Abstract

The fast growth of mobile networks has greatly enriched our life disseminating information and providing communications at any time and anywhere. However, at the same time, when people gather and exchange useful information, they also receive unwanted data and content, such as spam and unsolicited commercial advertisements. SMS (Short Message Service) spam is one typical example of unwanted content, which has caused a serious problem to mobile users by intruding on their devices, occupying device memories and irritating the users.

The easy accessibility and simplicity of Short Message Services (SMS) have made it attractive to malicious users thereby incurring unnecessary costing on the mobile users and the Network providers' resources.

SMS has reached more than 6 billion users globally with approximately 9.5 trillion SMS sent globally in 2009. This tremendous growth in mobile devices has made SMS a very attractive area to malicious organizations for carrying out illegal activities and influencing security risks such as SMS spam, Phishing, License to Kill Spyware, Malware, and privacy issues to mobile data.

Based on the statistics from the Singapore Police Force, from January till June 2020, the amount cheated through scams have increased by more than 8 million dollars

Contents

List of Figures

List of Tables

Chapter 1

Problem Definition

Short Message Service (SMS) is one of the well-known communication services in which a message sends electronically. Spam messages include advertisements, free services, promotions, awards, etc. People are using the ubiquity of mobile phone devices is expanding day by day as they give a vast variety of services by reducing the cost of services. Short Message Service (SMS) is one of the broadly utilized communication service. This has prompted an expansion in mobile phones attacks like SMS Spam. In this problem, preliminary results are mentioned or explained herein based on real time dataset(by extracting messages from phones and making dataset).

Chapter 2

Related Work

A weighted feature enhanced HMM for spam SMS filtering

Pre-process

In order to remove redundant information for better processing, every SMS message is first pre-handled toward the start. The pre-process incorporates tokenization and stops word expulsion.

Tokenization

Tokenization is to extract the words from an SMS message body. For some oriental languages, such as Chinese, Korean and Japanese, words are firmly associated without spaces in a sentence. In this manner, the SMS messages in these languages must be first partitioned into words and punctuations sequence through a division interaction.

SMS word weighting

SMS words might contain diverse semantic data and the SMS word weighting calculation can allocate a legitimate weight worth to each word. A word weight value demonstrates the shot at recognizing a spam or ham SMS message with the word's event. A word that conceivably happens in the spam set has a more modest negative weight and a word which in all likelihood shows up in the ham set has a greater positive weight. It indicates the HMM to think about additional significant words and be prepared unequivocally ideal.

SMS property prediction

While foreseeing the SMS property, the SMS messages in the testing set are first pre-handled as same as the preparing set. After tokenization and stop-word evacuation, the testing SMS messages become word arrangements with the first request. Then, at that point, these words acquire the word loads in the preparing set. They are subbed by the heaviness of a similar word in preparing a set. For the words that don't exist in the preparation set, known as obscure words, weight 0 is applied. So the word arrangements become weight successions which are the perception groupings. Individually, each weight grouping of the testing SMS messages takes care of the prepared HMM which deciphers spam or ham name arrangement for each testing SMS message thus.

Soft techniques for SMS spam classification: Methods, approaches and applications

Importance of spam detection

The continuous escalation of mobile devices over the years has given users an unbeatable communication experience which has increased users' performance efficiently. The most popular and widely used service of the Global System for Mobile communication (GSM) is the Short Message Services known as SMS. This fast and ever-growing service has reached more than 6 billion users globally with approximately 9.5 trillion SMS sent globally in 2009. This tremendous growth in mobile devices has made SMS a very attractive area to malicious organizations for carrying out illegal activities and influencing security risks such as SMS spam, Phishing, License to kill Spyware, Malware, and privacy issues to mobile data. So the SMS span detection is very important area.

Aim

The aim of this paper is to identify and review existing state of the art methodology for SMS spam based on some certain metrics: AI methods and techniques, approaches and deployed environment and the overall acceptability of existing SMS applications.

Content Based Approach

It involves the uses of words or character frequency mostly called bag of words for document representation. Here, the frequencies of words are used

as features within a classification method for detecting spamicity. Spamicity is the factor of the frequency of occurrence of the same words in the token database with assigned values to each word between the range 0.0 to 1.0. In Bag of words-based approaches, the sequence of words and their semantic relations are not considered.

