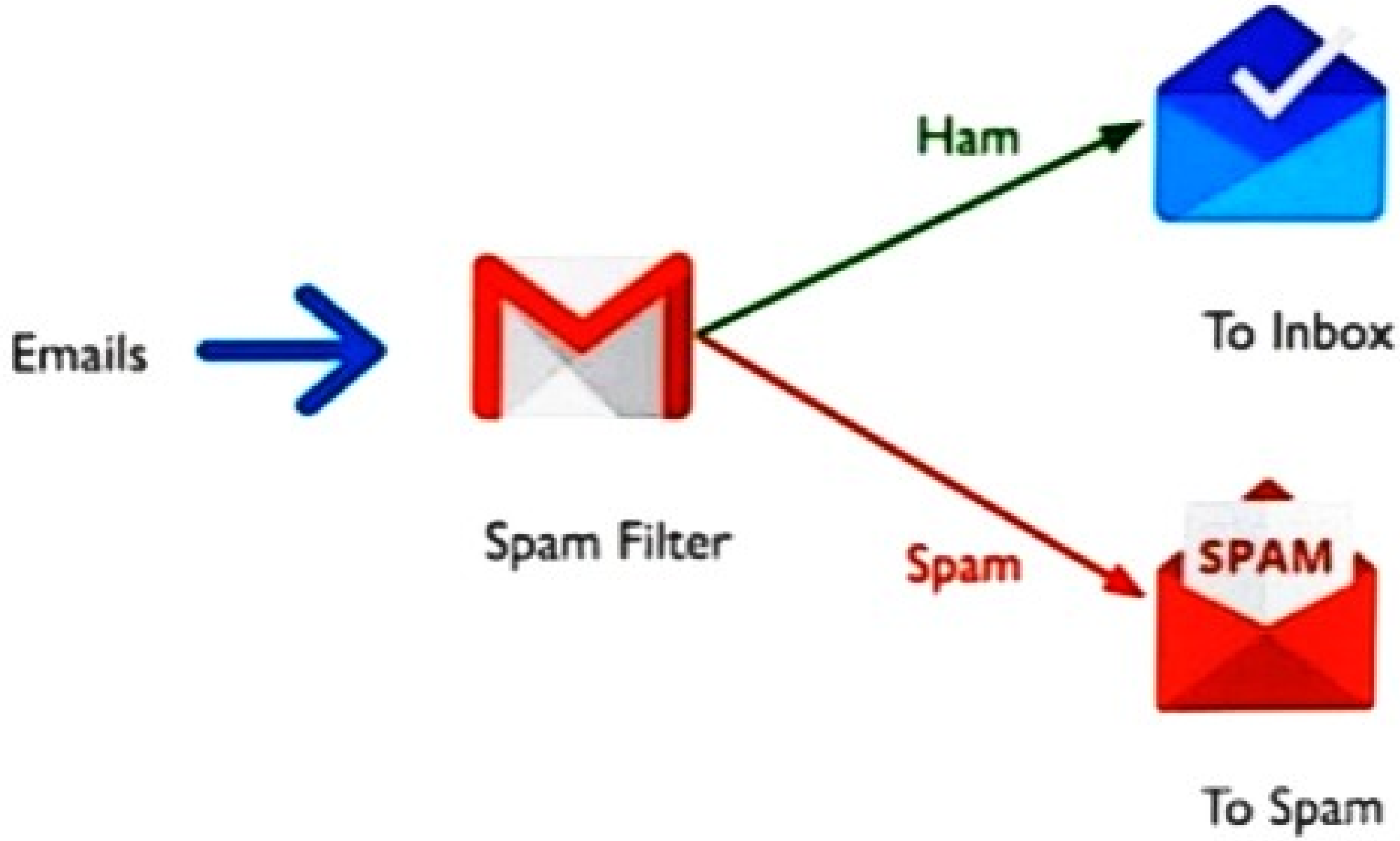
Scanned with AnyScanner

Introduction

In today's globalized world, email is a primary source of communication. This communication can vary from personal, business, corporate to government. With the rapid increase in email usage, there has also been increase in the SPAM emails. SPAM emails, also known as junk email involves nearly identical messages sent to numerous recipients by email. Apart from being annoying, spam emails can also pose a security threat to computer system. It is estimated that spam cost businesses on the order of $100 billion in 2007. In this project, we use text mining to perform automatic spam filtering to use emails effectively. We try to identify patterns using Data-mining classification algorithms to enable us classify the emails as HAM or SPAM

