Chinese Spam Classification based on Deep Learning

Zhengyang Chen *

Department of computer science and technology, Beijing University of Chemical Technology, Beijing, 100000, China

* Corresponding Author Email: 2020010082@stu.cdut.edu.cn

Abstract. In the 21st century, it is an information age where individuals and organizations use computers to work and they commonly use email to chat with each other, share knowledge or their wishes or discuss business. However, in this context, spam is harassing our lives. Spam is a type of email that is delivered to lots of numbers of recipients without asking them for their wishes. Spam wastes our time reading useless messages and takes up our email addresses and network traffic. The problem of spam has become increasingly important and is growing over the years. The emergence of machine learning technology can simulate human intelligence. Use machine learning to identify the construct of pictures and words. Through this technology, this paper can classify the kind of different messages. Spam is very common nowadays; it can easily send lots of message to occupying a large amount of network bandwidth. Access such kind of messages, this article uses four models: *Decision Tree, Support Vector Machine (SVM)*, Naive Bayes model, and Logistic Regression to identify spam to prevent it into our electronic mail.

Keywords: Spam classification, machine learning, SVM.

1. Introduction

In this 21st, it is the time of information era, individuals and organizations use the computer to work and they are common to use emails to chat with each other's and share knowledge or their wish or discuss business [1]. However, Spam is harassing our lives with this situation. Spam is a kind of email that be delivered to lots of recipients without the receiver's admission [2]. Spam is a waste of our time reading useless information and occupation our email address and network flow. The problem of spam is becoming more and more important which has been increasing over years. In recent research, 40% of emails are spam.15.4 million spam are sent every day which cost users about 355 million dollars per year [3]. In order to get rid of spam, using machine learning method use some models so the computer can distinguish spam automatically to prevent spam in the email. Machine learning is a way of computational algorithms that are focused to create to let machines learning from the nearby environment in order to emulate the human's mind and predict the results [4]. In general, Machine learning is selecting a model and training from the data that has been given and predicting future data from the data that has been known and models. Machine learning is also the core of Artificial Intelligence (AI). Nowadays, lots of people are searching on this subject and successful design some effective models. However, Chinese and English have lots of differences in orthography, syntax, semantics, and phonetics are different, so the method of English language spam is hardly achieved in the Chinese language [5]. So, it is important to design a model to identify the Chinese language model. There are many methods have been proposed to handle spam by using machine learning. The methods of identifying spam can be distinguished into two groups. One is called static methods and the other is called as dynamic methods. Static methods record the email address to identify spam and dynamic methods are considered the content of the email to identify normal email or spam [1]. In this essay, dynamic methods are used to identify spam.

The Text Retrieval Conference (TREC) is a meeting that host every year which is looking forward to providing the infrastructure necessary to help and support people to do research on the information retrieval community for the amount of email and text retrieval methods [6]. Trec06c is one of the public databases which is provided by this organization and from reality email in 2006. In this paper, SVM, Naive Bayes model, Logistic Regression and Decision Tree will be used to predict Chinese

email and differentiate it as normal email and spam and compare with other models to select the best model to identify spam. The database of those model is trec06c.

Scikit-learn is a machine learning tool based on Python. Scikit-learn library is a kind of Python module that has a lot of the most advanced machine learning algorithms so that medium and small scale supervised, and unsupervised problems can be solved by Scikit-learn library [7]. This library encourages on leading machine learning to become a package so people can call models in a function without having to design it [8]. Supervised learning means the models have some data and features to let models select and predict the results [9]. Unsupervised learning means there is no data to train so select models directly [10]. Using this package can help people more easily to learn machine learning and use machine learning models to achieve the target. Tensorflow is an interface that can help people achieve machine learning models to achieve in Linux systems [11].

To investigate and compare which model can help people prevent spam in their email without detecting normal email.

In this paper, SVM, Decision tree mode, Naive Bayes model, and Logistic Regression will be used to predict Chinese email and differentiate it as normal email and spam and compare it with other models to select the best model to identify spam. The database of those models is trec06c.This kind of model can be used in Python by the Scikit-learn package.

