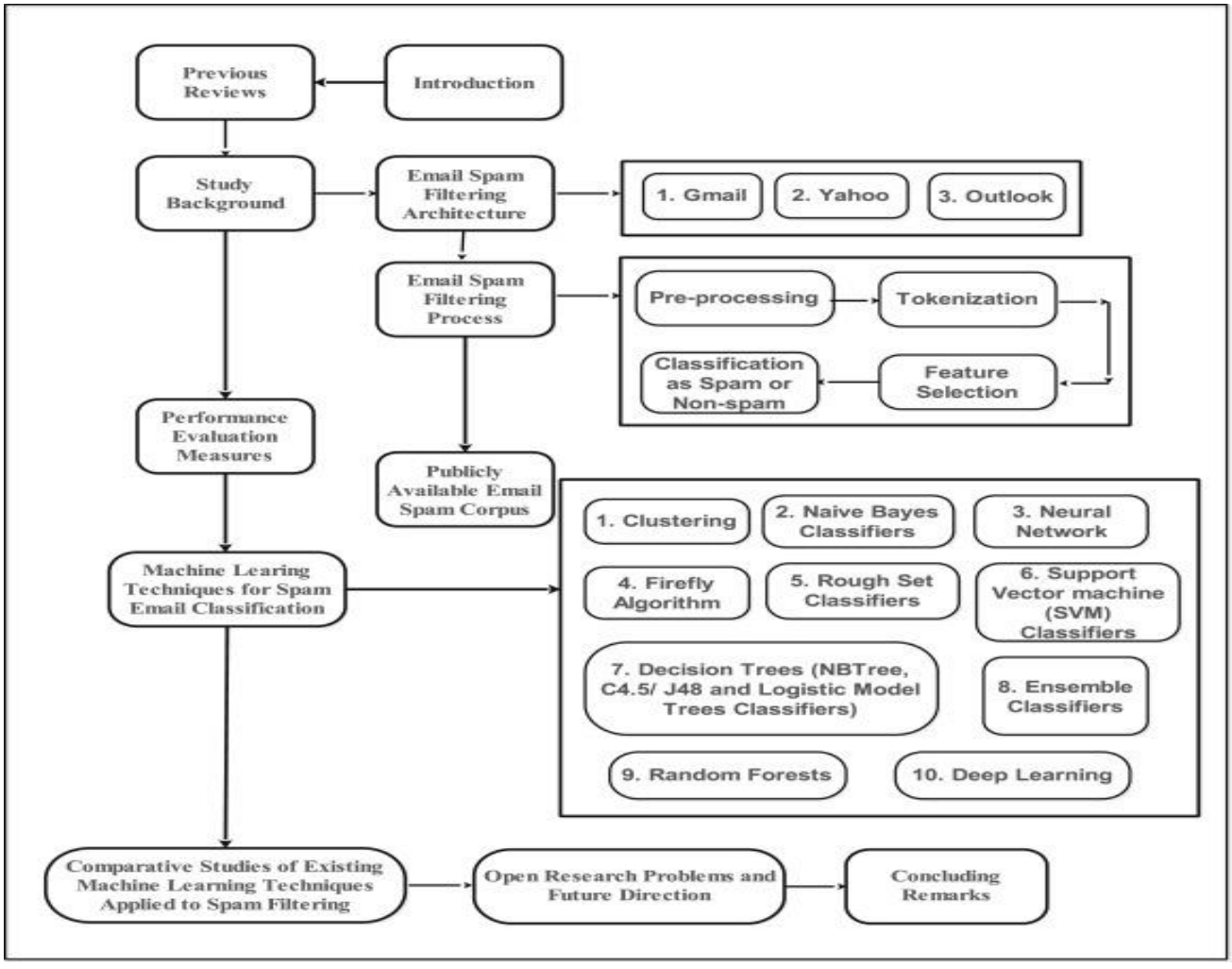
PROJECT TITLE: BUILDING A SMARTER AI-POWERED SPAM CLASSIFIER

Selecting a machine learning algorithm:

Machine learning approach have proved to be more efficient than knowledge engineering approach. No rule is required to be specified, rather a set of training samples which are pre-classified email messages are provided. A particular machine learning algorithm is then used to learn the classification rules from these email messages . Several studies have been carried out on machine learning techniques and many of these algorithms are being applied in the field of email spam filtering. Examples of such algorithms include Deep Learning, Naïve Bayes, Support Vector Machines, Neural Networks, K-Nearest Neighbour, Rough sets, and Random Forests.



