Three-Class Classification of Persian Emails by Naïve Bayes Algorithm

NasimVasfi-Sisi, and Mohammad-Reza Feizi-Derakhshi

***Abstract*—** Due to increasing information on the internet, people communicate with each other via e-mail addresses more than before. But sometimes, some people abuse this and send emails with unhelpful content for user that makes the user's time and money wasted. According to this fact that the content of emails mostly are in text, so text classification algorithms have been used for this purpose. In this article, Persian emails are classified into three classes using Naïve Bayesian algorithm and results are evaluated by precision, recall and F-measure. The obtained results show that classifying Persian emails into three classes using Naïve Bayesian algorithm based on precision, recall and F-measure respectively obtains 86.725%, 85.15% and 85.775%.

***Keywords*—** Email classification, Naïve Bayes , IG.

I. Introduction

IN past decade, by rapid development of the internet, e-mails have been became one of the fastest, cost-effective and easiest ways of communication. Nowadays, emails are increasing on internet exponentially, but unfortunately the efficiency and economical nature of emails have been abused. In real world, emails are divided into ham emails and spam emails. Ham emails in most cases have useful and applicative content and are helpful for user, but spam emails in most cases have unhelpful and redundant content[1].

II. Naïve Bayesian Algorithm

The Bayesian theory has been proposed by Tomas Bayes in (1702-1761). The Bayesian theorem is a probabilistic calculation method and an event which would be happened is dependent on an event which was happened before. This theory has self-learning ability in intelligent system which is widely used [2]. Bayesian theory can be used to predict future events according to present events based on statistic and probability theory. Suppose that A is an event in an instance space S, B_1,B_2, …, B_n are mutually incompatible and form a single event and P(A) >0 , P(B_i) > 0 and (i=1,2,3,…,n) and the Bayesian equation is calculated as in (1) [2]:

$$P(B_i \mid A) = \frac{P(B_i)P(A \mid B_i)}{\sum_{j=1}^{n} P(B_j)P(A \mid B_j)} \qquad i=1,2,\ldots,n \qquad (1)$$

Based on (1), Bayesian classification computation is applied in anti-spam field. Email filter is used to classify text with the probability of whether text belongs to ham or spam letters. By Bayesian method we can decide which class the received mail belongs to. Suppose that there are m instance spaces $\{c_1, c_2, ...c_n\}$ email e, has n features $\{t_1, t_2, ...t_n\}$ and is given to $c_k (k = 1,2,..., m)$, the probability that e belongs to c_k would equal to $P(c_k \mid e)$. Based on Bayesian equation, the probability of $P(c_k \mid e)$ is as in (2) [2]:

$$P(C_k \mid e) = \frac{P(C_k)P(e \mid C_k)}{P(e)} \qquad k=1, 2,\ldots ,m \qquad (2)$$

and

$$P(e) = \sum_{j=1}^{m} P(C_j)P(e \mid C_j) \qquad (3)$$

$$P(e \mid C_k) = P(t_1, t_2, ..., t_n \mid C_k) \qquad (4)$$

Where $P(C_k)$ is called the prior probability and $P(C_k \mid e)$ is called the next probability, so (4) has been converted into (5) as follows:

$$P(e \mid C_k)P(t_1, t_2, ..., t_n \mid C_k) = \prod_{i=1}^{n} P(t_i \mid C_k) \qquad (5)$$

Where $P(t_i \mid C_k)$ could be taken from training set, although Naïve Bayesian method is not really accurate, but good results can be achieved [3].

III. Classifying Emails Into Three Classes

Email classification process was always that each email that comes from any user can belong to two classes; spam and ham email. But the performed work is that we consider the third class called User email in addition to both ham and spam email classes and we want to study the precision of classification by classifying emails into three classes. Now, we define each class.

