Spam Classifier

Shreya Hunur[1], *Rohit Yadav*[2], *Hashmitha Katta*[3]

[1]San Jose State University
[2]San Jose State University
[3]San Jose State University

shreyanagaraddy.hunur@sjsu.edu, rohitkumar.yadav@sjsu.edu, hashmitha.katta@sjsu.edu

Abstract

Because of its global accessibility, relatively quick message transfer, and low sending cost, SMS is one of the most popular and widely used modes of communication. With technological advancements and an increase in content-based advertising, the use of Short Message Service (SMS) on phones has increased to the point where devices are occasionally flooded with many spams SMS. Spam is a term that is commonly used to describe junk or unsolicited messages or information. As a result, a spam SMS is any junk message delivered to a mobile device via text messaging. The danger is the same whether the spam is sent via email or SMS. Spam may result in the disclosure of personal information, invasion of privacy, or unauthorized access to mobile device data.

Most Internet users openly despise spam, but enough of them respond to commercial offers that spam remains a viable source of income for spammers. While most users want to do the right thing to avoid and eliminate spam, they require clear and simple guidelines on how to behave. Despite all the measures taken to eliminate spam, it has not been completely eradicated. Furthermore, if the countermeasures are overly sensitive, even legitimate SMS will be deleted.

Index Terms: speech recognition, human-computer interaction, computational paralinguistics

1. Introduction

Users now have personal and confidential information such as contact lists, credit card numbers, photographs, passwords, and much more stored in their smartphones, making them vulnerable to cyber-attacks via spam SMS. This allows hackers engaged in unethical activities to access smartphone data without the end-knowledge, user's jeopardizing the user's privacy. This could result in both monetary and functional losses. Spam messages appear to be on the rise, causing not only annoyance but also critical data loss for some users. Aside from that, SMS spam can be a driving force for malware and keyloggers.

The widespread distribution of this type of problem, as well as less effective security control measures, has inspired many researchers to develop a set of techniques to assist simple formula, giving them an advantage over other techniques in particular sense. We intend to compare various classifying techniques on various datasets gathered from previous research works and evaluate them based on their accuracies, precision, recall, and CAP Curve. A comparison of traditional machine learning techniques and deep learning methods will be carried out.

2. Methods

2.1. Pre-Processing

2.1.1. Data Cleaning :

Data cleaning is very crucial process in NLP. In this only alphabetic characters are extracted from text removing the punctuation and numbers, and then all characters are converted into lowercase. This cleaned text will be then used in further processing.

2.1.2. Tokenization :

Tokenization is breaking the raw text into small chunks. Tokenization breaks the raw text into words, sentences called tokens. These tokens help in understanding the context or developing the model for the NLP. The tokenization helps in interpreting the meaning of the text by analyzing the sequence of the words.

2.1.3. Removing stop words :

Stop words are a set of commonly used words in a language. Examples of stop words in English are "a", "the", "is", "are" and etc. Stop words are commonly used in Text Mining and Natural Language Processing (NLP) to eliminate words that are so commonly used that they carry very little useful information.

2.1.4. Stemming and Lemmatization

The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. However, the two words differ in their flavor. Stemming usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes. Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma .

2.1.5. Bag of Words

A bag-of-words is a representation of text that describes the occurrence of words within a document. It involves two things: A vocabulary of known words. A measure of the presence of known words.

2.1.6. TF-IDF

TF-IDF stands for "Term Frequency — Inverse Document Frequency". This is a technique to quantify words in a set of documents. We generally compute a score for each word to signify its importance in the document and corpus. This method is a widely used technique in Information Retrieval and Text Mining.

TF-IDF = Term Frequency (TF) * Inverse Document Frequency (IDF)

Terminology

t — term (word)

d — document (set of words)

N — count of corpus

corpus — the total document set

2.2. Model Building:

1. Build a model from vectors obtained from TFIDF
2. Split dataset into train and test data
3. Train model using train data
4. Validate model using test data
5. Calculate performance metrics

3. Comparisions

4. Example Analysis

5. Conclusions

6. Acknowledgements

7. References