EMAIL SPAM DETECTION USING PYTHON & MACHINE LEARNING

PRESENTED BY:
Srividya Polneni
Sindhuja Yerramalla

PROBLEM STATEMENT

- These days, all official and sensitive communication is made through emails and Spam emails are a major issue on the internet. It is easy to send an email that contains spam messages by spammers.
- Spam fills our inbox with several irrelevant emails. Spammers can steal our sensitive information from our devices like files, and contact. Even though we have the latest technology, it is challenging to detect spam emails.

ABSTRACT

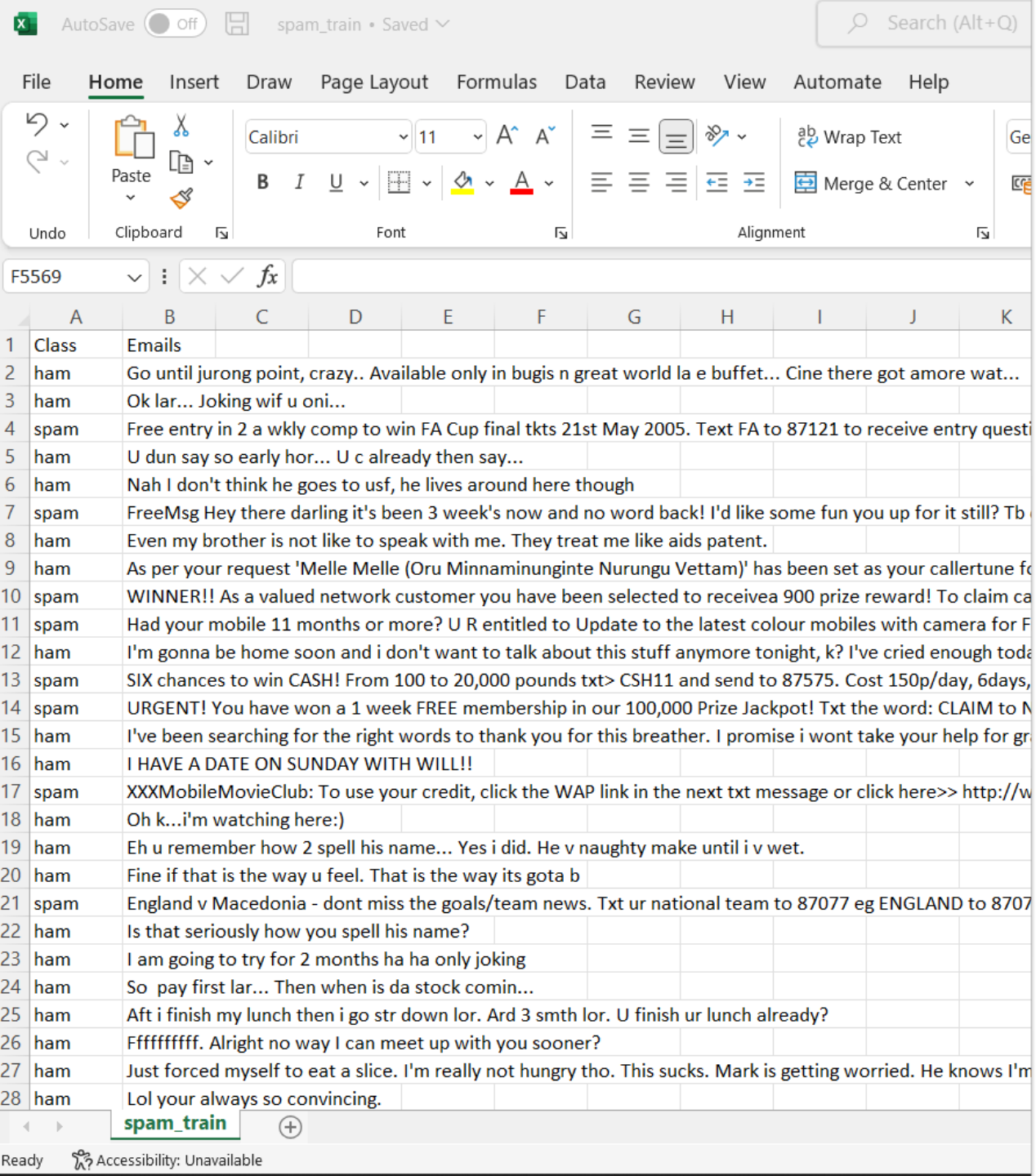
- Automatic e-mail filtering seems to be the most effective method for countering spam at the moment.
- Only several years ago most of the spam could be reliably dealt with by blocking e-mails coming from certain addresses or filtering out messages with certain subject lines.
- Spammers began to use several tricky methods to overcome the filtering methods like using random sender addresses and/or appending random characters at the beginning or the end of the message subject line.
- Machine learning techniques are being used to automatically filter spam e-mail at a very successful rate.

