Journal of Engineering and Applied Sciences 13 (22): 9638-9642, 2018
ISSN: 1816-949X
© Medwell Journals, 2018

YouTube Spam Comments Detection Using Artificial Neural Network

Thulfiqar Abd, Hussein Altabrawee and Samir Qaisar Ajmi
Faculty of Science, Al Muthanna University, Al Muthanna, Iraq

Abstract: YouTube is considered as one of the most popular video sharing websites that is growing very fast. Because of its popularity, it attracts different types of spammers who publish unwanted spam videos and comments. Spam comment can be defined as the comment that is not relevant to the specific content of a web page. In general, spam comments could be used to publish messages for online marketing, believes of religious, political ideas and links to spam websites that harm the computer of the user. This study presents a YouTube spam comment detection model using a fully connected feed forward neural network. The dataset used to build the model has been obtained from the UCI machine learning repository. A comparison between the ANN Model's results and the results achieved by Alberto is presented and it has been found that the ANN Model is better than most of the models used by Alberto.

Key words: Artificial neural network, spam comment detection, ANN Model's, UCI machine, YouTube spam, popularity

INTRODUCTION

YouTube is a very successful video shearing company. It has more than one billion users and that almost one third of the users of the whole internet. They watch a billion hours of YouTube videos and generate billions of views daily. YouTube created local site versions in more than 88 countries around the world. In addition, YouTube can be viewed using 76 different languages. At 2016, the company paid 2 billion US dollars to the producers who chose to monetize claims, since, 2007. After the lunching of the monetization system, YouTube site was overwhelmed by very low quality content which can be considered as spam videos and spam comments (Alberto *et al*., 2015a, b). Spam comment can be defined as the comment that is not relevant to the specific content of the web page (Alsaleh *et al*., 2015). Comment spams have been used to publish specific unwanted content, declare sales, promote pornographic content, degrade the website reputation, making the website trustworthy by increasing the count of views (Chowdury *et al*., 2013). In order to detect spam comments, several techniques can be used. These techniques have beendivided into two groups; Detection techniques and prevention techniques. Prevention techniques include registering the users of the website, CAPTCHA, limiting the time for commenting, limiting the comments number for a specific post, controlthe comments number for the users or the user's IP address, blocking IP addresses and the prevention of the external website's links (Alsaleh *et al*., 2015). On the other hand, detection techniques areused to classify the comments into spam commentsand hum comments. Machine learning algorithms can be used to build detection models.

In this research, a fully connected feed forward neural network has been used as a classification technique in order to detect the spam comments on YouTube. The dataset used to build the ANN Model has been retrieved from the UCI machine learning repository and it is collected and used by Alberto *et al.* (2015a, b). In addition, the ANN performance measures have been compared with the measures obtained by Alberto *et al.* (2015a, b). They have used many classification techniques that include decision tree, naive bayes, k-nearest neighbors, logistic regression, random forests and support vector machine. The ANN Model achieved a higher performance than most of the techniques used by Chowdury *et al.* (2013).

Literature review: Much research has been done in the area of spam comments detection. Alberto *et al.* (2015) showed that many classification techniques can be used to find the spam comments in YouTube. These techniques include decision trees, logistic regression, random forests, linear and Gaussian support vector machines and Bernoulli Naive Bayes. They used a dataset collected from YouTube which has 1956 real user comments that are related to five most viewed YouTube videos. Alberto *et al.* (2015a, b) labeled each comment as a spam or ham. The text of each comment is used as the input features to the machine learning algorithms by using the

Corresponding Author: Thulfiqar Abd, Faculty of Science, Al Muthanna University, Al Muthanna, Iraq bag-of-words model. Different performance measures have been used to evaluate the classification models such as the accuracy rate, the spam caught rate, the blocked ham rate, the F-measure and the Matthews correlation coefficient. They found that most of the machine learning techniques achieved accuracy rates that were higher than 90% as well as the rates of the blocked ham were lower than 5% (Chowdury *et al*., 2013).

Alsaleh *et al*. (2015) have proposed a spam comment detection system that is based on machine learning techniques. They used a dataset created by Wei (2012) which contains seven thousands posts and 574.054 comments collected from blogs. The comments labeled by the blog users to indicate their informative level. Alsaleh *et al*. (2015) relabeled the comments either with a spam flag or a ham flag. They labeled 385 comments with a spam label and 8.277 with a hum label. They have used the synthetic minority oversampling technique to fix the imbalanced dataset. They used four machine learning techniques which include decision tree, random forest, support vector machine and artificial neural network. Ten folds cross validation method has been used to train and evaluate the classification models. They used twelve features: post-comment similarity, inter-comment similarity, interval between post and comment, number of words in the comment, number of sentences in the comment, comment length, phone information, E-mail information, URL link, black words list, stop words ratio and word duplication ratio. In addition, they used two features selection filters, CfsSubsetEval filter and FilteredSubsetEval filter. The classifiers have been trained using all features or the best features selected by the filters. They showed that the ANN Model and the random forest model have a better performance when compared to the other models.

