**Spam SMS Detection**

**Student Name**

**Institution of Affiliation**

Table of Contents

[Abstract 2](#_Toc84583683)

[Objective: 3](#_Toc84583684)

[**Challenges:** 3](#_Toc84583685)

[Dataset: 3](#_Toc84583686)

[Approach and methodology: 3](#_Toc84583687)

[Preprocessing: 3](#_Toc84583688)

[Feature vectors: 4](#_Toc84583689)

[Classification: 4](#_Toc84583690)

[Results of Investigations / Tests: 4](#_Toc84583691)

[References 5](#_Toc84583692)

**Abstract:** As we've seen, spam email has expanded tremendously in recent years. Additionally, spam on Instant Messaging systems (so-called SPIM) and Short Message Service (SMS), as well as mobile spam, is becoming a big issue.

SMS spam can be addressed in the same manner that email spam is addressed: through regulation, economics, or technology. A range of technical solutions, including Bayesian filters, are used to combat email spam. We investigate how Bayesian filtering techniques used to combat spam in email can be used to detect and block mobile spam. We've built two large SMS spam test data sets, one in English and one in Spanish. We tested various message representation methodologies and Machine Learning algorithms on them to see how well they worked. [3] Our findings suggest that Bayesian filtering algorithms can also be used to detect SMS spam.

# **Objective:**

Identifying spam and ham text messages/sms (non-spam).

# **Challenges:**

Spam SMS databases, in contrast to spam emails, are extremely small.

The length of SMS messages is limited, as is the number of qualities that can be classified.

Text messaging frequently uses abbreviations, casual language, and other written languages in the language of English[5].

# **Dataset:**

The UCI Machine Learning Repository has a dataset of spam SMS messages called the SMS Spam Dataset. This is what's included:

Spam messages gathered from the Grumbletext website totaling 425

• 3375 ham messages from the NUS SMS Corpus were selected as a subset.

450 ham radio communications are part of Caroline Tag's Ph.D. thesis.

There were 4827 ham messages and 747 spam messages totaling 5574 in this batch.

There is a message on each line in the collection. Each line begins with a ham/spam designation, followed by a tabspace (\t).

# **Approach and methodology:**

## **Preprocessing:**

70 percent of the dataset is used for training, while the other 30 percent is used for testing.

Each piece of training data is examined one at a time, with the label and message being separated after each read.

Tokenization has been used to make the communication more digestible. When it comes time to sort them, only the alphabetical tokens will be saved. As a sign of disrespect, punctuation and special characters are neglected. Token Frequency Count is stored in Bag of Words.[1]

• There isn't any word-stemming going on in this paragraph.

The number of dollar signs, numeric strings, and message length are all kept as additional message properties.

This is a great feature because spam generators aim to use as much of the message as possible without wasting it.

Tokens that appear less than five times or more than 500 times are now removed. This helps to normalize the features by removing offensive words.

In total, this step generates 1502 + 3 = 1505 distinct properties.

## **Feature vectors:**

The frequency counts are used as values in the feature vectors, which are then built for each message in turn. Each of the three extra properties is handled as a feature, and its value is assigned to every message.

## Classification:

In this instance, the Naïve Bayes classifier is used.

For feature vectors containing word counts, the NB multinomial event model is appropriate.

When the feature vectors' word counts are zero, Laplace Smoothing is used in NB.

**Results:**

It only takes one second to train a classifier on an i3-2330M with 6GB DDR3 RAM.

• Testing is much quicker.

• The classifier's accuracy was tested using 30% of the training data and came out at 97.43 percent.

# **Results of Investigations / Tests:**

In order to categorize the smsCorpus en.xml file, we trained the classifier using all of the training data. There are 65,257 messages in all.