Scanned with AnyScanner



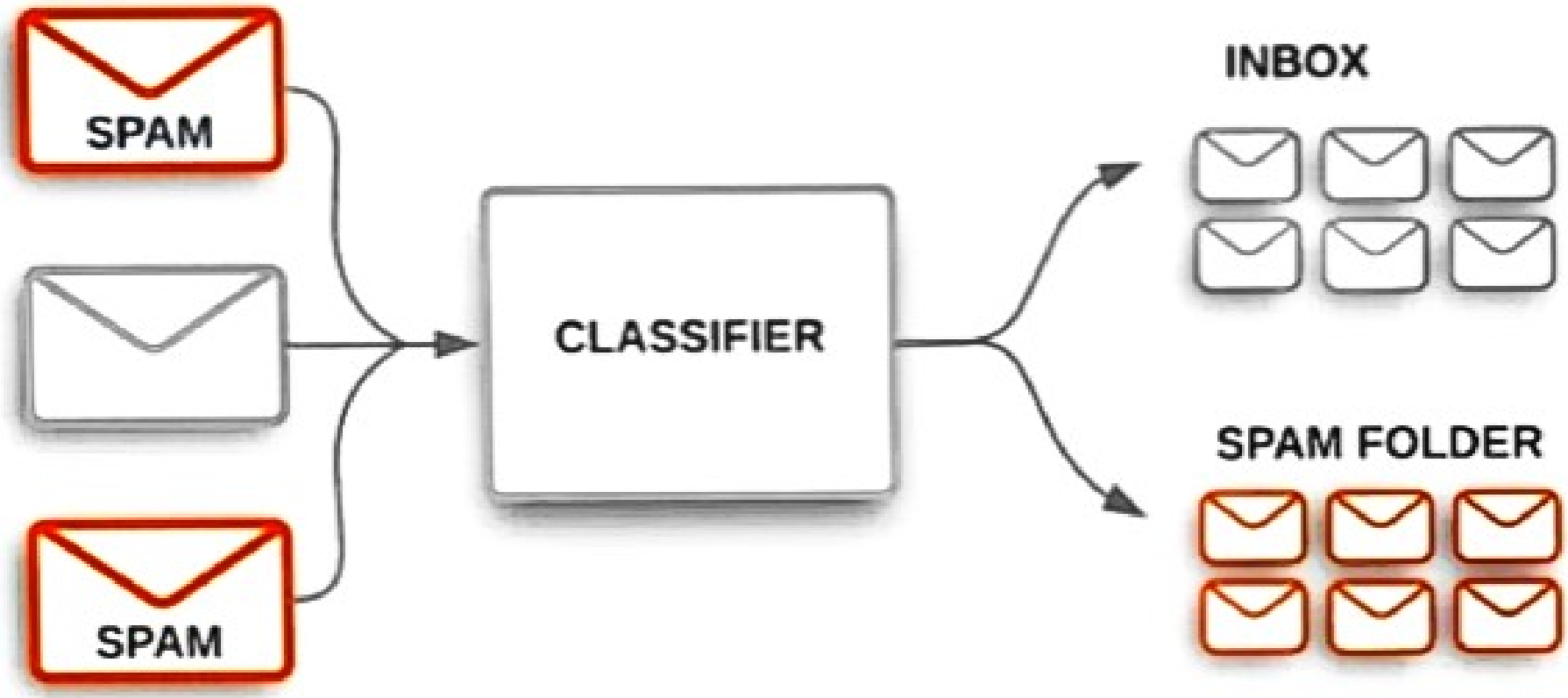
Scanned with AnyScanner

Algorithms

- Different algorithms used for email spam detection:-

1. Deep learning
2. Naive Bayes
3. Support Vector Machines
4. K-Nearest Neighbour
5. Rough Sets
6. Random Forests
7. Multinomial naive

Scanned with AnyScanner



Basic Image for spam filtration

Scanned with AnyScanner

Machine Learning

Arthur Samuel, an early American leader in the field of computer gaming and artificial intelligence, coined the term "Machine Learning " in 1959 while at IBM. He defined machine learning as "the field of study that gives computers the ability to learn without being explicitly programmed ".

- Machine learning is programming computers to optimize a performance criterion using example data or past experience .
- The field of study known as machine learning is concerned with the question of how to construct computer programs that automatically improve with experience.

Scanned with AnyScanner

```
# Import necessary libraries
import pandas as pd
from sklearn.model_selection import
train_test_split
from sklearn.feature_extraction.text
import TfidfVectorizer
from sklearn.naive_bayes import
MultinomialNB
from sklearn.metrics import
accuracy_score, confusion_matrix,
classification_report

# Load your dataset (assuming you
have a CSV file with 'text' and 'label'
columns)
data = pd.read_csv('spam_data.csv')

# Split the data into training and testing
sets
X = data['text']
```

Scanned with AnyScanner

```
y = data['label']
X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=0.2,
random_state=42)

# Text vectorization using TF-IDF
tfidf_vectorizer = TfidfVectorizer()
X_train_tfidf =
tfidf_vectorizer.fit_transform(X_train)
X_test_tfidf =
tfidf_vectorizer.transform(X_test)

# Train a Naive Bayes classifier
spam_classifier = MultinomialNB()
spam_classifier.fit(X_train_tfidf, y_train)

# Make predictions on the test data
y_pred =
spam_classifier.predict(X_test_tfidf)
```

Scanned with AnyScanner

```
# Evaluate the classifier
accuracy = accuracy_score(y_test,
y_pred)
confusion = confusion_matrix(y_test,
y_pred)
report = classification_report(y_test,
y_pred)

print(f'Accuracy: {accuracy}')
print('Confusion Matrix:')
print(confusion)
print('Classification Report:')
print(report)
```

Scanned with AnyScanner

```
Accuracy: 0.965
Confusion Matrix:
[[958   7]
 [ 26 134]]
Classification Report:
              precision    recall  f1-score   support

         ham       0.97      0.99      0.98       965
        spam       0.95      0.84      0.89       160

    accuracy                           0.97      1125
   macro avg       0.96      0.91      0.93      1125
weighted avg       0.97      0.97      0.97      1125
```

Scanned with AnyScanner

Conclusion

Spam is a major problem in today's world. Spam messages are the most unwanted messages the end user clients receive in our daily lives. Spam emails are available nothing but an ad for any company, any kind of virus etc. It will be too much. It is easy for hackers to access our system using these spam emails

Scanned with AnyScanner

Thank You

Scanned with AnyScanner