Milestone 2

```
import sys
sys.path.append('../../../src')

import glob, os
import numpy as np
import scipy as sp
import pandas as pd
from sklearn import decomposition, datasets
from matplotlib import pyplot as plt
from p1.models.utils import get_divided_data, get_data_comp, get_common_label, creat_pred_

%load_ext autoreload
%autoreload 2
```

Data Loading

```
# Get Files
data_dir = "../project_data/data"
data_dir_dict = dict()
for folder in os.listdir(data_dir):
    # print(folder)
    data_dir_dict[folder] = list()
    for file_name in glob.glob(f"{data_dir}/{folder}/*.csv"):
        # prprint(bag_of_words_k_folds)int(file_name)
        data_dir_dict[folder].append(os.path.abspath("./") + "/" + file_name)
        # print(os.path.abspath("./") + "/" + file_name)

# eval, test, train
# Load "bag-of-words" data
bag_of_words, glove, tfidf = list(), list(), list()
for i in range(3):
    bag_of_words.append(pd.read_csv(data_dir_dict["bag-of-words"][i]))
    glove.append(pd.read_csv(data_dir_dict["glove"][i]))
    tfidf.append(pd.read_csv(data_dir_dict["tfidf"][i]))
```

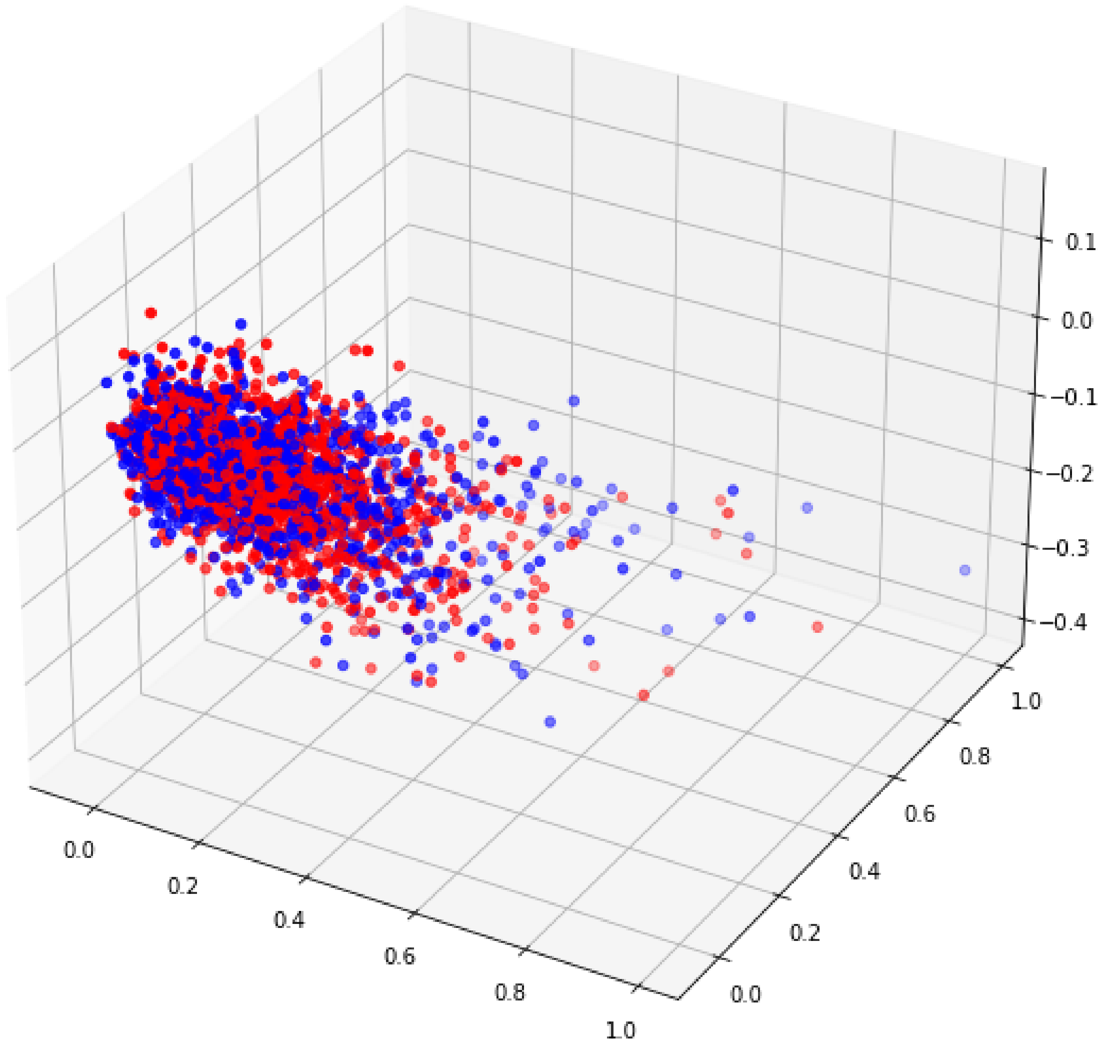
Glove Data Visualization

```
glove_train, glove_labels = get_data_comp(glove[2], labels=[1,0])
x = glove_train.to_numpy()

n_samps, n_features = x.shape
```

```
pca = decomposition.PCA(min(n_samps, n_features))
pca.fit(x)

s_vals = pca.components_[:, :3]
```

```
fig = plt.figure(figsize=[10, 10])
ax = plt.axes(projection='3d')
colors = ['b' if glove_labels[i] else 'r' for i in range(n_samps)]

y = np.dot(x, s_vals)
y = y/y.max()

ax.scatter3D(y[:,0], y[:, 1], glove_labels, c=glove_labels)
```

```
<mpl_toolkits.mplot3d.art3d.Path3DCollection at 0x70a4d0a02d30>
```



Discussion: I tried to visualize the dataset given its dot product with the principle components of the dataset. Need to do further visualization before I continue designing my final model.