Email Spam Classification Using Machine Learning

Harshitha Pothula
&
Ssrk Kasyap

Contents

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.

TECHNOLOGIES

Technologies Used:

- **Python**: Python is an interpreted, object-oriented, high-level programming language with dynamic semantics developed by Guido van Rossum

Libraries:

1. Numpy
2. Pandas
3. Sklearn
4. Matplotlib

Libraries

- **NumPy:-** NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. Moreover, NumPy forms the foundation of the Machine Learning stack.
- **Pandas:-** Pandas is one of the tools in Machine Learning which is used for data cleaning and analysis. It has features which are used for exploring, cleaning, transforming and visualizing from data.
- **Sklearn:-** Sklearn is intended to support research and teaching in NLP or closely related areas, including empirical linguistics, cognitive science, artificial intelligence, information retrieval and machine learning.
- **Matplotlib:-** Matplotlib is a low-level library of Python which is used for data visualization. It is easy to use and elulates MATLAB like graphs and visualization. This library is built on the top of NumPy arrays and consist of several plots like line chart, bar chart, histogram, etc.

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.

Dataset

A machine learning dataset is a collection of data that is used to train the model. A dataset acts as an example to teach the machine learning algorithm how to make predictions. dataset as "a collection of data that is treated as a single unit by a computer". This means that a dataset contains a lot of separate pieces of data but can be used to train an algorithm with the goal of finding predictable patterns inside the whole dataset.

How to train the data?

- Al training data will vary depending on whether you're using supervised or unsupervised learning. Unsupervised learning uses unlabeled data. Models are tasked with finding patterns (or similarities and deviations) in the data to make inferences and reach conclusion.
- With supervised learning, on the other hand, humans must tag, label, or annotate the data to their criteria, in order to train the model to reach the desired conclusion (output)Labeled data is shown in the examples above, where the desired outputs are predetermined.

Problem Definition

- Short Message (SMS) and email has grown into a multi-billion dollars commercial industry.
- SMS spam is still not as common as email spam.
- SMS Spam is showing growth, and in 2012 in parts of Asia up to 30% of text messages was spam.

Algorithms

Different algorithms that can be used for Email Spam Detection are:

1. Deep Learning
2. Naive Bayes
3. Support Vector Machine
4. K-Nearest Neighbour
5. Random Forest
6. Multinomial naive

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.

THANK YOU