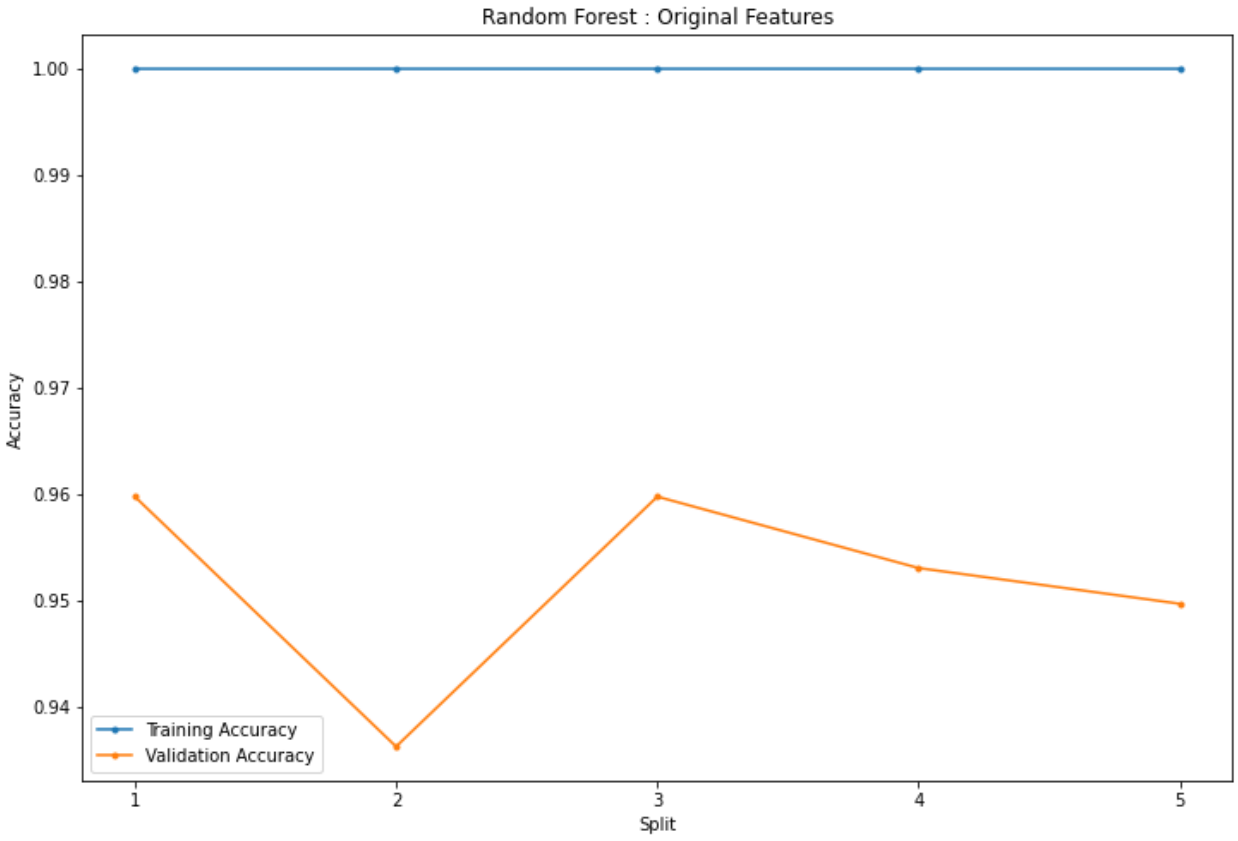
Christian Domacena
Assignment 2

**Preprocessing**
I used the preprocessing code from quiz 1 using nltk, porterstemmer. Also, I generated a new version of the "**news-train.csv**" to hold the cleaned text with its corresponding articleId and category named "**news-train-cleaned.csv**".