Non Content Based Approach

On the contrary, non-content based approaches utilize certain message characteristics or signature patterns as features for detecting anomalies within a network. These features could be based on a static measure (total number of messages sent per time); the size of the message and time stamp.

Hybrid Method

Hybrid approaches combine features from content and non-content based approaches for classification purposes.

Architecture

Based on the 83 selected studies, existing SMS spam filter architecture is based on three main layers which are: Client-based (solution resides on the mobile device), Server-based (solutions reside at the network provider's side, SMSC) and collaborative-based (solutions reside on both the client and the server layer). Quite a number of the selected studies solutions were based at the client side.

Dataset

The issue of ensuring user's privacy is a major factor affecting SMS data collection globally. Most of the existing English SMS corpus is based on reusability and compilation of smaller SMS databases to build more robust databases, examples of some databases formed from smaller ones include UCI corpus (comprises of NUS corpus, Caroline Tag and Grumble text), British English Corpus (comprises of Caroline Tag and Grumble text corpus).

Discussion, limitations and taxonomy

This survey emphasizes the overall research contributions on SMS spam detection and classification while also presenting a summary of existing studies based on their state-of-the-art-methodologies, approaches, architectures, status, SMS databases and existing anti-spam solutions. Classification of short messages is vital in ensuring that the users' security is preserved since mobile devices are important tools for daily activities. The significance of SMS spam classification is numerous. Based on the selected publications reviewed in this study, a lot of work has been done to solve the SMS spam problem using machine learning. However, some methods are yet to be fully explored such as Deep learning. This method is being maximized in areas like text mining, image processing, pattern recognition, etc. Based on the selected publication, this survey presents an overall summary of the pros and cons of the different areas involved in SMS spam classification as shown in Table above.

Conclusion and future direction

The analysis result obtained for each search strategies are as follows: Existing methods on SMS spam shows that Machine learning (ML), Statistical analysis and evolutionary methods are 49

Chapter 3

Requirements

The design of this project contains both hardware and software. The specifications are listed below.

Hardware

Leveraging the power of cloud computing, we have used Python 3 Google Compute Engine for the processing power. The provisioned server has 12GB Ram and 108GB of storage. the storage was used to upload the dataset and work on it.

Software

The algorithm was implemented in Google Collab in python using multiple libraries. The libraries include

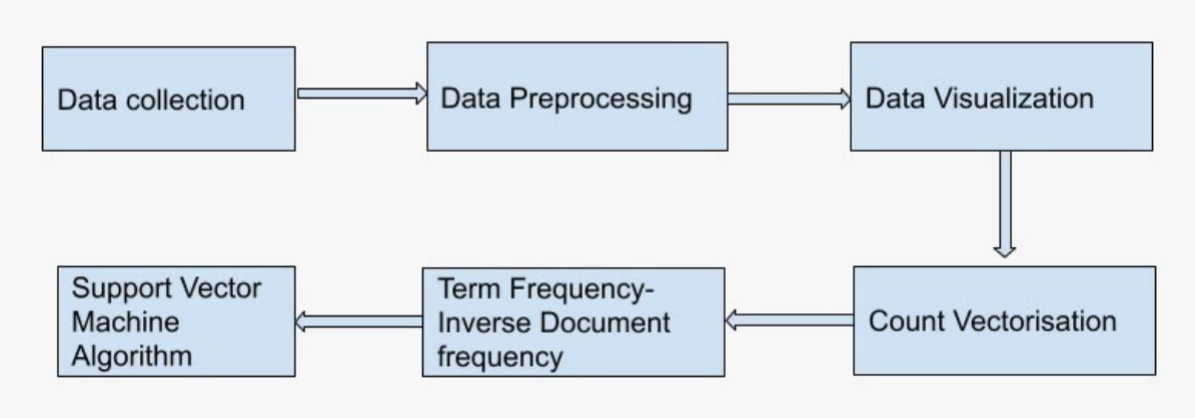
1. NUMPY
2. PANDAS

3. NLTK
4. SKLEARN
5. MATPLOTLIB
6. SEABORN

Chapter 4

Proposed System

Flow Chart



Data Collection

Data collection is an important step in the process of spam classification. Many datasets were available throughout the internet, but most of them were conversations from the US and UK. There were no datasets that are concerned with Indian SMS and how things work in India. As we all know in India there are a variety of ways to spam people through SMS. We get

random messages saying that we have won 10,000/- and to claim it we need to click a link or we are pre-approved for a loan. Taking into account all the types of spam messages we have curated a dataset keeping the Indian scenario in mind. This is a unique dataset and created from scratch.