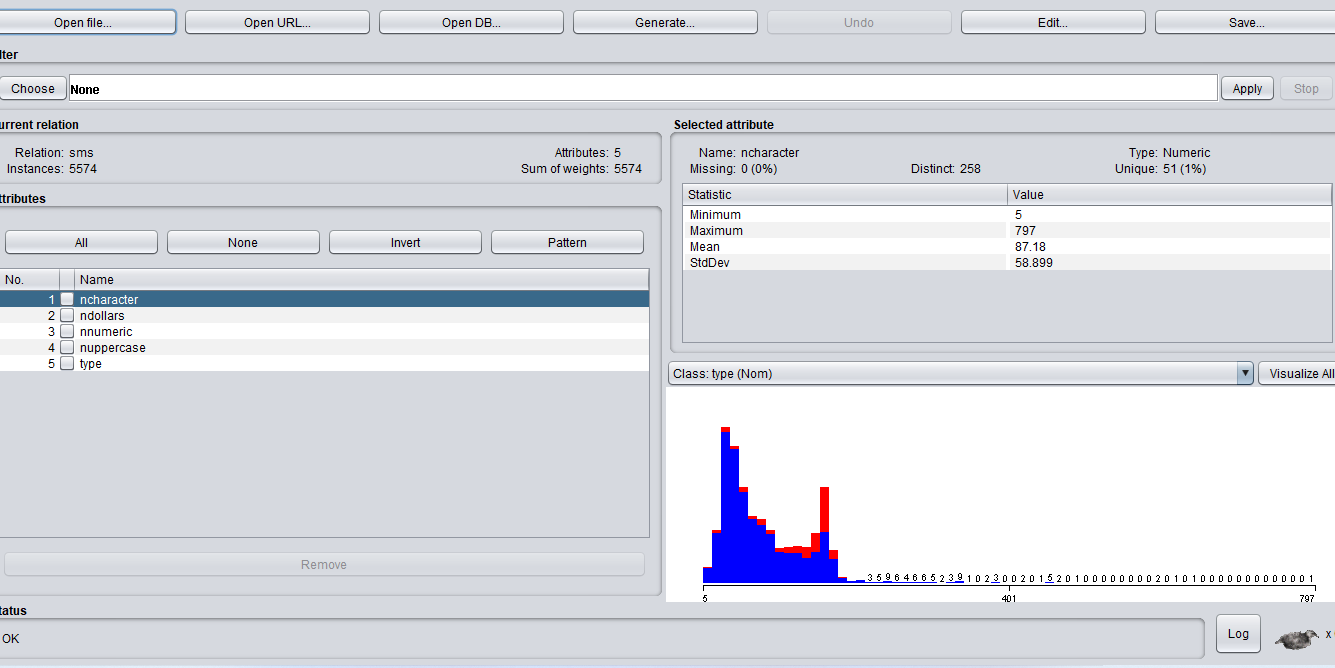
In the corpus, some messages appeared numerous times. It found 636 spam and 47,345 ham messages using this algorithm.

It was possible to get false positives (ham classified as spam) and false negatives (spam labeled as ham). However, false positives outnumbered false negatives in terms of frequency.

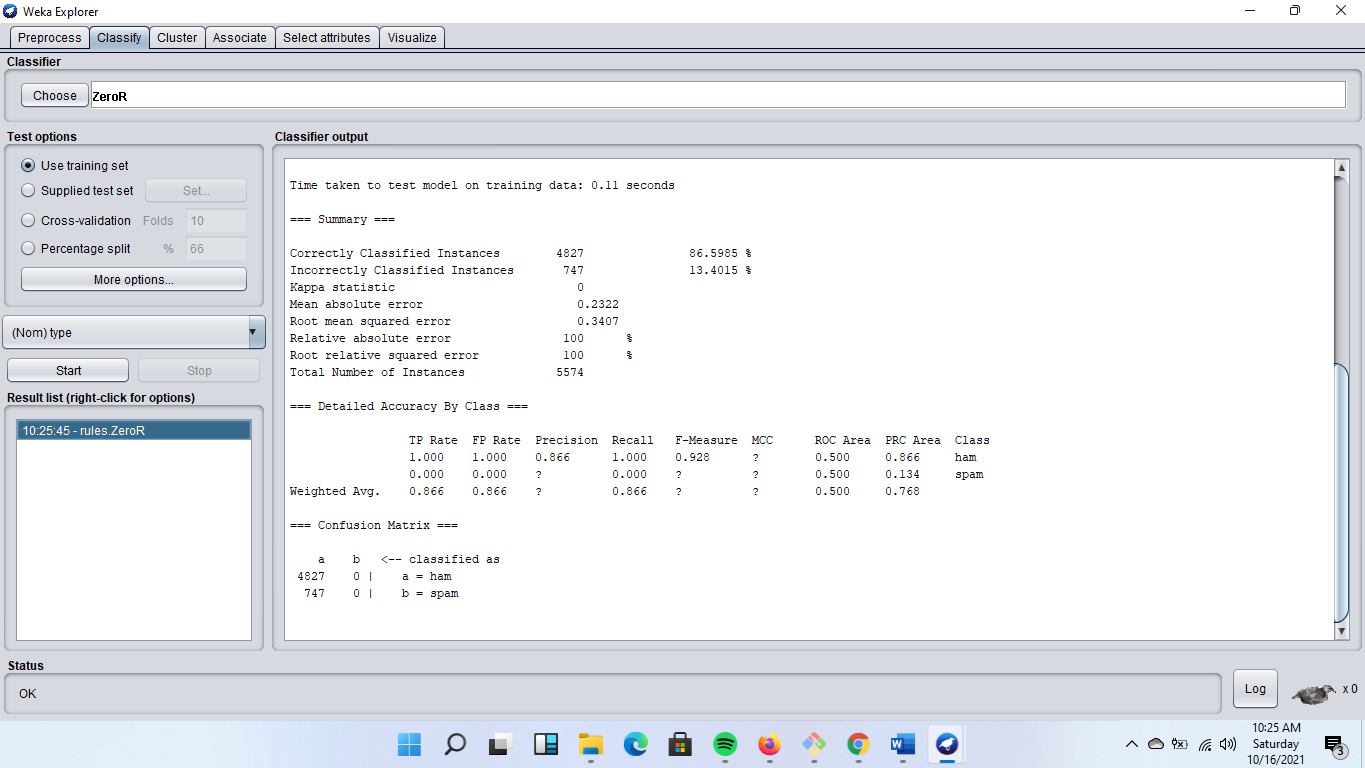
This is because the classifier connected the label "spam" with variables such as message length, numeric terms, and word features such as "call," "urgent," and so forth.

