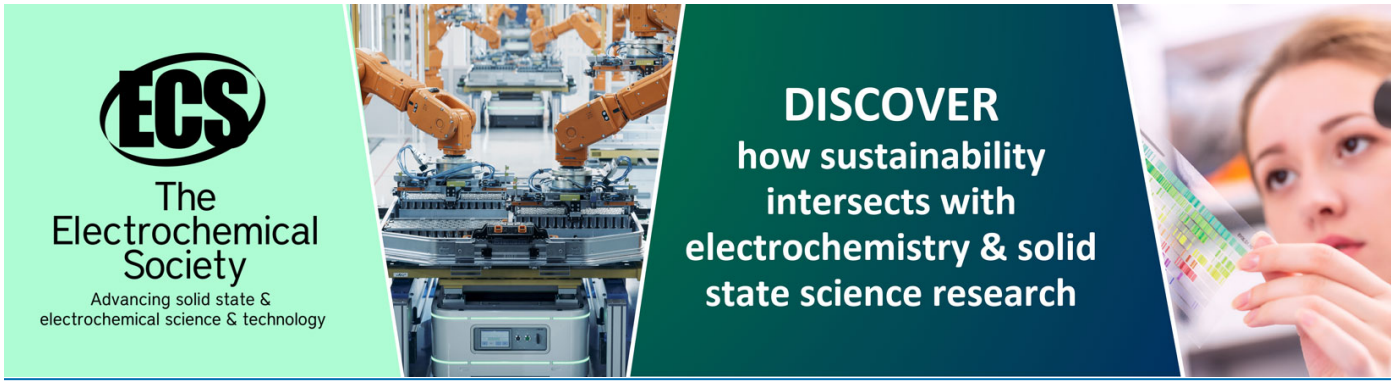
Spam comments prediction using stacking with ensemble learning



Spam comments prediction using stacking with ensemble learning

Arif Mehmood[1], Byung-Won On[2], Ingyu Lee[3], Imran Ashraf[1] ,Gyu Sang Choi[1]

[1]Department of Information and Communication Engineering, Yeungnam University, Gyeongbuk, Republic of Korea
[2]Department of Software Convergence Engineering, Kunsan National University, Republic of Korea
[3]Sorrel College of Business, Troy University, Troy, AL 36082, USA

Corresponding authors:castchoi@ynu.ac.kr[1] , on.byung.won@gmail.com[2]

Abstract. Illusive comments of product or services are misleading for people in decision making. The current methodologies to predict deceptive comments are concerned for feature designing with single training model. Indigenous features have ability to show some linguistic phenomena but are hard to reveal the latent semantic meaning of the comments. We propose a prediction model on general features of documents using stacking with ensemble learning. Term Frequency/Inverse Document Frequency (TF/IDF) features are inputs to stacking of Random Forest and Gradient Boosted Trees and the outputs of the base learners are encapsulated with decision tree to make final training of the model. The results exhibits that our approach gives the accuracy of 92.19% which outperform the state-of-the-art method.

1. Introduction

The Spam in current era of digital communication is a trick for diverting the traffic.The continuous growth of users on social networkslike Facebook, Twitter, YouTube, etc., is also a reason of massive information spread. This information usually spread through reviews and comments that have opened new avenues for spammers.The famous video social websites like YouTube, Daily motion and Vimeo are also infested with spam that normally embroils comments and links to some salacious or dating site or some irrelevant videos. These comments are usually generated automatically through bots software.Predicting spam information is a challenge for the researcher in the field of natural language processing [1].Typically, it is initiated for getting the people's attention through deceptive comments and reviews [2].

In this study, we address above mentioned challenges and propose a novel spam comments prediction model based on a stacking with ensemble learning on text features of TF/IDF [3]. In comparison to other methods which are single learning classifiers our proposed model is based on combining the classifier functionality to achieve the maximum accuracy in prediction. We apply stacking that is a mechanism in which the output of the classifiers at level one will be used as training data for another classifier to approximate the same target function. At level one, we apply Random Forest (RF) [4] and Gradient Boosting Tree (GBT) [5] and for level two we use simple Decision Tree (DT).