NasimVasfi-Sisi, Young Researchers and Elite Club, Shabestar Branch, Islamic Azad University, Shabestar, Iran, (nasim_vasfi@yahoo.com)

Mohammad-Reza Feizi-Derakhshi , Department of Computer Engineering, University of Tabriz, Tabriz, Iran, (mfeizi@tabrizu.ac.ir)

A. Spam email class definition

Emails which have unhelpful and mostly redundant content are called spam emails.

B. Ham email class definition

Emails which have useful and important information are called ham emails.

C. User email class definition

This class includes those emails that are considered as ham emails from some people's viewpoints and as spam emails from others' viewpoints. For example, some product advertisement emails that initially considered as spam emails by user, but when he/she gets into body and text email and reads it; sees a product in it and wants to buy that product, while at first thought it was a spam email. That's why; we consider the third class as User three classes. So the user sends the emails that come in to his/her inbox to one of these three classes. In summary it can be said that we could not transfer User emails to their related class at first glance and this type of emails require more study and the user transfers the received email to the given class based on his/her own opinion.

IV. Providing A Dataset

According to the performed studies, since there is not any Persian email in three classes'dataset, at first we must collect a Persian email three classes dataset. To collect this data set we have used emails of 4 users and at last 1600 emails have been collected. Users have sent their received emails based on their own opinion into one of three classes of Spam emails, Ham emails and User emails.

Emails consist of different parts such as title, body, and sent date, sent time, sender's and receiver's email addresses and to separate email's different parts we have used labels or HTML tags. We will study only email's body and title parts for three class type of classification.

There are different types of methods for selecting features such as information gain, document frequency, document reverse frequency, Term frequency, correlation coefficient, mutual information, CHI and etc. where in this article the information gain method has been used to select features to detect three classes which would be defined later.

V. Information Gain Method

Information gain is usually used as a good criterion in machine learning. This method evaluates the presence or non-presence of that word in a document according to the number of information observed for each class [4]. This method uses two positive and negative correlation of $P(t,c)$ and $P(\bar{t},c)$. If c and t represent class and the given feature, respectively, the formulation of this method would be as (6):

$$IG(t,c) = P(t,c) * Log \frac{P(t,c)}{P(t) * P(c)} + \bar{P}(t,c) * Log \frac{\bar{P}(t,c)}{\bar{p}(t) * P(C)} \quad (6)$$

Where, $P(t,c)$ is the value of probability where in a document x of training set, the feature t has appeared in class c and $P(\bar{t},c)$ is value of probability where in a document x of training set, feature t has not been appeared in class c and $P(C)$, $P(t)$ and $P(\bar{t})$ represent class c's occurrence probability, feature occurrence probability and not occurrence probability of feature t, respectively.

For each word at each class c, this value is calculated and finally the maximum or mean of those values are considered as that word's information gain and their highest ones are selected [5] , [7]. After selecting the best features, Naïve Bayesian method has been used for classification [6].

VI. Evaluation Criteria

In text classification problems, usually precision, recall and F-measure are used whose equations are as in (7), (8) and (9) [7].

$$\Pr ecision = \frac{TruePositive}{TruePositive + FalsePositive} \quad (7)$$

$$\Re call = \frac{TruePositive}{TruePositive + FalseNegative} \quad (8)$$

$$F_1 = \frac{2 * (precision * \Re call)}{precision + recall} \quad (9)$$

VII. Tests And Results

According to this fact that in emails classification, the number of emails and features are important to diagnosis classes and if the number of features is so high or so low, then the system would not be able to detect the email of each class, so we have performed some tests to determine these two values with evaluation criteria in order to study the results in average for a given number of features and emails for three class case. Based on tables 1, 2 and 3, we study a different number of emails such as 60, 65, 70 ad 75 emails with 80, 90, 100 and 150 number of features with all three criteria.