The element of this article will be constituent of four parts: 1. Overview of the state of research about spam identification and problems in the field of high-dimensional data processing. 2. Analyze and introduce different models (algorithms, current state of research, methods). 3. introduce the process of success models. 4. Conclusion.

2. Models

In this charter, SVM, Naive Bayes model, *Decision Tree*, and Logistic Regression will be introduced in this essay and discuss their state and model processing.

2.1. SVM

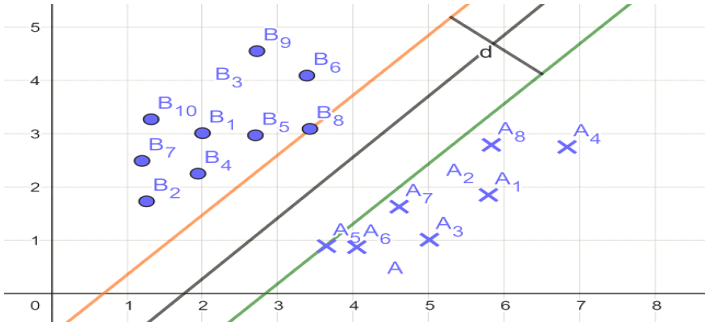

Figure 1. The map of SVM

Figure 1 shows the idea of linear Support Vector Machines in the picture. In conclusion, linear Support Vector Machines use linear to distinguish items into different classes.

Support Vector Machines, known as SVM was first support in 1963 by Vladimir N. Vapnik and Alexander Y. Lerner. SVM had soon become a very important part of statistical learning theory after it was proposed and in 1991, nonlinear Support Vector Machines were first founded by Bernhard E. Boser and Isabelle M. Guyon and Vapnik.

When solving the classification problem, input data are given as X = {x1, x2, ..., xn}. Learning objectives are given as Y = {y1, y2...yn}. All of the input data contains many features. All of these feature constitutes the feature space and learning objectives are classified variables representing negative class and positive class. If the feature space has a decision boundary that can divide the object as a negative class and positive class, the decision boundary is calculated as $w^T X + b = 0$, and point to plane distance is calculated as $y_i(w^T X_i + b) \geq 1$, w and b are the vectors, and the intercept is respectively.

2.2. Naive Bayes model

The Naive Bayes model is based on the Naive Bayesian algorithm. The Naive Bayesian algorithm is dependent on the principle of the Bayes principle. Bayes principle was proposed by Thomas Bayes in the 18 century. The Naive Bayesian algorithm uses probability statistics knowledge to classify the database. The principle of the Bayesian method is the combination of prior probability and posterior probability. Prior probability means the possibility of people guessing from experience. Posterior probability means from lots of test. The possibility comes to a stable number. Naive Bayes makes these two combines so that avoids the subjective bias of using only prior probability. Also, it can avoid using sample information alone which will cause the overfitting phenomenon. When the dataset is large, the Bayesian classification algorithm can get high accuracy, and the algorithm is relatively simple. Nowadays, the Naive Bayes model has become the most popular classification algorithm in the world.

In the dataset D= {d1, d2, ..., dn}. The feature attribute set of the corresponding sample data is X= {x1, x2...xn}. Class variables Y = {y1, y2, ..., yn}, x1, x2...x3 and random and independent. Based on naive Bayes, posterior probability can be calculated from prior probability: $P(Y \mid \mathrm{X}) = \frac{P(\mathrm{Y})\mathrm{P}(\mathrm{X} \mid \mathrm{Y})}{\mathrm{P}(\mathrm{X})}$. Using category Y, the formula can be further expanded as: $P(X \mid Y = \mathrm{y}) = \prod_{i=1}^{d} P(x_i \mid Y = y)$. By these two above formulas, the result can be calculated as: $P(y_i \mid x_1, x_2, ..., x_d) = \frac{P(y_i)\prod_{j=1}^{d} P(x_i \mid y_i)}{\prod_{j=1}^{d} P(x_j)}$ which is the posterior probability. Naive Bayes does not learn the joint probability distribution of input or output directly, instead, it finishes work by learning the prior probability and conditional probability of the class.

2.3. Decision tree

The Decision Tree is a kind of Decision analysis method which was depended on the occurrence probability of different situations, Decision tree is a graphical method of intuitive use of probability analysis.