Radulescu *et al*. (2014) have built a spam comments detection system using machine learning techniques, topic detection and natural language processing techniques. The system is divided into three models, the feature extraction model, the topic extraction model and the post-comment similarity model. The training dataset, used by Mishne *et al*. (2005), contains 1024 comments. On the other hand they used datasets from YouTube and the Daily Telegraph Websites during the system evaluation process. They labeled each data row manually, Decision tree, Naive Bayes and support vector machines were used as machine learning algorithms. Many features related to the comments have been used to differentiate between spam and ham comments. They include: the links count, the whitespaces count, the number of sentences, the punctuation marks count, word duplication, the ratio of the stop words, the number of non ASCII characters, capital letters count and the number of the newlines in the comment. The decision tree model achieved the best results.

Ammari *et al*. (2011) worked on creating user models or profiles that enriched by the characteristics of the users which are obtained from the social websites. As a first step they focused on filtering and identifying the noisy web content such as spam comments that are irrelevant to a specific domain. They proposed semantically enriched machine learning system that can detect and filter the spam comments related to YouTube job interview videos. In their research, two datasets have been used. The first dataset contains a clean and highly relevant, experimentally controlled, YouTube comments. It was used in order to create semantically enriched bag of words that represents the ground truth vocabulary of the job interview activity. The second dataset contains publicly unlabeled comments. They labeled each unlabeled comment in the dataset based on the comment's relevance score which is computed using the bag of words. Then, the labeled dataset was used to train the classification models. The second dataset contains 1,159 YouTube comments that are related to seventeen job interview videos while the first dataset contains 193 clean and highly relevant user-guided comments. Two machine learning techniques have been used to create the classifiers, C4.5 Decision Tree (DT) and Naive Bayes multinomial. They found that the decision tree model, C4.5 is better in classifying the noisy comments. On the other hand, the Naive Bayes classifier had lower false positive rate than the C4.5 classifier.

Song *et al*. (2014) proposed new technique to detect social spam comments. They used word based features, topic based features and user based features. The dataset contains 6.407 videos that have 6,431,471 total comments. There were 481.334 spam comments in the dataset. In their research, support vector machine has been used to train the classification model. They usedprecision, accuracy, recall, F1-measure and receiver operating characteristic curve as an evaluation and performance metrics. Their model achieved 91.17 accuracy, 78.43 F1 score and 87.75 ROC.

Ezpeleta *et al*. (2017) have built a spam comments filtering method. They focused on sentiment analysis features and personality recognition features. They used a dataset contains 6,431.471 comments and the spam comments equals 481,334. In addition, a dataset that includes 1,000 Spam and 3,000 Ham was added. Seven different classifiers have been used to build the model.

They were applied to four datasets (original dataset, polarity dataset, personality dataset and combined dataset). They achieved 82.55% accuracyby applying the model on the combined dataset (Ezpeleta *et al*., 2017).

Mehmood *et al*. (2018) proposed spam comment detection model depend on staking. The features Term Frequency/Inverse Document Frequency (TF/IDF) are used in the model. They used random forest and gradient boosting tree in the first level then they used decision tree classifier at level two. The data set has been obtained from UCI machine learning repository. The model achieved 92.19 accuracy.

Alberto *et al*. (2015a, b) have presented an overall study of machine learning mechanisms to detect unwanted comments automatically in the blogs. They used eleven classification techniques. The data collected from (Mishne *et al*., 2005) the data has been divided into three features groups: the first group has 11,901 attributes obtained from text message, the second group has 2,442 attributes obtained from posting metadata and the third group has 13,776 attributes obtained from text message and posting data.

MATERIALS AND METHODS

The proposed system: The following diagram Fig. 1 shows the main steps and components of the proposed machine learning system.

The first step is collecting the data from the data sources. In our case, the data has been collected from the UCI machine learning repository. The second step is preprocessing the data in order to get a bag of words representation of each comment in the dataset. In the third step, the result of the second step, the training and testing dataset is fed to the machine learning algorithm. The machine learning algorithm builds a model using the training data and tests the model using the test data. Finally, the machine learning algorithm produces a trained model or a trained classifier that can take as an input a new data row and predicts its label.