TABLE I
STUDYING THE CASE OF CLASSIFYING INTO THREE CLASSES OF MEAN EMAILS OF 4 USERS BASED ON PRECISION CRITERION AND NAÏVE BAYESIAN METHOD

150	100	90	80	feature / Email
85.25 %	86.3%	86.55%	83.225 %	60
81.75 %	81.55 %	86.725 %	85.575 %	65
76.9 %	73.85 %	80.625 %	80.875 %	70
75.6 %	74.075 %	74.2%	73.25 %	75

TABLE II
STUDYING THE CASE OF CLASSIFYING INTO THREE CLASSES OF MEAN EMAILS OF 4 USERS BASED ON RECALL CRITERION AND NAÏVE BAYESIAN METHOD

150	100	90	80	features / Email
80.525 %	84.075 %	84.1%	81.125%	60
81.1%	81.15 %	85.15 %	83.775%	65
76.525 %	73.475 %	78.4%	78.425%	70
74.65 %	71.175 %	72.35 %	71.175%	75

TABLE III
STUDYING THE CASE OF CLASSIFYING INTO THREE CLASSES OF MEAN EMAILS OF 4 USERS BASED ON F-MEASURE AND NAÏVE BAYESIAN METHOD

150	100	90	80	features / Email
79.375 %	83.675 %	83.525%	80.6%	60
80.65 %	80.825 %	85.775%	83.5%	65
76.275 %	72.975 %	78.075%	78.1%	70
74.45 %	71%	71.875%	70.77 %	75

Based on tables 1, 2 and 3, we study a different number of emails such as 60, 65, 70 ad 75 emails with 80, 90, 100 and 150 number of features with all three criteria.

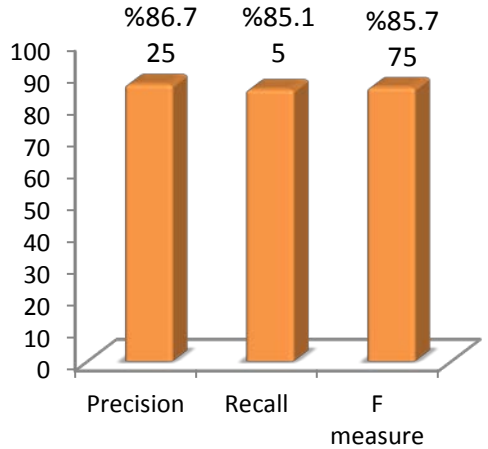

Fig. 1 Represents the diagram of Precision, Recall and F-measure for three class case

As Fig. 1 shows, for three class case emails by using Naïve Bayesian method the values of 86.725%, 85.15%,and 85.775% have been obtained for precision, recall and F criteria, respectively.

VIII. CONCLUSION

Emails are growing in internet exponentially and the usual process of classifying emails is in this way that emails are classified into two classes of spam and ham emails, while in this article we have classified emails into three classes of spam, User and ham. We have studied the three class classification of emails by Naïve Bayesian algorithm and with three precision, recall and F measure and the obtained results of tests are 86.725%, 85.15%, and 85.775% based on precision, recall and finally F measure, respectively.

REFERENCES

[1] L. Diao, C. Yang, "Training Anti-Spam Models with Smaller Training Set Via Svm Way". In International Conference on Electronics and Information Engineering (ICEIE).2, IEEE. 2010, pp. 101-105.
[2] T. Oda, "A Spam-Detecting Artificial Immune System". Master Thesis. Carleton University Ottawa, Canada. January, 2005.
[3] W. Jiansheng, and Z. Xingwen, "Improvement of Chinese Spam filtering method based on Bayesian Classification". In 2nd International Conference on Future Computer and Communication, 1. 2010.
[4] X. Yan, G. J. "A Study on Mutual Information-based Feature Selection for Text Categorization". Journal of Computational Information Systems , 2007, 1007-1012.
[5] Y. Yang, and J. Pedersen, "A comparative study on feature selection in text categorization", 1997, pp. 412-420.
[6] T. Mitchel, "Machine Learning" ,(second ed.), McGraw-Hill Science/Engineering/Math, 1997.
[7] S. Jalili and M. Bitarafan, "Increase the efficiency of text categorization based on the improved feature selection method" , University College of Engineering,4(3),2006 , pp. 313-328.(In Persian).