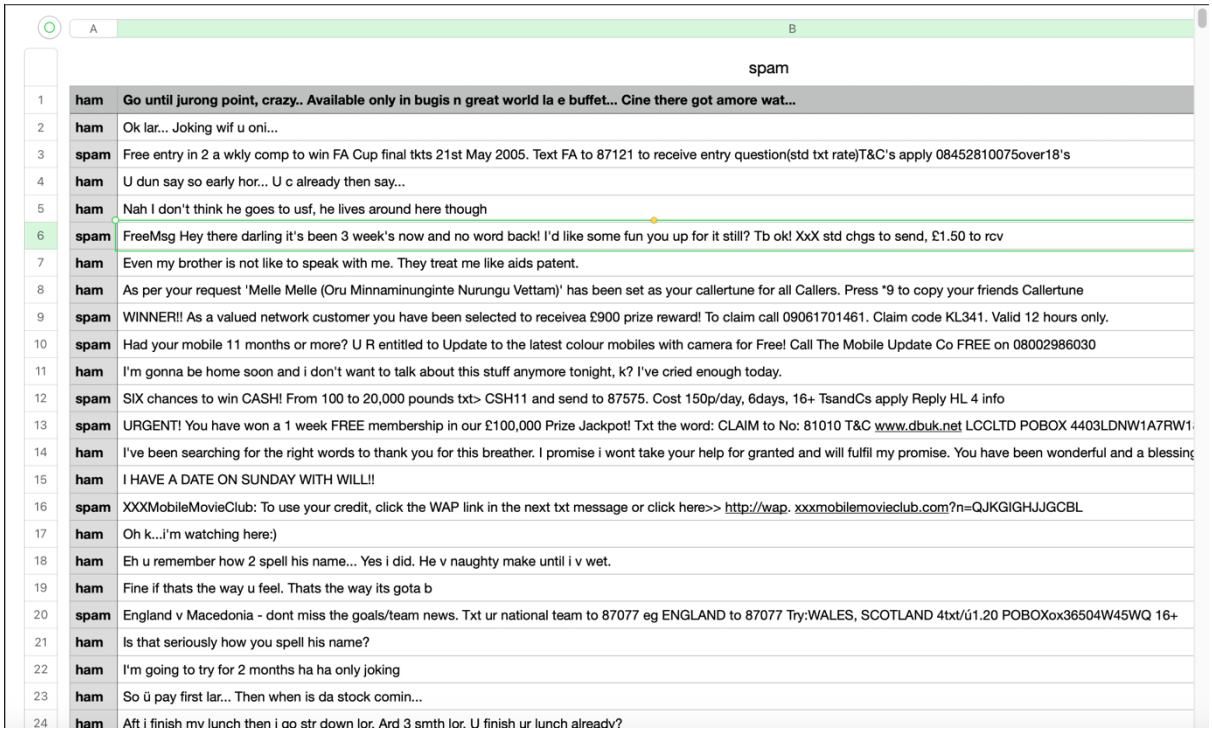
Building a Smarter AI-Powered Spam Classifier: Designing a Web Application to Classify Spam Messages Using TF-IDF, Multinomial Naive Bayes, and Other NLTK Libraries with Iterative Improvement to Enhance Accuracy, Precision, Recall, and F1-score.

1. Data Collection:

-Download the "spam.csv" dataset from Kaggle's SMS Spam Collection Dataset (https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset).