Preprocessing

Preprocessing is a data mining technique that transforms raw data into an understandable format. Real-world data is always incomplete and that data cannot be sent through a model. That would cause certain errors. That is why we need to preprocess data before sending it through a model.

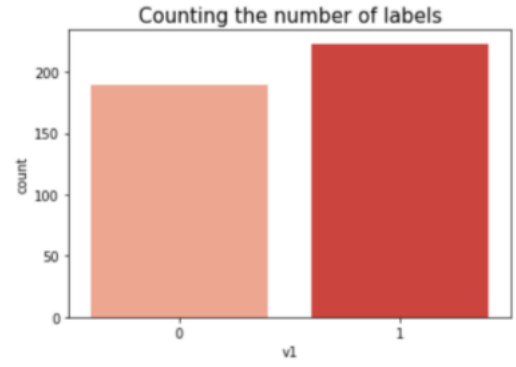
The first step is to find if there are any missing values. Since we are the ones that created the dataset, there were no missing values. In this project we are dealing with language, so we applied some natural language pre-processing techniques like removal of punctuation and stop words removal. Punctuations like full stops and commas act as noise to the model, so it is advised to remove punctuations. Words like “a”, “the”, “is”, “are” also act as noise to the model. In NLP this comes under a section known as stop words. NLTK has a predefined set of words called stop-words. The data is iterated over and the stop words are removed to avoid noise for the model.

In pre-processing, we also decide the target variable for the model. The dataset contains 3 columns namely spam/ham, message and length. We set the spam/ham as the target variable and renamed the columns as v1, v2 and length.

4.1 Data Visualisation

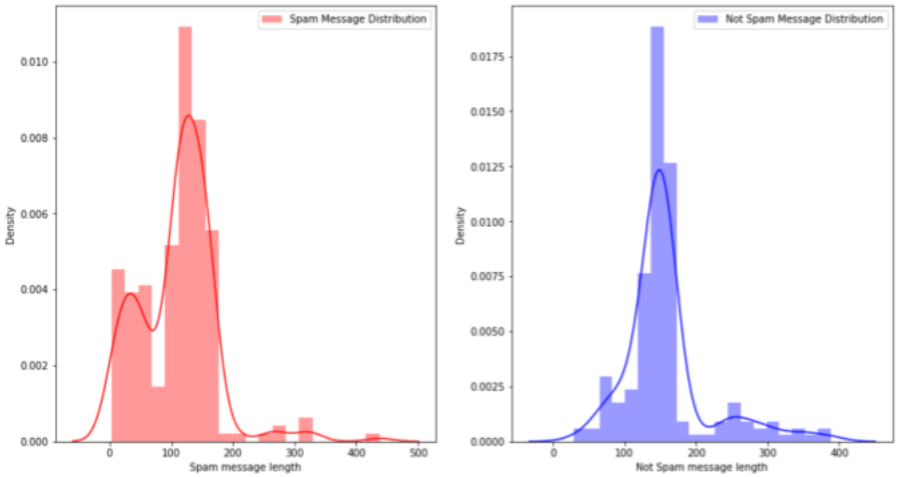
Data visualization helps to tell stories by curating data into a form easier to understand, highlighting the trends and outliers. A good visualization tells a story, removing the noise from data and highlighting useful information. Effective data visualization is a delicate balancing act between form and function.

1. The first step is to see the distribution of the target variables. This gives us a fair idea of the no of spam messages and no of not spam messages. this gives us a picture of how evenly the messages are distributed and hints about the performance of the dataset.




```
1    223
0    189
Name: v1, dtype: int64
```

2. The length of a message gives a reading on the correctness of a message. a message too short containing a link is often spam, or a message too long persuading you into a product are also spam. In the below figure, we have plotted 2 graphs showing the length of the message vs no of messages with that length.