## Q1

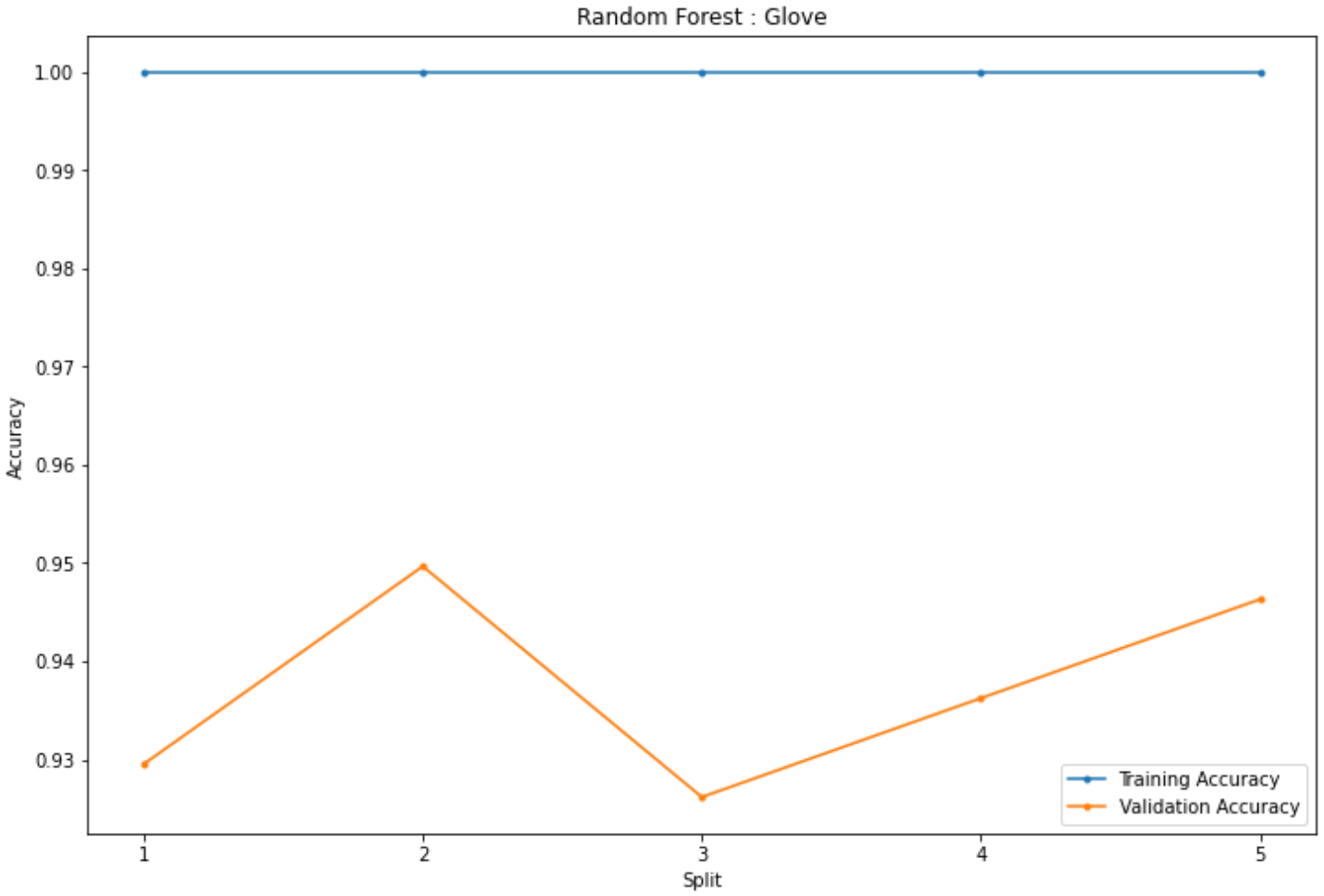
A. Random forest with **original features**

| n-estimators = 100 | Training | Validation |
|---|---|---|
| Avg. Accuracy | 1.0 | 0.9516778523489933 |
| Avg. Std. Deviation | 0.0 | 0.008647046125318842 |



