

Assignment Report - Machine Learning Laboratory

Team Leader: 20z214 - Dhanush Gowdhaman

Team Members: 20z201 - Aadhithyashri, 20z234 - Nashita V, 20z250 - Shivani Sri S, 20z261 - Vivekanand

Date: 07.02.2023

Problem Statement:

A Spam message is any unsolicited digital communication message that is sent in bulk, and is often malicious.. A Ham message is any message that is not Spam. In other words, they are legitimate messages. With the rise in frequency and complexity of spam attacks, it is becoming increasingly difficult for users to identify and protect themselves from these attacks. It is essential to have a solution to automatically classify SMS messages as either spam or ham to enhance the security and privacy of users.

Dataset description:

The dataset¹ for this project has been taken from Kaggle.com which is an online community that lets users collaborate and publish datasets, use notebooks and compete in challenges. The dataset available at Kaggle.com titled **SMS Spam Collection Dataset** comprises a corpus that has been collected from free or free for research purposes sources at the Internet. The collection of SMS messages have been tagged as spam or legitimate.

- 425 SMS spam messages from Grumbletext, a UK forum where users make public claims about SMS spam. The messages have been manually extracted from the website where users often don't include the very message that they are reporting and hence, the process of identifying the right messages for the dataset is challenging and time consuming as it requires close examination of many pages.
- 3,375 SMS ham messages (a subset of a dataset of about 10,000 messages) of the NUS SMS Corpus², a dataset of ham messages collected for research purposes. This corpus was collected by Tao Chen and Min-Yen Kan³. This data was gathered for research purposes at the Department of Computer Science at the National University of Singapore. The messages mostly came from Singaporeans, with a significant portion being students at the university. Participants were informed that their contributions would be publicly available before volunteering to provide the message.

- 450 SMS ham messages from Caroline Tag's PhD Thesis⁴.
- 1,002 SMS ham messages and 322 SMS spam messages from SMS Spam Corpus v.0.1 Big⁵.

The dataset is contained in a CSV file - **spam.csv** (503.66 kB). This file contains one message per line and each line has two columns - v1 and v2.

- v1 - This column contains the label that classifies the message as ham or spam.
- v2 - This column contains the raw text or message

Statistics of dataset:

- Ham - 87%
- Spam - 13%
- No. of unique values = 5169

Tools to be used

- **Coding Environment:** Google Colab/ VSCode
- **Language:** Python
- **Libraries:**
 - Scikit-learn - It is an open source library for Python that is built on top of NumPy, SciPy, and Matplotlib. It provides a user-friendly interface for performing tasks like classification, regression, clustering, and dimensionality reduction. It is favored for its ease of use and performance in MACHine Learning applications.
 - Pandas - It is an open source data manipulation and analysis library for Python that provides flexible data structures like dataframes and series, to make working with relational data easier. It's used to load data from sources like CSV, SQL, and is useful in working with large datasets
 - Numpy - It is an open source library for Python that is used for numerical computing and performs fast array-oriented operations.
 - Matplotlib - It is an open source data visualization library for Python that creates static, animated and interactive visualizations for a wide range of plots like bar charts, histograms, scatter plots, and more.

Challenges Faced

- A large, diverse dataset is required for our project for training as spam messages can have very different content from one another, which can be challenging to accurately identify all types of spam.
- Some spam messages try to evade spam filters by using words and phrases that resemble or are used in ham messages.
- To make sure the class distribution is not imbalanced, i.e., the number of spam messages are much smaller than the number of ham messages.

Contribution of Team Members

This table contains the work assigned to each member.

| Roll Number | Name | Contributions |
|--------------------|---------------|------------------------------|
| 20z201 | Aadhithyashri | Data Visualization |
| 20z214 | Dhanush | Model Creation |
| 20z234 | Nashita V | Data Preprocessing |
| 20z250 | Shivani Sri S | Model Creation (Alternative) |
| 20z261 | Vivekanand | Data Visualization |

References

1. Kaggle.com - SMS Spam Collection Dataset
<https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset>
2. Chen, T., Kan Min-Yen (2015-03-09). The National University of Singapore SMS Corpus. ScholarBank@NUS Repository. [Dataset].
<https://doi.org/10.25540/WVM0-4RNX>
3. Tao Chen and Min-Yen Kan (2013). Creating a Live, Public Short Message Service Corpus: The NUS SMS Corpus. Language Resources and Evaluation, 47(2)(2013), pages 299-355. URL: <https://link.springer.com/article/10.1007%2Fs10579-012-9197-9>
4. Caroline Tag's PhD Thesis - A CORPUS LINGUISTICS STUDY OF SMS TEXT MESSAGING. URL: <https://etheses.bham.ac.uk/id/eprint/253/1/Tagg09PhD.pdf>
5. SMS Spam Corpus v.0.1 Big. URL: <http://www.esp.uem.es/jmgomez/smsspamcorpus>
6. SMS Spam Corpus v.1. URL: <https://www.dt.fee.unicamp.br/~tiago/smsspamcollection/>
7. Suparna Das Gupta et al 2021 J. Phys.: Conf. Ser. **1797** 012017 - SMS Spam Detection Using Machine Learning. URL:
<https://iopscience.iop.org/article/10.1088/1742-6596/1797/1/012017/pdf>
8. SMS Spam Detection using Machine Learning Approach - Abhishek Patel, Priya Jhariya, SudalaguntaBharath, Ankita wadhawan. URL: <https://ijert.org/papers/IJCRT2104653.pdf>
9. NLP: Spam Detection in SMS (text) data using Deep Learning. URL:
<https://towardsdatascience.com/nlp-spam-detection-in-sms-text-data-using-deep-learning-b8632db85cc8>
10. Nilam Nur Amir Sjarif, Nurulhuda Firdaus Mohd Azmi, Suriyati Chuprat, Haslina Md Sarkan, Yazriwati Yahya, and Suriani Mohd Sam. 2019. SMS Spam Message Detection using Term Frequency-Inverse Document Frequency and Random Forest Algorithm. Procedia Comput. Sci. 161, C (2019), 509–515. URL:
<https://doi.org/10.1016/j.procs.2019.11.150>