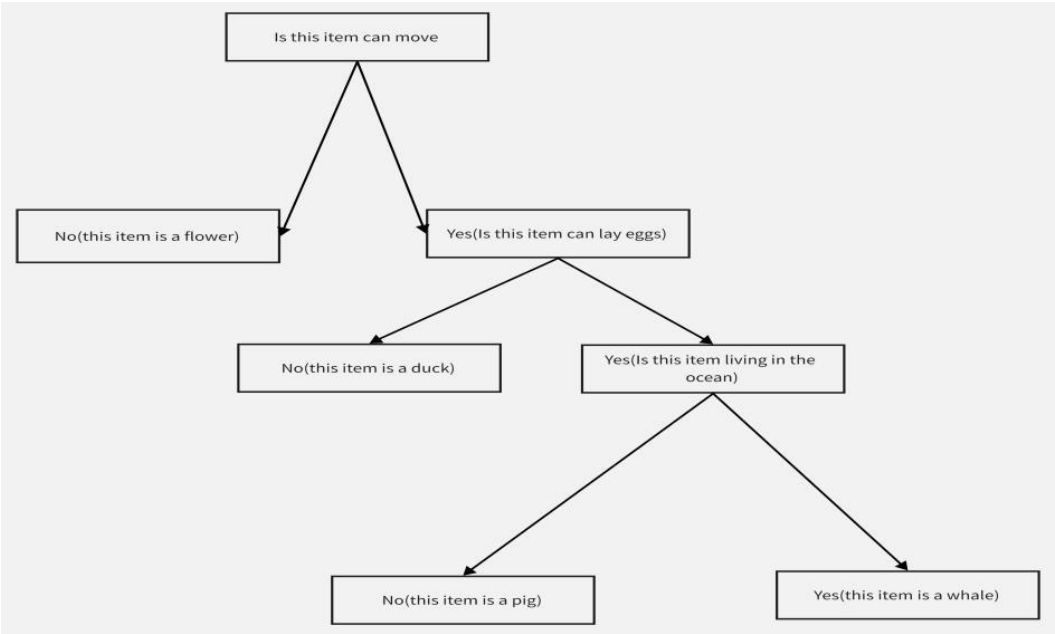

Figure 2. Example of a decision tree. (Photo credit: Original)

As can be seen from Figure 2, the decision tree uses some feature to identify items into several groups. Decision tree is a kind of prediction model. Decision tree shows maps as a dendrogram. In this dendrogram tree between object features and object values. Every node in this tree is a feature. Every path leads to a possible value of the node, and each leaf represents the value of the feature represented by the path from the root node to the leaf node. The decision tree has only one output.

2.4. Logistic Regression

Logistic Regression is a method of Probability that is common in data mining, and economic projection. Logistic Regression is a generalized linear regression analysis model to help identify items. It is used to sort objects into 2 kinds of things.

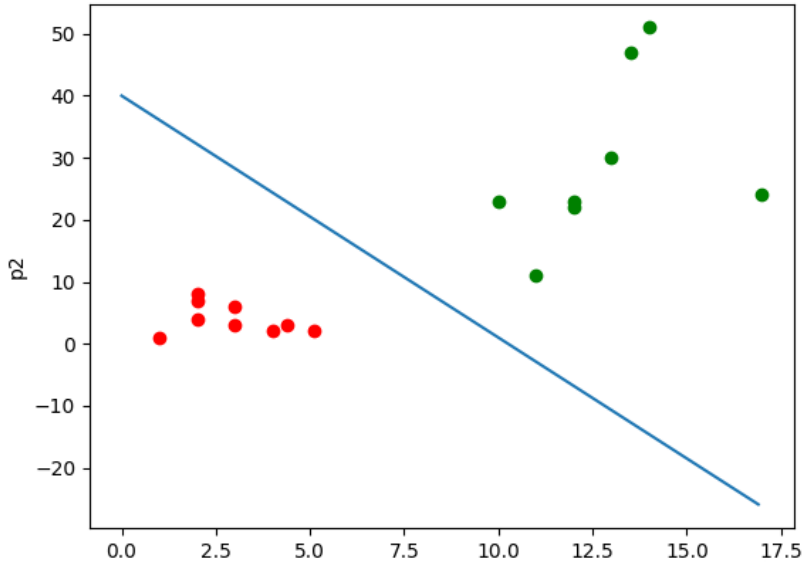

Figure 3. Examples of Logistic Regression. (Reference: https://blog.csdn.net/weixin_60737527/article/details/124141293)

