UDC 004.41

# EMAIL FILTERING AND SPAM FIGHTING

**© Fedorova S.A., Agureeva A.V.**

*Samara National Research University, Samara, Russian Federation*

e-mail: feodorowa.sof@yandex.ru

Nowadays, e-mail is not only an effective way to send and receive data. With its help, the user visits various Internet sites. However, sometimes viewing an important email becomes a challenge for many laymen. Among the huge number of unnecessary letters, even an IT authority may get lost.

Furthermore, our email can turn into spam, and we will never know why. The question is, how do spam filters work? Is it possible to adapt modern systems to your needs? These questions formed the basis and the goal of our research – to determine the best algorithm and create our own model based on it.

To achieve this goal, we had to solve the following tasks:

1. To analyse and describe some spam filtering algorithms;
2. To study the methods for the quality of spam filter classification;
3. To develop a mathematical model using machine learning methods for spam filtering;
4. To create and test a software package in Python based on the developed models.

The research consists of several parts: theoretical and practical.

Firstly, the following methods of automatic classification were chosen and analyzed: naive Bayes classifier, logistic regression, and word vectorization. These methods serve as a background for many spam filters. In addition, the authors compared the metrics that would be further used to evaluate the chosen algorithms. The composition of the metrics includes: Accuracy, Precision, Recall, F-measure, Matthews correlation coefficient (MCC). It should be noted, that the algorithms which are widely used in popular companies such as Mail.ru, Gmail.com and Yandex.ru were also analyzed.

Secondly, the authors collected and processed databases of spam messages, which formed the basis of the training sample. Spam databases were searched on the Kaggle.com, archieve.org sites, as well as using a Google search. Three English databases were found with a total capacity of 14 299 letters. This database is universal, as it contains spam on examples of e-mail and SMS messages [1–3].

At the next stage, a comparison was made between the naive Bayes classifier and logistic regression on the obtained sample using a static model such as n-grams(n-grams are a sequence of n consecutive words in a text). The results of the comparison of classifiers are presented in the table.

In the future, to create our own spam filter, we began to use logistic regression, because it copes with the task better than the naive Bayes classifier, and for any set of n-grams.

Next, based on the collected data using vectorization and logistic regression, we created our own spam filter. To do this, we trained the model on the received sample and evaluated the constructed model according to various metrics.

The next step is to evaluate the quality of the developed spam filter. For classification, 25 emails from personal mail were selected, some of which are spam, and some are ordinary letters.

As a result of comparing and evaluating the results, it was found that the spam filter classified 5 regular emails as spam, while ours, on the contrary, classified 4 spam emails as regular.

Comparing the quality scores of the two algorithms, we find that the developed spam filter has a higher accuracy – 0.667, and the Gmail spam filter – 0.444, but at the same time has a lower value of such an indicator as f-measure, for the developed spam filter – 0.444, and for Gmail it is 0.533. Also, during the study, a trend was revealed that the developed algorithm copes better with large messages than Gmail spam filter.

After developing, applying and evaluating spam filtering algorithms, we considered another question regarding the applicability of the constructed model to other samples, for example, to letters in another language.

In order to determine the number of letters in the sample required to create a robust spam filter model, we took the entire mixed sample of 14 299 letters we had and began to train the model sequentially on a part of objects from 10 to 14 299 elements.

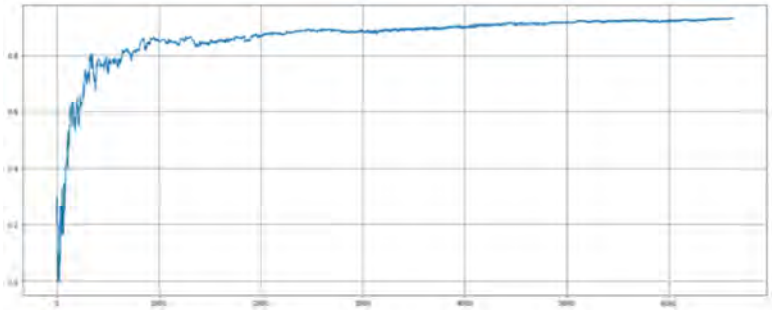

*Figure –- The dependence of the f-measure on the number of letters in the training set*

As you can see from Figure, at the very beginning, the quality is constantly changing, because each new letter greatly affects the quality of the classification. In the future, a logarithmic increase in the quality of the model is visible, and after reaching 2500 letters, the quality of the model grows more smoothly. And we are ready to say that if there are about 2500 letters in the sample, then the algorithm should work stably.

To sum up, a mathematical algorithm based on word vectorization and logistic regression was developed. The study proved that the given spam filter is an effective tool, which can be used in various fields.

## References

1. Рябенко Е., Слесарев А., Кантор В., Соколов Е., Драль Э. Спецкурс «Машинное обучение и анализ данных». Лекция «Прикладные задачи анализа данных» / Д.П. Ветров, Д.А. Кропотов. URL: https://ru.coursera.org/learn/data-analysis-applications.

2. Кемаев Ю.А. Исследование и разработка моделей векторного представления слов. URL: http://seminar.at.ispras.ru/wp-content/uploads/2012/07/pres_diploma.pdf.

3. Ветров Д.П. Спецкурс «Байесовские методы машинного обучения». Лекция 2 «Вероятностная постановка задач классификации и регрессии. Байесовские решающие правила. Обобщенные линейные модели» / Д.П. Ветров, Д.А. Кропотов. URL: http://www.machinelearning.ru/wiki/images/7/78/BayesML-2009-2a.pdf.