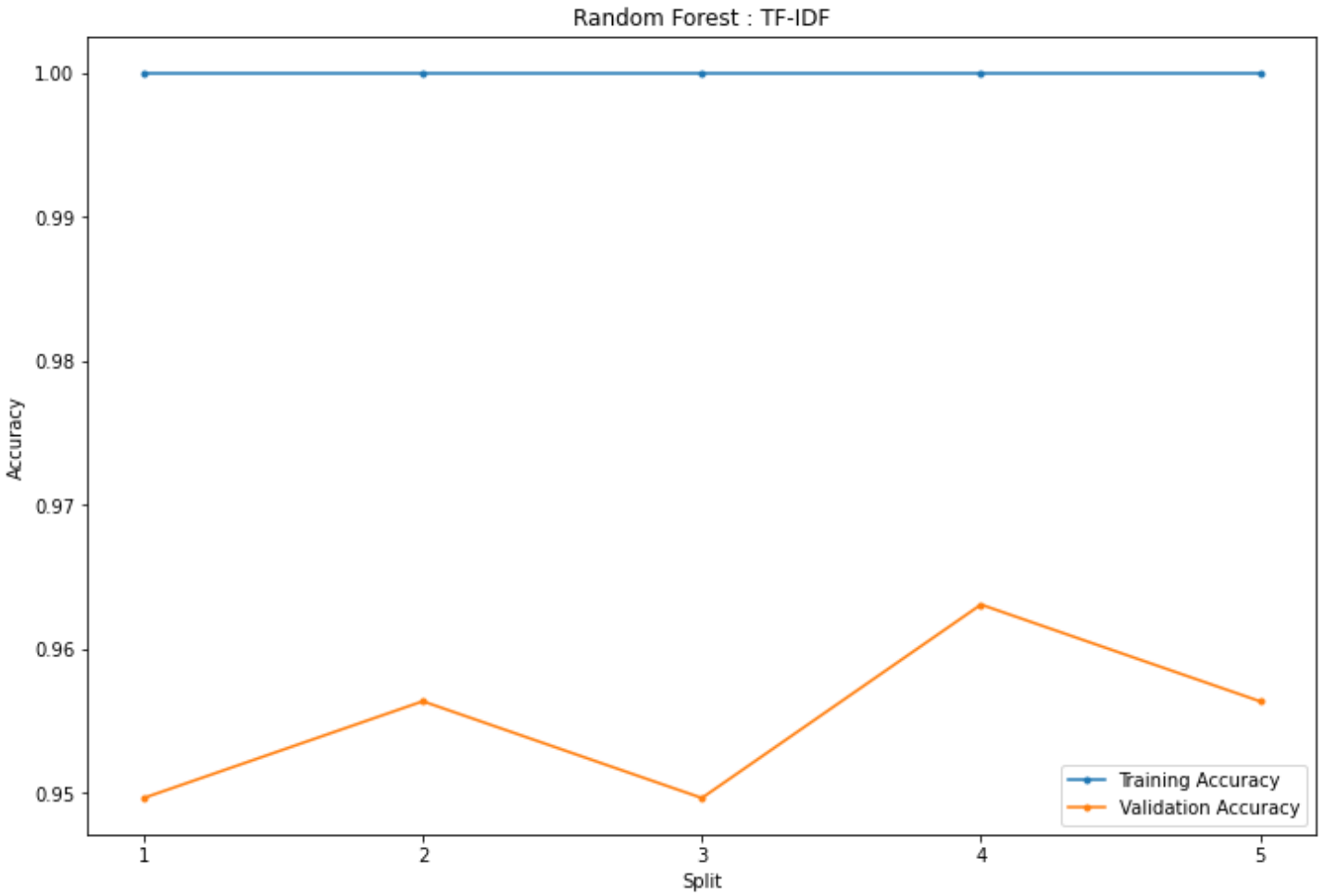
B.
Random forest - **Glove**

| n-estimators = 100 | Training | Validation |
|---|---|---|
| Avg. Accuracy | 1.0 | 0.9375838926174496 |
| Avg. Std. Deviation | 0.0 | 0.009153142078513988 |



Random Forest - TFIDF Vectors

| n-estimators = 100 | Training | Validation |
|---|---|---|
| Avg. Accuracy | 1.0 | 0.9550335570469798 |
| Avg. Std. Deviation | 0.0 | 0.00502235890842139 |



**Q2.**

1. I used sklearn's GridSearchCV to get the best parameter for my classification model.

Parameters

| C | kernel |
|---|---|
| 1 | rbf |

2. I used glove to generate features from q1

| c=1, kernel='rbf' | Training | Validation |
|---|---|---|
| Avg. Accuracy | 0.9380872483221477 | 0.9261744966442954 |
| Avg. Std. Deviation | 0.0014433431656111569 | 0.011816655611851684 |