From Figure 3, a linear divide points into two groups. One of above the linear and the other is under the linear. Since there are two classes, one class is set as 0 and the other class is set as 1. Each input data x^{(I)} is mapped to a value between 0 and 1 by a function. And if it's greater than 0.5, it's belonged to 1, otherwise, it belongs to 0. Moreover, the function needs an undetermined parameter. By using samples for training, this parameter can become a stable number and accurately predict the data in the training set. In general, sigmoid is the most common function in Logistic Regression. The formula of the sigmoid is: $f(x)=\frac{1}{1+e^{-x}}$, and cross-entropy loss is used to solve the parameters: $J(w)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}(h(x_i)-y_i)^2$.

3. The Accuracy of Different Kinds of Models

3.1. The achievement of models in python library

The project of model prediction uses the Python to achieve. Scikit-learn is a free Python library. It can achieve lots of machine learning algorithms. NumPy is an open-source scientific computing library in Python. Using these two libraries can easily handle data into training. Flanker is a library that is used to analyze emails, use flanker can help to process the original database.

3.2. The Structure of The Dataset and Result

In trec06c database, its separate email has such structure:

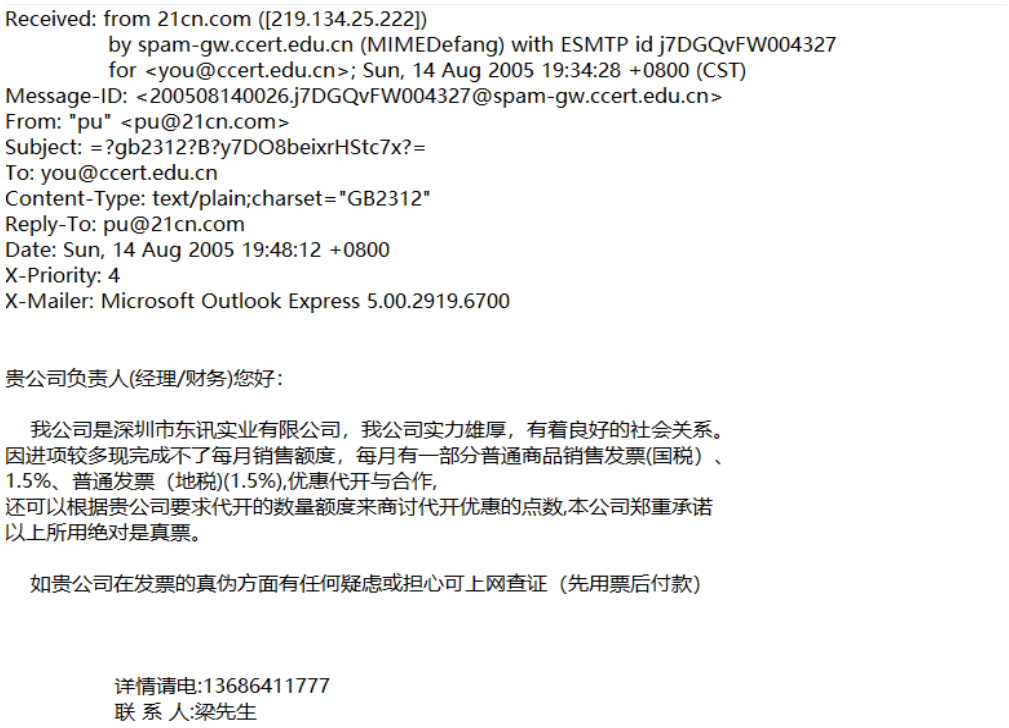

Figure 4. One of the emails in trec06c. (Photo credit: Original)

Figure 4 illustrates the information about the email structure in the trec06c database.