The system components

Artificial neural networks: An Artificial Neural Network (ANN) represents a machine learning technique that can model a nonlinear and complex relationship between a set of descriptive features and a target variable. Its design inspired by the biological neurons architecture and it has many desirable characteristics such as fault tolerance, speed and scalability with parallel computation. The artificial neural network can be defined as an interconnection of nodes, neurons and it has three main parts: the topology of the network, the node character and the learning rules (Zou *et al*., 2008).

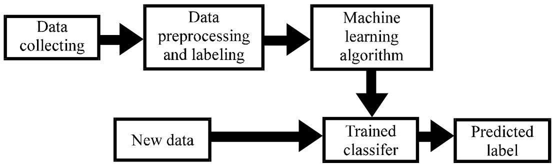

Fig. 1: The main steps and components of the proposed system

The node character specifies the way by which the node processes the signals. The node character includes determining the input of the node, the output of the node and the associated weights and the activation function used by the node. On the other hand, the network topology specifies how the nodes or the neurons are connected. The learning rules specify the way of initializing and adjusting the weights during the learning process. Each ANN neuron or node receives many inputs by weighted connections coming from other nodes (Zou *et al*., 2008).

There are many choices for the activation function used by the nodes such as the sigmoid, the rectifier linear unit and the tanh activation function. The network topology usually consists of an input layer, hidden layers and an output layer. Each layer has many neurons. Furthermore, the network topology can be classified into feedforward network and feedback network. In the feedforward network, the nodes are connected by a one way connection with no loop goes back. While in the feedback network, the output connection of the nodes could be the input connection to previous layer nodes or to the same layer nodes. There are two well-known approaches for learning the network, nearest neighbor methods and error correction methods (Zou *et al*., 2008).

The experiment

Dataset and data sources: The dataset used in this research has been used by Alberto *et al*. (2015). The dataset contains user's comments related to the top five most viewed YouTube videos. The data extracted directly from YouTube at 2015. They labeled each comment manually as a spam or ham. Each row in the dataset contains the video ID, the author of the comment, publication date and the comment text. The dataset is available at UCI machine learning repository under the name "YouTube spam collection data set. The data was divided into five sub datasets: Psy, KatyPerry, LMFAO, Eminem and Shakira.

Data preprocessing and machine learning software: To build the machine learning model, the text comment has been used only as an input to the machine learning algorithm. Each text has been converted to bag of word representation. RapidMiner studio has been used as a machine learning software.

Accuracy and performance measures: In this research, well known accuracy and performance measures have been used such as accuracy rate, Spam Caught rate (SC), Blocked Ham rate (BH), F1 measure and MCC.

Models implementation: In this research, RapidMiner studio has been used as a machine learning software which offers several useful features used to build the ANN Model. Those features include parameter optimization, regularization, data preprocessing, data sampling and measuring the model's performance.

The first step of the implementation was using a grid parameter optimization operator that was responsible of finding the best values of the ANN Model learning rate and the ANN Model L2 regularization. The minimum value of the learning rate was 0 while the max value was 1.3 and the number of the steps was 100 steps on the linear scale. The minimum value of the L2 regularization was 0 while the max value was 1 and the number of the steps was 10 on the linear scale.

Inside the optimization operator, a split validation operator has been used which was responsible of dividing the data into the training and the testing datasets. The split ratio set to 0.7 and the linear sampling has been used to divide the data. The main reason for using the linear sampling is to insure that the earliest comments are used in the training process while the newer comments are used in the testing process.

Inside the split validation operator an ANN operator has been used which was configured to have two hidden layers. Each hidden layer has 50 neurons that use a rectifier activation function. The 25 epochs have been used to train the ANN Model. A performance operator has been used to measure the model's accuracy and the classification performance.

RESULTS AND DISCUSSION

Table 1 shows the results of applying the ANN classification model on each one of the 5 datasets.When comparing these results withthe results obtained by Alberto *et al.* (2015a, b) for the Psy dataset, the ANN Model achieved higher accuracy, F1 measure and MCC; Similar blocked Ham rate as (Alberto *et al.*, 2015a, b) which equals to 0. On the other hand, Alberto's Models achieved slightly higher SC rate than the ANN Model.

Table 1: The ANN Model's results

Dataset name	Spam caught rate (SC) %	Blocked Ham rate (BH) %	F1 measure	MCC	Accuracy (%)
Psy	93.88	00.00	0.9684	0.9439	97.14
Katy perry	97.87	5.17	0.9583	0.9240	96.19
LMFAO	98.21	5.33	0.9565	0.9235	96.18
Eminem	98.41	00.00	0.9920	0.9851	99.25
Shakira	100.00	7.41	0.9661	0.9302	96.40

For the Katy Perry dataset, the ANN Model achieved higher accuracy, SC, F1 measure and MCC than Alberto's Models while it achieved lower BH than Alberto's Models. For the LMFAO dataset, the ANN Model achieved higher SC and MCC than Alberto's Models while it achieved lower ACC, BH and F1 measure than Alberto's Models.