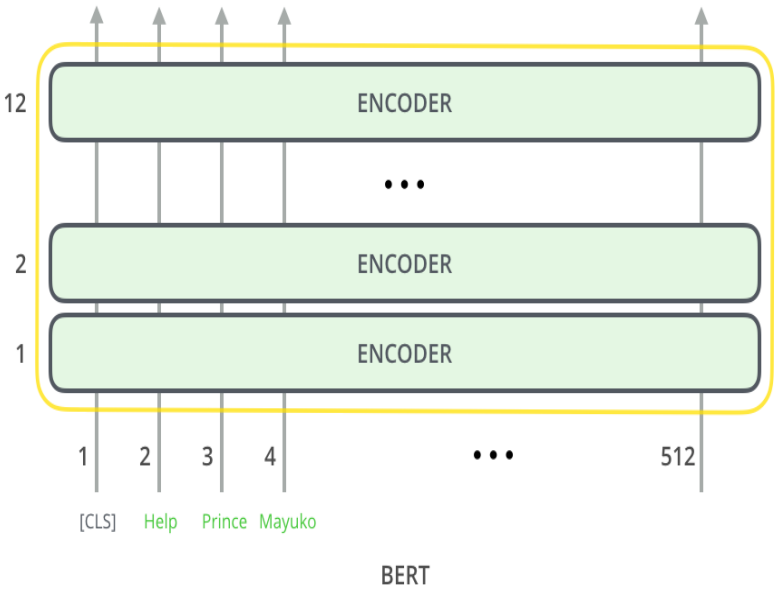
Training the model

To classify spam messages, the BERT model is best for this task as it contains many versions.

In this work, the researchers used the second approach where the finetuning process is made by using linear layers, dropout layers , batch normalization layers , Rectified Linear Unit (ReLU) activations, and log softmax activation with Xavier initialization of weights added at the end in the classifier part of the pre-trained model. The main reason for adding the dropout layer is to avoid overfitting, batch normalization is used to reduce internal covariate shift, and Xavier initialization will help converge the designed model faster

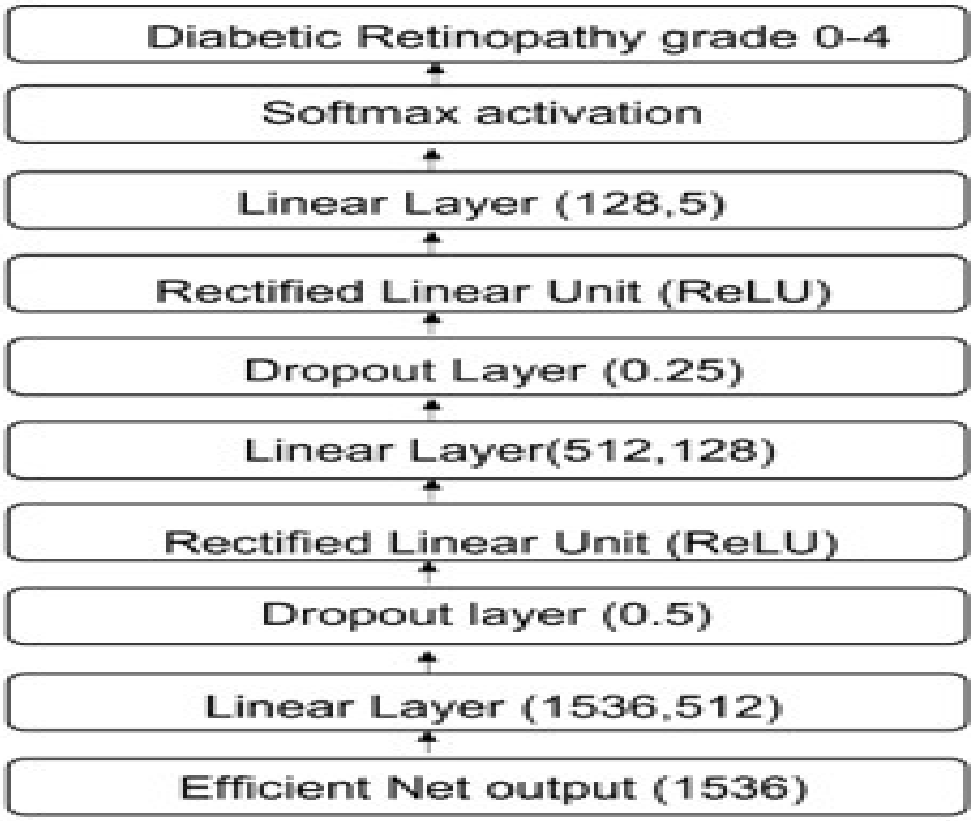
There are two ways to use BERT in our applications. The first approach is to train the model from scratch using the pretrained weights as initial weights, which requires massive data samples and more computational resources as all the weights are updated after each step. In the second approach, all the pretrained weights are not updated and require fewer data samples and fewer computational resources. This approach was used.