CONT..

- First, data collection and representation are mostly problem specific (i.e. e-mail messages), second, e-mail feature selection and feature reduction attempt to reduce the dimensionality (i.e. the number of features).
- Finally, the e-mail classification phase of the process finds the actual mapping between the training set and the testing set.
- Machine Learning approach includes lots of algorithms that can be used in e-mail filtering like Naïve Bayes, K-nearest neighbour, Support Vector Machine, and classifiers. In conclusion, we try to summarize the performance results of the few machine learning methods in terms of spam precision and accuracy.

IMPLEMENTED METHODS

- In this system, we are implementing Natural Language Processing (NLP) like **TF-IDF** is one of the simple and robust methods to understand the context of a text.
- Term Frequency and Inverse Document Frequency (TF-IDF) are used to find the related content and important words and phrases in a larger text. Implementing TF-IDF analysis is very easy using Python.
- Computers cannot understand the meaning of a text, but they can understand numbers. The words can be converted to numbers which is called feature extraction.
- Later these features are trained with machine learning techniques such as **Support Vector Machine, K-Nearest Neighbors, and Naive Bayes Classifiers** for Spam email detection.



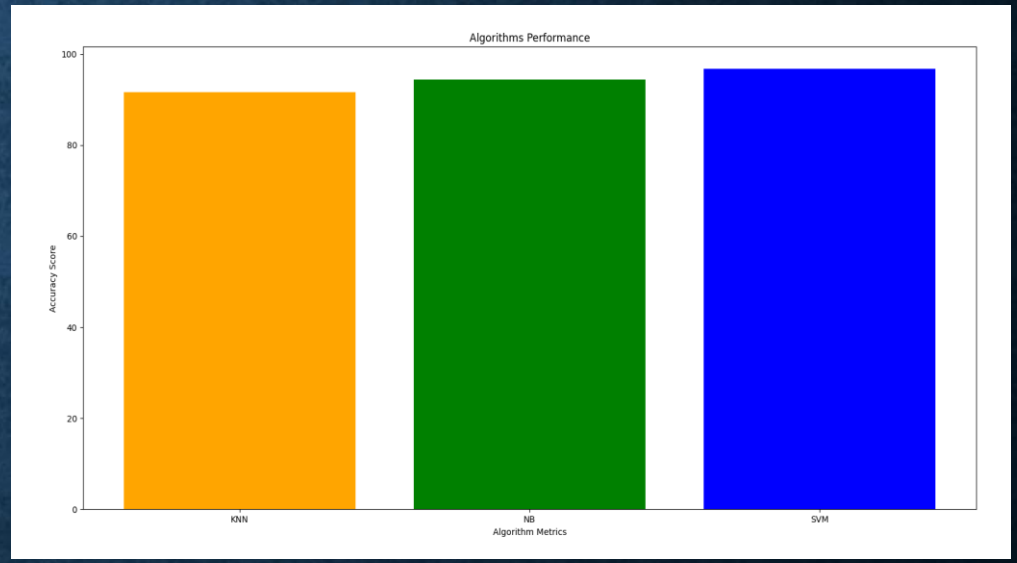
DATA SETS USED

- In this system, we are using the Email Spam Classification dataset which is accessed from the Kaggle repository https://www.kaggle.com/code/harshsinha1234/email-spam-classification-nlp/data.
