

Sri Lanka Institute of Information Technology

Information Retrieval and Web Analytics – IT

Assignment 2

Sentiment Analysis – Spam Identification

Group Details

1. D. G. Y. C. K. Alwis – IT17181402
2. W. M. N. A. Fernando – IT17175012
3. R. M. M. G. Rathnayake – IT17182188

Introduction

What is Spam?

Spam is electronic junk mail or junk newsgroup postings. Some people define spam even more generally as any unsolicited [email](https://www.webopedia.com/TERM/E/e_mail.html). However, if a long-lost brother finds your [email address](https://www.webopedia.com/TERM/E/e_mail_address.html) and sends you a message, this could hardly be called spam, even though it is unsolicited. Real spam is generally email advertising for some product sent to a [mailing list](https://www.webopedia.com/TERM/M/mailing_list.html) or [newsgroup](https://www.webopedia.com/TERM/N/newsgroup.html). – source ( <https://www.webopedia.com/TERM/S/spam.html> )

Why it is called spam?

There is some debate about the source of the term, but the generally accepted version is that it comes from the Monty Python song, "*Spam spam spam spam, spam spam spam spam, lovely spam, wonderful spam*". Like the song, spam is an endless repetition of worthless text. Another school of thought maintains that it comes from the computer group lab at the University of Southern California who gave it the name because it has many of the same characteristics as the lunch meat Spam:

* Nobody wants it or ever asks for it.
* No one ever eats it; it is the first item to be pushed to the side when eating the entree.
* Sometimes it is tasty, like 1% of junk mail that is useful to some people.

The term spam can also be used to describe any "unwanted" email from a company or website -- typically at some point a user would have agreed to receive the email via subscription list opt-in -- a newer term called [graymail](https://www.webopedia.com/TERM/G/graymail.html) is used to describe this particular type of spam.

Spam vs Ham

Nowadays, it's likely that everyone knows what [Spam](https://cwiki.apache.org/confluence/display/SPAMASSASSIN/Spam) means, in the context of e-mail. The use of the word "Ham", on the other hand, is relatively new and sometimes confusing.

"Ham" is *e-mail that is not*[*Spam*](https://cwiki.apache.org/confluence/display/SPAMASSASSIN/Spam). In other words, "non-spam", or "good mail". It should be considered a shorter, snappier synonym for "non-spam".

Its usage is particularly common among anti-spam software developers, and not widely known elsewhere; in general, it is probably better to use the term "non-spam", instead.

Dataset specifications:

Dataset Name: dataset.csv

Number of records: 10881

Description: Two columns which has email and class (spam or ham)

Used Models: 1. Naive Bayes Model

2. Support Vector Machine

3. K-NN Model

Group Contribution

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Name | Data analysis (finding, collecting data) | Data refinement | Building model – Gaussian Classifier | Building Model – SVM classifier | Building Model – K-NN classifier | Summary |
| D. G. Y. K. Alwis |  | Checkmark |  |  | Checkmark | Checkmark |
| R. M. M. G. Rathnayake | Checkmark | Checkmark |  | Checkmark |  | Checkmark |
| W. M. N. A. Fernando | Checkmark | Checkmark | Checkmark |  |  | Checkmark |

Summary

This report explains about three scores represented by text classification models such as Naive Bayes, Support Vector Machines, K-Nearest Neighbors. Those models scored between 73% - 74%. Although Naive Bayes model has scored slightly high. Therefore, I recommend Naive Bayes algorithm is the best model to spam email identification.

We faced some difficulties in data cleaning process because of several reasons. Such as modern words, words do not include in dictionary…. etc. Failures of data cleaning process can affect for accuracy of the models. So, it’s the reason for accuracy getting lower.