers, batch_first=True)
        elif model_type == "lstm":
            self.layer = nn.LSTM(input_size, hidden_size, num_layers, batch_first=True)
        else:
            self.layer = nn.RNN(input_size, hidden_size, num_layers, batch_first=True)

        #dropout layer
        self.dropout = nn.Dropout(0.3)

        # Fully connected layers
        self.fc = nn.Linear(hidden_size, output_size)

    def forward(self, x):
        batch_size = x.size(0)
        hidden = self.init_hidden(batch_size)

        out, hidden = self.layer(x, hidden)
        # Stack up the model output
        out = out.contiguous().view(-1, self.hidden_size)

        out = self.dropout(out)
        # Only use the output from the last time step
        out = self.fc(out)
        return out, hidden

    def init_hidden(self, batch_size):
        hidden = torch.zeros(self.num_layers, batch_size, self.hidden_size).to(device)
        return hidden
```

```
time: 1.53 ms (started: 2023-10-19 23:56:41 +00:00)
```

```
input_size = 300
hidden_size = 10
output_size = 2

net3 = RNNModel(input_size, hidden_size, 10, output_size, model_type="rnn").to(device)
criterion = nn.CrossEntropyLoss()
optimizer = torch.optim.Adam(net3.parameters(), lr=0.01)

model3 = train_and_evaluate(
    model=net3,
    train_data_loader=train_data_loader,
    valid_data_loader=valid_data_loader,
    optimizer=optimizer,
    loss_fn=criterion,
    device=device,
    num_epochs=25,
    checkpoint=True,
    path="simple_rnn_w2v_feat.pt"
)
```

0% 0/624 [00:00<?, ?it/s]

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
<ipython-input-65-6b1bcedac76e> in <cell line: 9>()
      7 optimizer = torch.optim.Adam(net3.parameters(), lr=0.01)
      8
----> 9 model3 = train_and_evaluate(
     10     model=net3,
     11     train_data_loader=train_data_loader,
```

1 frames

```
<ipython-input-63-a604df9676f1> in train_loop_fn(data_loader, model, optimizer, loss_fn, device)
     19         labels = batch['target'].detach()
     20
---> 21         all_emb = torch.stack(embeddings).to(device, dtype=torch.float32, non_blocking=True)
     22         all_lb = torch.stack(labels).to(device, dtype=torch.long, non_blocking=True)
     23

TypeError: stack(): argument 'tensors' (position 1) must be tuple of Tensors, not Tensor
```

SEARCH STACK OVERFLOW

```
time: 134 ms (started: 2023-10-19 23:56:42 +00:00)
```

## GRU

```
input_size = 300
hidden_size = 10
output_size = 2

net4 = RNNModel(input_size, hidden_size, 10, output_size, model_type="gru").to(device)
criterion = nn.CrossEntropyLoss()
optimizer = torch.optim.Adam(net4.parameters(), lr=0.01)

model4 = train_and_evaluate(
    model=net4,
    train_data_loader=train_data_loader,
    valid_data_loader=valid_data_loader,
    optimizer=optimizer,
    loss_fn=criterion,
    device=device,
    num_epochs=25,
    checkpoint=True,
    path="gru_w2v_feat.pt"
)
```

## LSTM

```
input_size = 300
hidden_size = 10
output_size = 2

net5 = RNNModel(input_size, hidden_size, 10, output_size, model_type="lstm").to(device)
criterion = nn.CrossEntropyLoss()
optimizer = torch.optim.Adam(net5.parameters(), lr=0.01)

model5 = train_and_evaluate(
    model=net5,
    train_data_loader=train_data_loader,
    valid_data_loader=valid_data_loader,
    optimizer=optimizer,
    loss_fn=criterion,
    device=device,
```

```
    num_epochs=25,
    checkpoint=True,
    path="lstm_w2v_feat.pt"
)
```

### Conclusion

1. **Feature Representations**:
   - TF-IDF outperforms Word2Vec across all models.
   - Averaged Word2Vec is better than Concatenated Word2Vec.
2. **Model Comparisons**:
   - SVM outperforms Perceptron consistently.
   - MLP with averaged Word2Vec performs better than RNN, GRU, and LSTM with Word2Vec.
3. **Recurrent Models**:
   - RNN, GRU, and LSTM show similar performance with Word2Vec embeddings.
4. **Overall Performance**:
   - Highest accuracy (~87%) is achieved with SVM using TF-IDF.

Other:

- Averaging Word2Vec embeddings seems a more effective representation
- SVM model is better at capturing the non-linear relationships in the data compared to the Perceptron
- TF-IDF may capture important information more effectively than Word2Vec embeddings

## References


1. https://www.kaggle.com/code/abhishek/bert-multi-lingual-tpu-training-8-cores
2. https://www.kaggle.com/mishra1993/pytorch-multi-layer-perceptron-mnist
3. https://www.kaggle.com/code/arunmohan003/sentiment-analysis-using-lstm-pytorch
4. https://pytorch.org/docs/stable/generated/torch.nn.RNN.html
5. https://pytorch.org/tutorials/intermediate/char_rnn_classification_tutorial.html
6. My own HW1 python file submission
7. https://piazza.com/class/llm91seaknw3j6/post/408


## THE END