Architecture of BERT



.

Final model architecture for classifier



The combined model performed the best with precision and recall values close to maximum values. The combination minimized the false positives and true negatives. Falsepositive is considered ham message classified as spam message whereas true negative is considered spam message classified as ham message. Adding the dropout layer before and after the batch normalization layer could avoid the differences of false-positive and true-negative in the trained model on the combined dataset. The dropout layers produced higher precision and recall values with better accuracy and F1- score. Accuracy and F1-score are the metrics for evaluating the model performance

Spam filters are usually evaluated on large databases containing ham and spam messages that are publicly available to users. An example of the performance measures that are used is classification accuracy (Acc). It is the comparative number of messages rightly classified, the percentage of messages rightly classified is used as an added measure for evaluating performance of the filter. It has however been highlighted that using Accuracy as the only performance indices is not sufficient.

Other performance metrics such as recall, precision and derived measures used in the field of information retrieval must be considered, so also is false positives and false negatives used in decision theory. This is very important because of the costs attached to misclassification. When a spam message is wrongly classified as ham, it gives rise to a somewhat insignificant problem, because the only thing the user need to do is to delete such message. In contrast, when a non-spam message is wrongly labeled as Spam, this is obnoxious, because it indicates the possibility of losing valuable information as a result of the filter's classification error.

This is very imperative especially in settings where Spam messages are automatically deleted. Therefore, it is inadequate to evaluate the performance of any Machine Learning algorithm used in spam filter using classification accuracy exclusively. Furthermore, in a setting that is lopsided or biased where the number of spam messages utilized for testing the performance of the filter is very much higher than that of ham messages, the classifier can record a very high accuracy by concentrating on the detection of spam messages solely. In a real world environment where there is nothing like zero probability of wrongly categorizing a ham message, it is required that a compromise be reached between the two kinds of errors, depending on the predisposition of user and the performance indicators used.

Evaluating its performance

Machine learning algorithms have been extensively applied in the field of spam filtering. Substantial work have been done to improve the effectiveness of spam filters for classifying emails as either ham (valid messages) or spam (unwanted messages) by means of ML classifiers. They have the ability to recognise distinctive characteristics of the contents of emails. Many significant work have been done in the field of spam filtering using techniques that does not possess the ability to adapt to different conditions; and on problems that are exclusive to some fields e.g. identifying messages that are hidden inside a stego image. Most of the machine learning algorithms used for classification of tasks were designed to learn about inactive objective groups. The authors in posited that when these algorithms are trained on data that has some data that have been poisoned by an enemy, it makes the algorithms susceptible to a number of different attacks on the reliability and accessibility of the data. As a matter of fact, manipulating as minute as 1% of the training data is enough in certain instances . Though it might be strange to hear that the data supplied by an enemy is used to train a system, it does happen in some real world systems.

Examples include spam detection systems, spam connection, financial fraud, credit card fraud, and other unwelcome deeds where the earlier deeds of the enemy are a major origin of training data. The unfortunate thing is that a good number systems are

re-trained regularly using the new instances of undesirable activities. This serves as a launching pad for attacker to launch more attacks on such system.

One of the open problem that needs to be addressed is handling of threat to the security of the spam filters. Though some attempt have been made to address this problem. For example, the threat model for adaptive spam filters proposed by categorises attacks based to whether they are causative or exploratory, targeted or indiscriminate, and if they are meant to interrupt reliability or accessibility. The purpose of causative attack is to trigger error in categorisation of messages, whereas an exploratory attack aims to determine the classification of a message or set of messages. An attacks on integrity is meant to have a negative influence on the classification of spam, on the other hand, attacks on accessibility is meant to have a negative influence on the c classification of ham. The fundamental purpose of a spammer is to send spam which cannot be seized by the filter (or user) and labeled as spam. There are other potential capabilities of attack which all depend entirely on the ability to send random messages grouped as spam. A larger percentage of spam filters are nevertheless susceptible to different kinds of attack. For example, Bayes filter is susceptible to mimicry attack . Naïve Bayes and AdaBoost also demonstrated endless deterioration to adversary control attack.