The proposed model has been evaluated on data sets of UCI Machine Learning Repository [6] about the YouTube Spam Collection. Results of experiments show that our proposed model significantly outperforms the state-of-the-art methods.
The rest of the paper is organized in the following gmanner. Section 2 presents related work. Section 3 gives details of our proposed stacking model for spam comment prediction. Section 4reports experimental results and comparison with other models and section 5 concludes this work. In the end future works in this domain is highlighted.

2. Related work

Spam data prediction is one of the key security challenges in the field of social networks as well as cyber world in general [7] [8].Usually for detection and filtrations of unwanted information blacklists are used e.g. Twitter uses the blacklists filtering module BotMakerin anti-spam system [9].Social media spam is a type of spam information which spreads rumors on the social media [10].

Freedom of free comment in social media is largely misused by spammers especially in YouTube Sureka [11] proposes a data mining based methodology to predict spam comments in YouTube forums. Radulescu, Dinsoreanu, and Potolea [12] proposes a methodology based on various features such as discontinuous text, inadequate and vulgar content and unrelated and out-of-context comment. They mainly focus on the features extraction but our approach is using simple text features TF/IDF and most importantly the combining of the ensemble learning. Alberto, Lochter, and Almeida [13] introduces an online tool called TubeSpam for automatically detecting spam comments posted in the comments sections of YouTube videos. The authors also evaluates different state-of-the-art classification techniques and conclude that decision trees, logistic regression, Bernoulli naïve Bayes, random forests, and support vector machines are statistically equivalent. In our proposed methodology, we uses different learning technique that not been used previously.

3. Proposed methodology

Our methodology is based on multi learner models and usually it is called stacking. The selections of models at the stage of base learners and at accumulate stage is novel to the best of our knowledge. The flow diagram of proposed methodology is shown in Fig. 1.The proposed methodology has been tested on YouTube spam comments data sets at UCI repository. Data set includes 1956 spam and ham documents that are further divided in the ratio of 70% for training and 30% for test purpose. For training we have total of 1370 documents that include 666 ham and 301 spam. Similarly, division for test we have 586 that include 285 ham and 301 spam documents.