Message content that incorporated other language elements (hindi, Chinese) typed in English and mixed with conventional English text caused false positives and negatives. However, laplace smoothing can handle some of these words because they aren't visible in the training data.

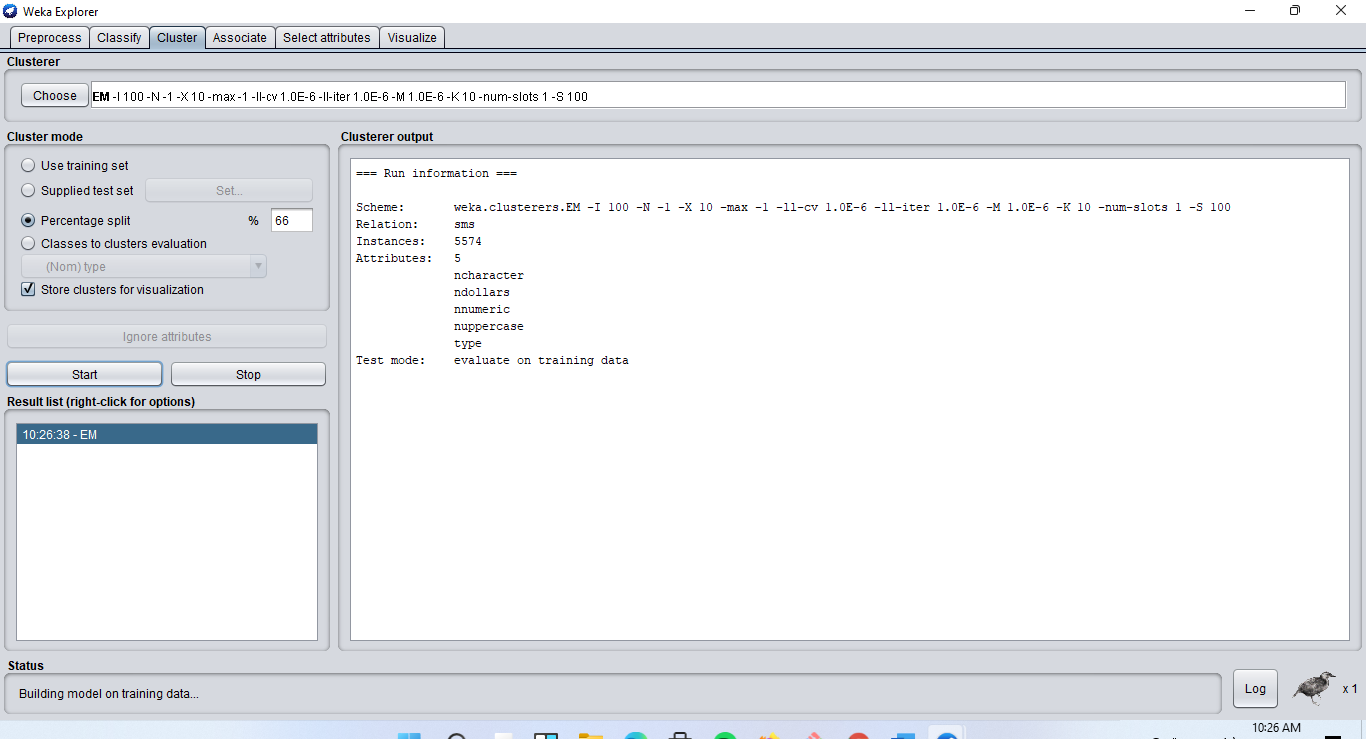
Adding elements such as the ratio of capital to tiny words, spam terms to normal words, the consideration of various currency symbols, and the number of URLs in a message might boost the scenario's accuracy.