	A	B
		spam
1	ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
2	ham	Ok lar... Joking wif u oni...
3	spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's
4	ham	U dun say so early hor... U c already then say...
5	ham	Nah I don't think he goes to usf, he lives around here though
6	spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, £1.50 to rcv
7	ham	Even my brother is not like to speak with me. They treat me like aids patent.
8	ham	As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your friends Callertune
9	spam	WINNER!! As a valued network customer you have been selected to receivea £900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only.
10	spam	Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile Update Co FREE on 08002986030
11	ham	I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried enough today.
12	spam	SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info
13	spam	URGENT! You have won a 1 week FREE membership in our £100,000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C www.dbuk.net LCCLTD POBOX 4403LDNW1A7RW1
14	ham	I've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will fulfil my promise. You have been wonderful and a blessing
15	ham	I HAVE A DATE ON SUNDAY WITH WILL!!
16	spam	XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> http://wap. xxxmobilemovieclub.com?n=QJKGIGHJJGCBL
17	ham	Oh k...i'm watching here:)
18	ham	Eh u remember how 2 spell his name... Yes i did. He v naughty make until i v wet.
19	ham	Fine if thats the way u feel. Thats the way its gota b
20	spam	England v Macedonia - dont miss the goals/team news. Txt ur national team to 87077 eg ENGLAND to 87077 Try:WALES, SCOTLAND 4txt/ú1.20 POBOXox36504W45WQ 16+
21	ham	Is that seriously how you spell his name?
22	ham	I'm going to try for 2 months ha ha only joking
23	ham	So ü pay first lar... Then when is da stock comin...
24	ham	Aft i finish my lunch then i go str down lor. Ard 3 smth lor. U finish ur lunch already?

2.Data Preprocessing:

-Load the dataset from "./src/data/spam.csv."

```
data = pd.read_csv(r"./src/data/spam.csv",sep="\t",names=["label", "message"])
```

3. Feature Extraction:

- Apply TF-IDF (Term Frequency-Inverse Document Frequency) to convert the tokenized words into numerical features.

- Set the maximum number of features to 2500, meaning you'll create a TF-IDF matrix with 2500 columns representing the most important words in your dataset.

```
self.porterStemmer = PorterStemmer()
corpus = []
for i in range(0, len(self.data)):
    review = re.sub('[^a-zA-Z]', ' ', self.data['message'][i]).lower().split()
    review = [self.porterStemmer.stem(word) for word in review if not word in stopwords.words('english')]
    review = ' '.join(review)
    corpus.append(review)

self.TfidfVectorization = TfidfVectorizer(max_features=2500)
```

4. Model Selection:

- Choose the Multinomial Naive Bayes algorithm as your initial machine learning model. Import the necessary libraries from scikit-learn.
- Split your preprocessed data into a training set and a test set for model evaluation.
- Train the Multinomial Naive Bayes model on the TF-IDF-transformed training data.

```
X = self.TfidfVectorization.fit_transform(corpus).toarray()

y=pd.get_dummies(self.data['label']).iloc[:,1].values

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

self.spam_detect_model = MultinomialNB().fit(X_train, y_train)

y_predict=self.spam_detect_model.predict(X_test)
```

5. Evaluation:

- Use the trained model to make predictions on the test dataset.
- Calculate various performance metrics, including:
  - Accuracy: The proportion of correctly classified messages.
  - Precision: The proportion of true spam messages among the messages classified as spam.
  - Recall: The proportion of true spam messages correctly classified as spam.
  - F1-score: A harmonic mean of precision and recall, which balances both metrics.
- Assess the model's performance using these metrics to determine its initial effectiveness.

```
self.currectAccuracyScore = accuracy_score(y_predict, y_test)
self.currentPrecisionScore = precision_score(y_predict, y_test)
self.currentRelayScore = recall_score(y_predict, y_test)
self.currentf1Score = f1_score(y_predict, y_test)
```

6. Iterative Improvement:

- When a new message is predicted using your initial model, update your dataset by appending the prediction result ("spam" or "ham") along with the message.
  - Periodically, or when enough new data is collected, refresh the model by repeating steps 2 to 5 using the updated dataset.

```
self.data.loc[len(self.data)]={"label":prediction,"message":new_message}
self.refreshModel()
```

7. Deployment:

- Once you have achieved satisfactory performance, you can deploy your web application.
  - Develop a user interface for users to input messages and receive spam classification results.
  - Host the web application on a server or cloud platform.
  - Ensure the model is retrained periodically with new data to maintain its accuracy and effectiveness.