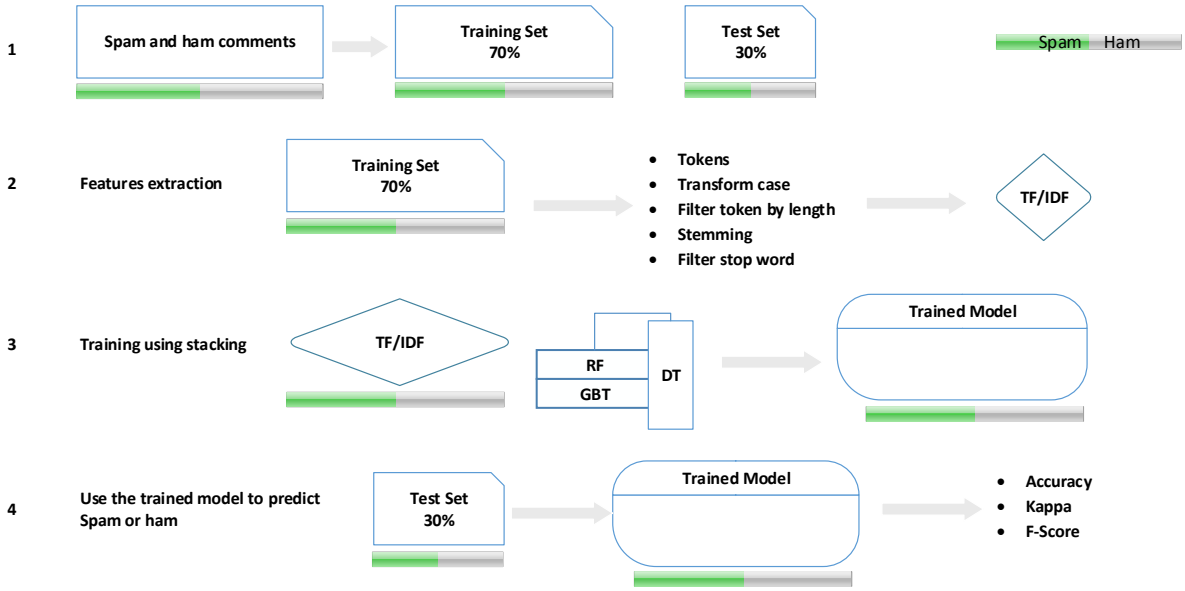


Figure 1. Proposed methodology

We use RF and GBT at base learners. Both models are ensemble learning models and arelater combined with the DT model. We extract only TF/IDF text features as described in (1) for training purpose. The raw spam text has been filtered through different process like tokenization, transform case, stemming, filter token by length and applying the stop words. Porter stemming algorithm[14] is used for stemming and filter token with a length of minimum of 04 characters and maximum of 25 characters.

$$W_{i,j} = tf_{i,j} \times log(\frac{N}{df_i}) \quad (1)$$

Where term frequency i.e., $tf_{i,j}$ is the number of occurrences of i in j and df_i is the number of documents containing i while N is total number of documents.

4. Results and Discussion

We choose the best of ensemble learning models at stage one and they cover the deficiency of each other [15].We keep the TF/IDF features for training and testing the proposed methodology. The same features also tested on various single state-of-the-art learning model and same combination as our proposed approach but Naïve Base (NB) at stage2 in stacking. However, we find that the proposed methodology outperforms other tested models.

The results are shown in Table 1. GBT also shows good results working alone and performs well with RF. However, at stage 2 it does not show good results when working with NB and its performance is not good enough to meet our scheme. The other models like RF and Support Vector Machine (SVM) do not perform well when using the TF/IDF features. The reason our strategy outperforms is that both RF and GBT are ensemble learner based on DT. In addition, we also use DT at stage 2 in stacking which improves the results.

Table 1. Different models accuracy comparison

Methodology	Accuracy
GBT	90.66%
RF	83.02%
SVM	81.49%
(GBT+RF)-NB	91.00%
Proposed	92.19%

The percentage of accuracy, kappa value and F-scores are used as the performance parameters to evaluate the accuracy of all classifiers. Fig. 2 exhibits that the performance of the proposed methodology is significantly better than the other selected models.

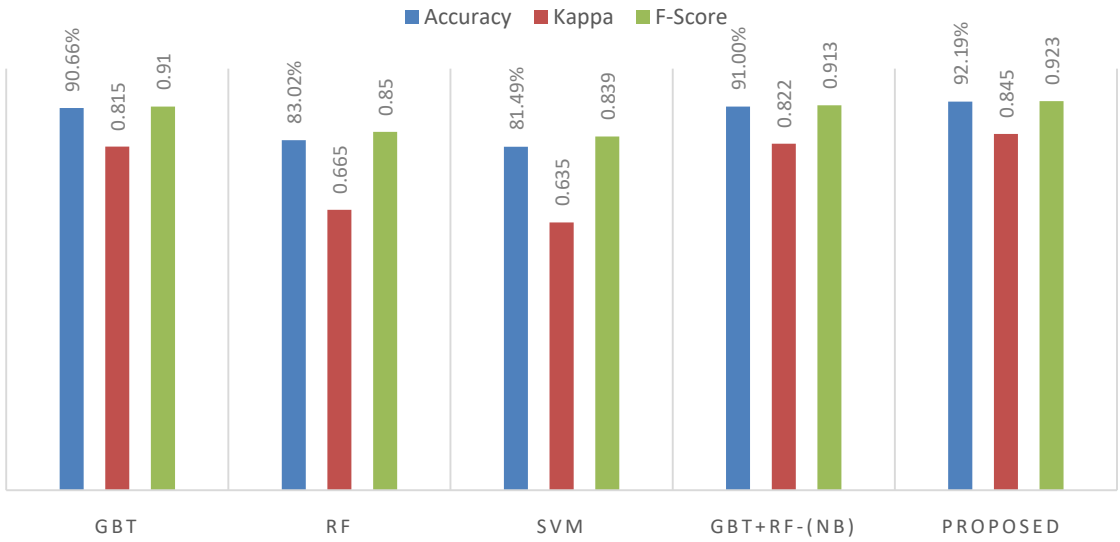


Figure 2. Performance comparison

The percentage of accuracy, kappa value and F-scores are used as the performance parameters to evaluate the accuracy of all classifiers. Fig. 2 exhibits that the performance of the proposed methodology is significantly better than the other selected models. Confusion matrix is shown in Table 2 that exhibits the strong agreement on both ham and spam. The performance of the model in terms of precession and recall are 0.888 and 0.972 respectively.