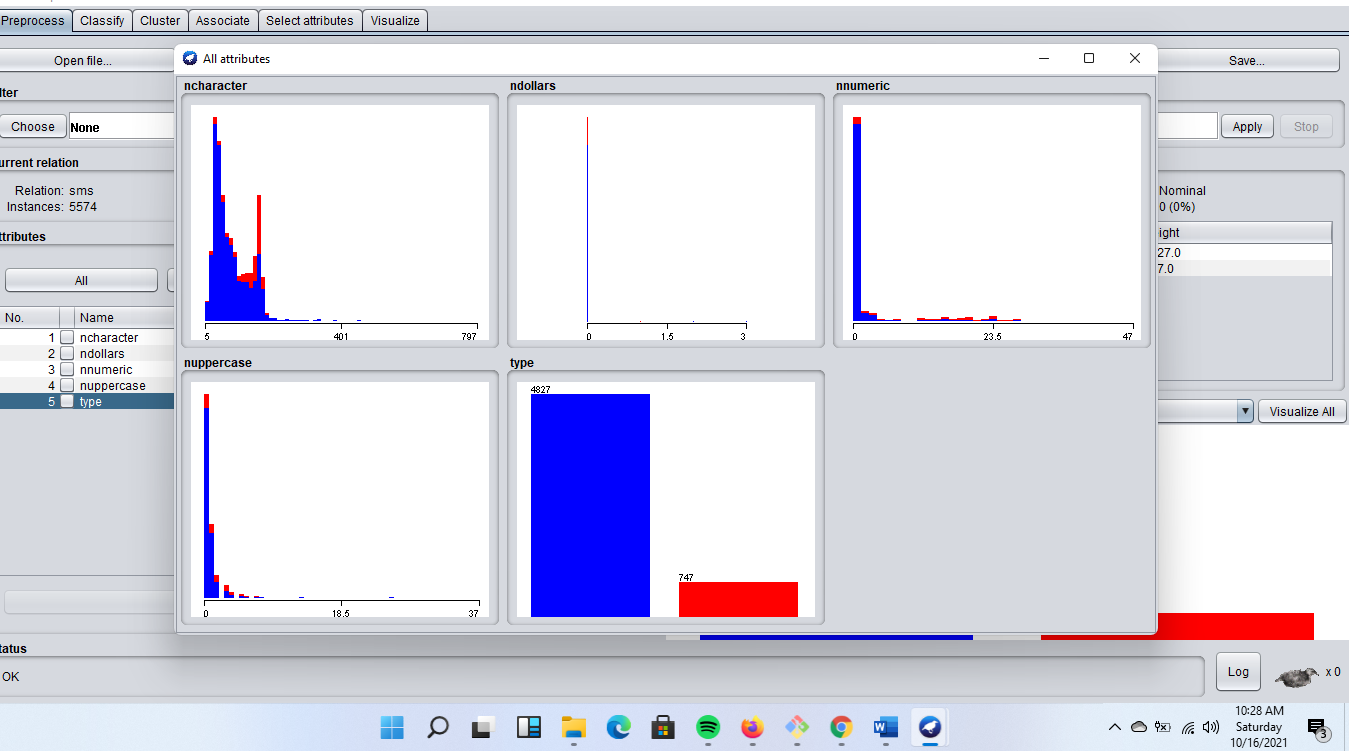
Classifying the data



Training the dataset



Visualization of the data



# **References**

[1] Alzahrani, A., & Rawat, D. B. (2019, April). Comparative Study of Machine Learning Algorithms for SMS Spam Detection. In *2019 SoutheastCon* (pp. 1-6). IEEE.

[2] Agboola, O. S. (2020). Spam Detection Using Machine Learning.

[3] Oluwatoyin, O., Bodunde, A., Titus, G., & Ganiyu, A. (2019). An Improved Machine Learning-Based Short Message Service Spam Detection System. *International Journal of Computer Network & Information Security*, *11*(12).

[4] Kawade, K. O., & Oza, K. S. (2018). Content-based SMS spam filtering using machine learning technique. *International Journal of Computer Engineering and Applications*, *7*, 4.

[5] Poomka, P., Pongsena, W., Kerdprasop, N., & Kerdprasop, K. (2019). SMS spam detection based on long short-term memory and gated recurrent unit. *International Journal of Future Computer and Communication*, *8*(1).