For the Eminem dataset, the ANN Model achieved higher accuracy, F1 measure and MCC; Similar blocked ham rate as (Alberto *et al.*, 2015a, b) which equals to 0. On the other hand, Alberto's Models achieved slightly higher SC rate than the ANN Model. For the Shakira dataset, the ANN Model achieved higher F1 measure; similar SC rate as (Alberto *et al.*, 2015a, b) which equals to 100. On the other hand, Alberto's Models achieved slightly higher ACC, BH and MCC rate than the ANN Model.

To summarize, the ANN Model achieved better accuracy, F1 measure and MCC than Alberto's Models in most of the datasets while it achieved lower or equal BH rate in most datasets.

CONCLUSION

YouTube is considered as one of the most popular video sharing websites that is growing very fast. Because of its popularity, it attracts different types of spammers who publish unwanted spam videos and comments (Chowdury *et al.*, 2013).

In this research, 5 datasets were obtained from UCI machine learning repository that were collected and used by Alberto *et al.* (2015a, b) to build a spam classifier.

The main goal of this research was to detect YouTube spam comments by using ANN Model and compare its results with the results achieved by Alberto's Models. The ANN results have indicated that The ANN Model achieved better accuracy, F1 measure and MCC than Alberto's Models in most of the datasets while it achieved lower or equal BH rate in most datasets.

RECOMMENDATIONS

As future research, more data would be collected as well as more classification methods could be used in order to get better results.

REFERENCES

Alberto, T.C., J.V. Lochter and T.A. Almeida, 2015b. Post or block? Advances in automatically filtering undesired comments. J. Intell. Rob. Syst., 80: 245-259.

Alberto, T.C., J.V. Lochter and T.A. Almeida, 2015a. Tubespam: Comment spam filtering on Youtube. Proceedings of the 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), December 9-11, 2015, IEEE, Sorocaba, Brazil, ISBN: 978-1-5090-0287-0, pp: 138-143.

Alsaleh, M., A. Alarifi, F. Al-Quayed and A. Al-Salman, 2015. Combating comment spam with machine learning approaches. Proceedings of the 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA'15), December 9-11, 2015, IEEE, Miami, Florida, ISBN:978-1-5090-0286-3, pp: 295-300.

Chowdury, R., M.N.M. Adnan, G.A.N. Mahmud and R.M. Rahman, 2013. A data mining based spam detection system for youtube. Proceedings of the 2013 8th International Conference on Digital Information Management (ICDIM), September 10-12, 2013, IEEE, Dhaka, Bangladesh, ISBN: 978-1-4799-0615-4, pp: 373-378.

Ezpeleta, E., I. Garitano, I. Arenaza-Nuno, J.M.G. Hidalgo and U. Zurutuza, 2017. Novel comment spam filtering method on YouTube: Sentiment analysis and personality recognition. Proceedings of the International Conference on Web Engineering, June 5-8, 2017, Springer, Rome, Italy, ISBN:978-3-319-74433-9, pp: 228-240.

Mehmood, A., B.W. On, I. Lee, I. Ashraf and G.S. Choi, 2018. Spam comments prediction using stacking with ensemble learning. J. Phys. Conf. Ser., 933: 012012-012016.

Mishne, G., D. Carmel and R. Lempel, 2005. Blocking blog spam with language model disagreement. Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb'05), May 10, 2005, Chiba University, Chiba, Japan, pp: 1-6.

Radulescu, C., M. Dinsoreanu and R. Potolea, 2014. Identification of spam comments using natural language processing techniques. Proceedings of the 2014 IEEE International Conference on Intelligent Computer Communication and Processing (ICCP'14), September 4-6, 2014, IEEE, Cluj Napoca, Romania, ISBN:978-1-4799-6568-7, pp: 29-35.

Song, L., R.K.K. Lau and C. Yin, 2014. Discriminative topic mining for social spam detection. Proceedings of the 2014 Pacific Asia Conference on Information Systems (PACIS'14), June 24-28, 2014, Association for Information Systems, Atlanta, Georgia, USA., pp: 1-17.

Wei, J., 2012. Blog comments classification using tree structured conditional random fields. MSc Thesis, University of British Columbia, Vancouver, Canada.

Zou, J., Y. Han and S.S. So, 2008. Overview of Artificial Neural Networks. In: Artificial Neural Networks, Livingstone, D.J. (Ed.). Humana Press, New York, USA., ISBN:978-1-58829-718-1, pp: 14-22.