Table 2. Confusion Matrix of the proposed model

Confusion Matrix	Ham	Spam
Ham	278	35
Spam	8	265

5. Conclusion and future work

This paper proposes a new strategy for spam comments prediction using simple but famous text features of TF/IDF for novel selection of stacking model at base learner and final learning. This

strategy performance gives the accuracy of 92.19% that is significantly better as comparedto single model base or even stacking with NB.
We are planning to test the model on different spam data sets to evaluate the efficiency of the model in the future.

6. References

[1] F. Figueiredo, J.M. Almeida, M.A. Gonalves, F. Benevenuto.Trendlearner: early prediction of popularity trends of user generated content. Inf. Sci. (Ny), 349 (2016), pp. 172–187.
[2] D. Streitfeld.For 2 a Star, an Online Retailer Gets 5 Star Product Reviews26, , New York Times (2012)
[3] G. Salton and C.Buckley, 1988. Term-weighting approaches in automatic text retrieval. Information processing & management, 24(5), pp.513-523.
[4] L. Breiman,Random forests. Mob. Learn. 2001, 45, 5–32.
[5] J.H. Friedman, Greedy function approximation: A gradient boosting machine. Ann. Stat. 2001, 29, 1189–1232.
[6] UCI YouTube Spam Collection Data Set https://archive.ics.uci.edu/ml/datasets/YouTube+Spam+Collection(accessed on 01 July 2017).
[7] F. Norouzi, A. Dehghantanha, B. Eterovic-Soric, K.-K.R. Choo. Investigating social net-working applications on smartphones: detecting Facebook, Twitter,LinkedIn, and Google+ Artifacts ON Android and iOS platforms Aust J Forensic Sci (2015), 10.1080/00450618.2015.1066854
[8] D. Quick, B. Martini, K. Choo Cloud storage forensics Syngress Publishing, Waltham (MA) (2014)
[9] R. Jeyaraman Fighting spam with botmaker Twitter Engineering Blog, August (2014)
[10] F. Wu, J. Shu, Y. Huang, Z. Yuan Co-detecting social spammers and spam messages in microblogging via exploiting social contexts Neurocomputing, 201 (2016), pp. 51–C65
[11] A. Sureka Mining user comment activity for detecting forum spammers in youtube CoRR, abs/1103.5044 (2011)
[12] C. Radulescu, M. Dinsoreanu, R. Potolea Identification of spam comments using natural language processing techniques Intelligent computer communication and processing (iccp), 2014 ieee international conference on, IEEE (2014), pp. 29-35
[13] T.C. Alberto, J.V. Lochter, T.A. Almeida Post or block? Advances in automatically filtering undesired comments Journal of Intelligent & Robotic Systems, 80 (2015), pp. 245–259 https://doi.org/10.1007/s10846-014-0105-y
[14] M. Porter, Porter Stemming Algorithm. Available online: http://tartarus.org/martin/ PorterStemmer/ (accessed on 01 July 2017).
[15] A. Mehmood,B.W.On, I. Lee, and G.S. Choi,., 2017. Prognosis Essay Scoring and Article Relevancy Using Multi-Text Features and Machine Learning. Symmetry, 9(1), p.11.

Acknowledgments

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. NRF-2016R1A2B1014843) for the second author (Byung-Won On), and supported by the Ministry of Trade, Industry & Energy (MOTIE, Korea) under Industrial Technology Innovation Program. No.10063130, by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2016R1A2B4007498), and the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2017-2016-0-00313) supervised by the IITP (Institute for Information & communications Technology Promotion) for the fifth author (Gyu Sang Choi).