3. The next data visualization technique is the word cloud. This gives a rough estimate of the sentiment in the messages. Which words occur the most and which the least. As we can see most of the words with the highest occurrence are related to money and transactions and debit cards and OTP. This gives us a fair understanding of what are the contents of the dataset.

Data Split

Splitting data into training data and testing data is one of the most important steps. If omitted it can lead to overfitting or underfitting. Overfitting is a scenario where the error rate on training data is very small, but the model fails to generalize on the test data. This leads to a scenario where the error rate on training data is less and the error rate on test data is very high. On the other hand, we have underfitting where the model is not able to capture the patterns in the data.

Count Vectorization

Count Vectorization is a great tool which is used to transform the given text into a vector on the basis of the frequency count of each word that occurs in the entire text.

Key Observations

1. There are 1913 unique words in the document, represented as columns of the table.

2. There are 412 text samples in the document, each represented as rows of the table.

3. Every cell contains a number, that represents the count of the word in that particular text.

4. All words have been converted to lowercase.The words in columns have been arranged alphabetically

TF-IDF

TF-IDF is a method which gives us a numerical weightage of words which reflects how important the particular word is to a document in a corpus. Tf is Term frequency, and IDF is Inverse document frequency. Tf(Term Frequency): Term frequency can be thought of as how often does a word 'w' occur in a document 'd'. IDF(inverse document frequency): Sometimes, words like 'the' occur a lot and do not give us vital information regarding the document. To minimize the weight of terms occurring very frequently by incorporating the weight of words rarely occurring in the document. Combining these two we come up with the TF-IDF score.

SVM

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised

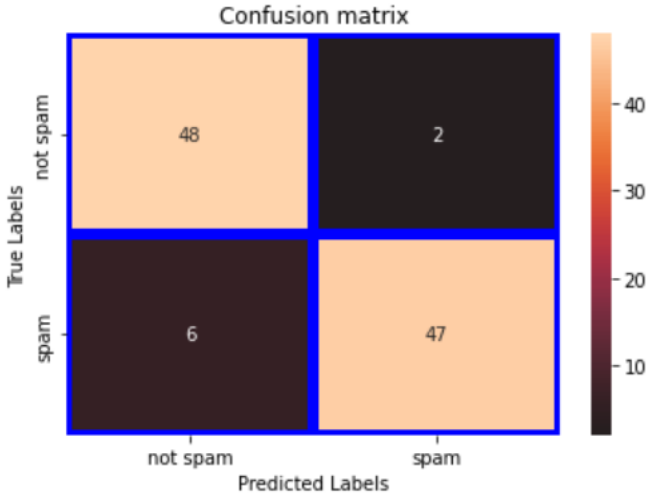
learning), the algorithm outputs an optimal hyperplane which categorizes new examples.

Chapter 5

Result and Analysis

SMS is one of the most significant communication methods between people, where SMS messages are two types: ham and spam, spam message are the undesirable messages that must be removed or blocked before the user receiving them. Therefore, in this research, machine learning algorithm were used in order to improve SMS spam detection process using count vectorization where we have transformed the vector on basis of frequency count of each word that exists in the entire text and based on Tf and IDF also used to caluculte the TF-IDF score and these result were given to the support vector machine and the output is predicted as spam or ham. The support vector machine algorithm gave the results for ham presicion, recall , f1-score ,support as 89%, 96%, 92%,50% and for spam presicion , recall , f1-score , support as 96% ,89%, 92%,53%. and the accuracy of 92%.

	precision	recall	f1-score	support
0	0.89	0.96	0.92	50
1	0.96	0.89	0.92	53
accuracy			0.92	103
macro avg	0.92	0.92	0.92	103
weighted avg	0.93	0.92	0.92	103



Chapter 6

Conclusion

The proposed problem statement meets the objective by achieving the accuracy of 92% using count vectorization and support vectorization algorithm. The future work that we can do is by adding more indian specific data set so that algorithm predicts more accurately and more powerful machine learning algorithm.

References

[1] https://doi.org/10.1016/j.engappai.2019.08.024.

[2] https://doi.org/10.1016/j.future.2014.06.010

[3] https://doi.org/10.1016/j.procs.2017.08.335

[4] https://doi.org/10.1016/j.procs.2017.08.335

Appendix A

Source code

Collab Link