- In this system, we are using the text format dataset which is not understood by the ML techniques. So, it is a challenge to convert the text format dataset into a numerical dataset using the NLP technique and ML algorithms
- This dataset contains multiple columns like text and spam. The **text** column will contain the email subjects and the **spam** column will contain 0 or 1. The '0' value indicates the **Ham** and the '1' value indicated the **Spam.**

RESULTS & CONCLUSION

Results

	Algorithm	Accuracy	Precision	Recall	F1_Score
1	KNN	91.63179916317992	63.703703703703695	47.175141242937855	51.343823979947075
2	NB	94.38135086670651	64.61739699149771	53.672316384180796	57.53936122357175
3	SVM	96.77226539151226	64.42606859556011	60.14813283726057	62.0542543588769



- So by the above results, we can conclude that the Algorithm performance is better in the SVM classification technique in all the features like Accuracy, Precision, Recall etc.,

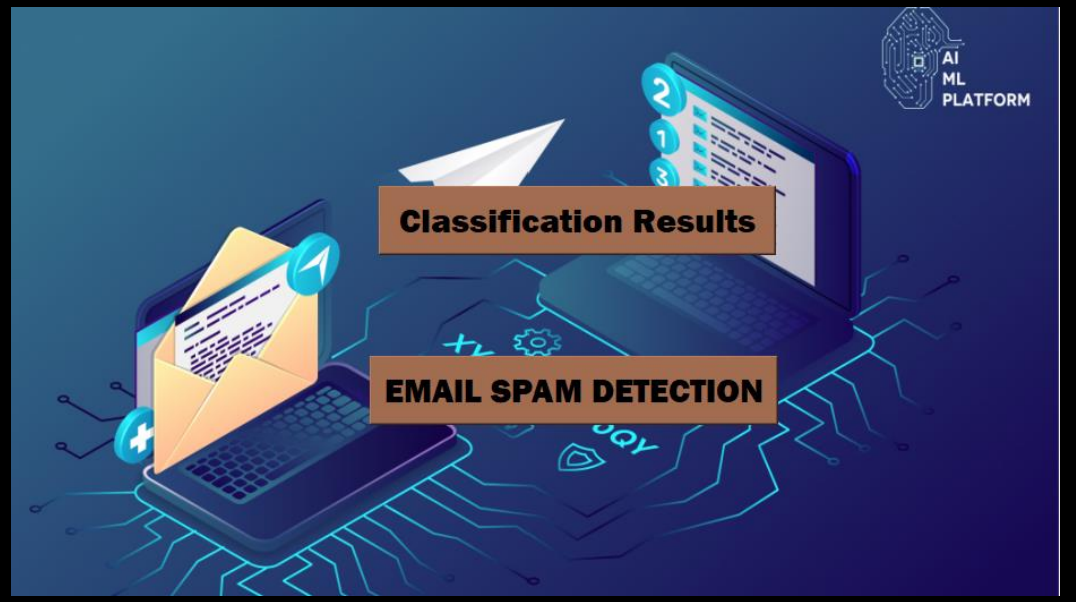
EMAIL SPAM DETECTION USING PYTHON & MACHINE LEARNING

Admin Login

User Name

Password

LOGIN



Enter Email Subject:

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 4403LDNW1A7RW18

Detection

Result : SPAM

Enter Email Subject:

Yup... Ok i go home look at the timings then i msg _ again... Xuhui going to learn on 2nd may too but her lesson is at 8am

Detection

Result : HAM

TEAM WORK!

- Contribution to work by each Individual

Team member: Srividya Polneni

- Dataset Selection
- Training Data

Team member: Sindhuja Yerramalla

- UI development
- Performance of Algorithms

THANK YOU!!