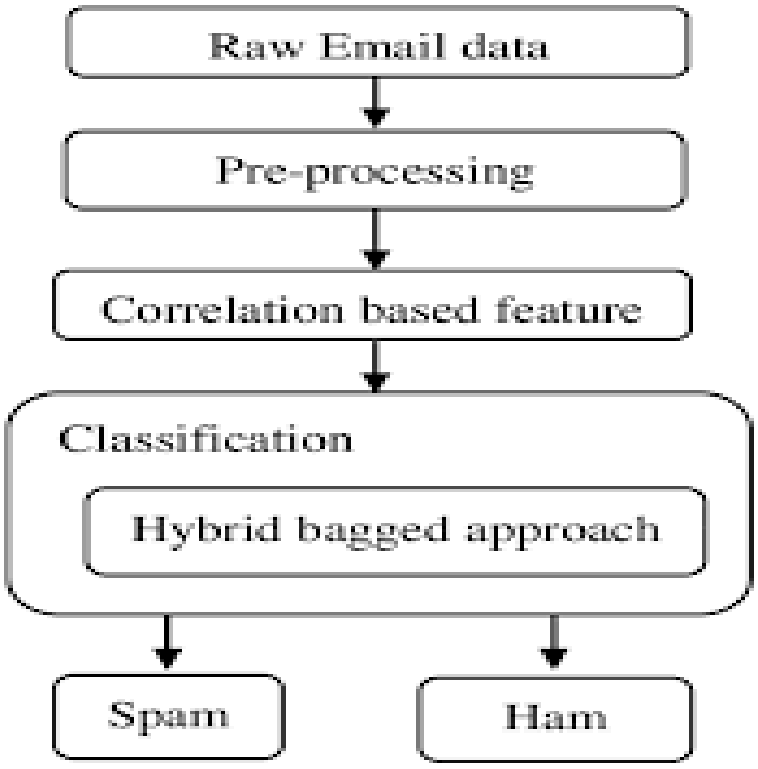


Figure. 1 Basic process for email filtering

We hereby suggest that the future of email spam filters lies in deep learning for content-based classification and deep adversarial learning techniques. Deep learning is a kind of machine learning technique that allows computers to learn from experience and knowledge devoid of explicit programming and mine valuable patterns from primitive data . The traditional machine learning algorithms finds it very hard to mine adequately-represented features because to the limitations that characterised such algorithms. The shortcomings of the usual machine learning algorithms include: need for knowledge from expert in a particular field, curse of dimensionality, and high computational cost. Deep learning have been applied to solve representation problem by creating several naive features to represent a complicated concept. Deep learning will be far more effective in solving the problem of spam email because as number of available training data is increasing, the effectiveness and efficiency of deep learning becomes more pronounced.

Deep learning models have the capacity to solve sophisticated problems by using intricate and huge models. Thus, they exploit the computational power of modern CPUs and GPUs. Deep learning is generally considered to be a black box since we have imperfect knowledge of the explanations behind its high performance. Despite the huge success of deep learning in solving many problems, it has been discovered lately that deep neural networks are susceptible to adversarial examples. Adversarial examples are unnoticeable to human but can effortlessly fool deep neural networks during the testing/deploying phase. The helplessness to adversarial examples becomes one of the foremost dangers for using deep neural networks in situations where safety is very crucial. Therefore, the adversarial deep learning technique is a great method that is yet to be exploited in email spam filtering.
Summarily, the open research problems in email spam filtering are itemized below:

- Lack of effective strategy to handle the threats to the security of the spam filters. Such an attack can be causative or exploratory, targeted or indiscriminate attack.

- The inability of the current spam filtering techniques to effectively deal with the concept drift phenomenon.

- Failure of many spam filters to reduce their false positive rate.

- Development of more efficient image spam filters. Most spam filters can only classify spam messages that are text. However, many savvy spammers send spam email as text embedded in an image (stego image) thereby making the spam email to evade detection from filters.

- The need to develop adapted, scalable, and integrated filters by applying ontology and semantic web to spam email filtering.

- Lack of filters that have the capacity to dynamically update the feature space. Majority of the existing spam filters are unable to incrementally add or delete features without re-creating the model totally to keep abreast of current trends in email spam filtering.