From Figure 4, all of the emails has the same construction. Since the dataset is currently stored as mail and indexed for spam labeling, using EML Analyze in the flanker library to first extract the sender, recipient, subject, time sent, content, and whether or not each message is spam labeled and transform those messages into .CSV file. Using sklearn.model_selection library to separate the dataset into train and test(7:3). Using textPase to get rid of useless words and set data into the train by using sklearn. In this project, accuracy, recall, and F1 Score are recorded to compare with other models. Accuracy means the accuracy of the different model predictions. Recall means the prediction possibility of right in the true right. F1-Score is also called balanced F Score which is the average of accuracy and recall. Using these 3 indices can distinguish the four models' performance on spam identifications.

Table 1. The accuracy and F1 Score of different kinds of models.

Margin	Naive Bayes mode	Support Vector Machine	Logistic Regression	*Decision Tree*
Accuracy	0.93	0.98	0.98	0.94
Recall	0.93	0.98	0.98	0.94
F1 Score	0.92	0.98	0.98	0.94

Table 1 shows the result of these four models regrading their accuracy, recall, and F1 Score.

4. Conclusion

Spam is a waste of our time reading useless messages that take up our email addresses and web traffic. The problem of spam has become increasingly important and has been increasing over the last few years. In recent studies, 40% of emails are spam. 15.4 million spam emails are sent every day, costing users around $355 million a year. Machine Learning is a way to succeed in artificial intelligence by letting machines learn knowledge automatically and it is important to choose models so that machine learning can be more accurate. Machine learning methods are proposed in order to get rid of spam, they are used to prevent spam from entering emails by using models that enable

computers to automatically distinguish spam and normal email. In conclusion, this project compares Naive Bayes mode, Decision Tree, Logistic Regression, and SVM on spam identification by using Python sklearn library to simulate the models and train them. Then use test data to let them predict and compare with the real situations. This process led people to use the best model to help identify spam and normal email in Support Vector Machine, Baive Bayes mode, Decision Tree, and Logistic Regression. In these four models, SVM and Logistic Regression have the highest accuracy, recall, and F1-score in these four models (98%) which means Logistic Regression and Support Vector Machine are very suitable to use in detecting spam attacks because these models can find spam accurately. This result can give us more advanced arithmetic on intelligent transportation applications. Using this kind of detection system can also help intelligent transportation to ignore useless visual information and help AI calculate faster.

References

[1] Awad, W. A., and ELseuofi, S. M 2011 Machine learning methods for spam e-mail classification. *International Journal of Computer Science & Information Technology (IJCSIT)*, **3**, 173-184.

[2] Galligan, L 2001 Possible effects of English-Chinese language differences on the processing of mathematical text: A review. *Mathematics Education Research Journal*, **13**, 112-132.

[3] Yu, B., and Xu, Z. B 2008 A comparative study for content-based dynamic spam classification using four machine learning algorithms. *Knowledge-Based Systems*, **21**, 355-362.

[4] Cormack, G. V 2008 Email spam filtering: A systematic review. *Foundations and Trends® in Information Retrieval*, **1**, 335-455.

[5] El Naqa, I., and Murphy, M. J 2015 What is machine learning?. In machine learning in radiation oncology, 3-11.

[6] Bratko, A., Filipic, B., and Zupan, B. 2006 Towards Practical PPM Spam Filtering: Experiments for the TREC *2006 Spam Track*.

[7] Kramer, Oliver. "Scikit-learn." *Machine learning for evolution strategies*. Springer, Cham, 2016. 45-53.

[8] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... and Duchesnay, E. 2011 Scikit-learn: Machine learning in Python. *The Journal of machine Learning research 12*.

[9] Turc I, Ming-Wei C, Kenton L, and Kristina T 2019 Well-Read Students Learn Better: On the Importance of Pre-Training Compact Models *Preprint* https://doi.org/10.48550/arXiv.1908.08962

[10] Alec R, Karthik N, Tim S, and Ilya S 2020 Improving Language Understanding with Unsupervised Learning OpenAI.

[11] Martín Abadi *et al*. 2015 TensorFlow: Large-scale machine learning on heterogeneous systems *Preprint* https://doi.org/10.5281/zenodo.4724125.