Spam Identification on Facebook, Twitter and Email using Machine Learning

Koushik Chowdhury

***Abstract*—**This decade has witnessed the growth of social media. Social media is consist of user-generated content such as post sharing, update profile, find new people. There are lots of popular social media such as Facebook, Twitter, Instagram, WhatsApp, etc. Among these social media, Facebook and Twitter are the most popular sites. Billions of people use these sites to share their expressions, opinions and thoughts. The user of these sites is witnessing various kind of unwanted social threats in every moment such as insults, hate speeches, malicious links, nudity, bulk messages, etc. These social threats are known as social spam. Not only in social media, but also in the electronic mails, people have received these kinds of social threats. So, it is necessary to identify the spam content. In this paper, I discussed three different datasets that are made of spam texts and non-spam texts. Then, I applied the various machine learning classifier to find out the best classifier. My analysis shows that the support vector machine produces the best results. In the end, I identified when a tweet or a Facebook comment or an email is 'Spam' content or 'Not spam' content. I also employed neural networks like MLP classifier and simple CNN architecture to detect spam contents.

***Keywords*—**Spam, classifier, SVM, MLP, dataset.

I. Introduction

According to data from statista, more than three billion people do social networking via social media [1]. Facebook has more than two billion users and Twitter has around 400-500 million users. Also, Millions of people communicates via electronic mails every day. On this medium, people share their events of public and private life. The main concerns of these media relate to their privacy, security, ethical aspect, abuse information and more. People can unconsciously click on a malicious URL that can interfere with their social life. It causes various problems such as profile hack, information leakage, abusive information, etc. Identifying this type of spam is important to prevent the misuse of information. In this paper, I have tried to detect these types of spams. To do that I have worked with various classification algorithms. Classification algorithms are supervised learning method. It classifies data item into the predefined class label [2]. The classification algorithm builds a model and that model help to predict future data trends. There are several techniques for data classification such as Decision tree, logistic regression, random forest, gradient boosting, k-nearest neighbors, support vector machine, etc. With classification, the generated model will be able to predict a class for given data depending on previously learned information from historical data or recorded data [3]. In this research, I have collected three different datasets of three different mediums. Three mediums are Facebook, Twitter and Email. I have analyzed out datasets with logistic regression, k-nearest neighbor, gradient boosting, random forest and support vector machine. According to accuracy and AUC value, I have found that the support vector machine is the best classifier for all three datasets. I have used MLP classifier as well as convolutional neural network to find out which datasets provide the highest accuracy. The Twitter dataset provides the highest accuracy and Facebook dataset provide the lowest accuracy.

Koushik Chowdhury, Saarland, Germany, Email: s8kochow@stud.uni-saarland.de

II. Background Study

A. Literature Review

Classification is one of the most used technique in statistical learning. Many researches have shown the way and purposes of classification algorithms for different fields in their researches. In this research, I have gone through some of the states of art research papers that recently published. Most of the research paper is about prediction, measuring performance, forecasting or determining the future situation. One of the research paper has tried to detect social spam by using traditional classifier [4]. This paper detects social spam based on Facebook data. Some researchers have applied convolutional neural network to detect the deceptive opinion spam [5]. They also compare their result with regression neural network and general regression network. They have collected their opinion data from the hotel, restaurant and doctor. Another group of researchers from Saarland University worked with 2016 election foreign state-sponsored accounts data on Twitter [6]. They employed the traditional classification algorithms as well as demonstrate a neural network technique. I have gathered information about how to classify data successfully and apply classification algorithm based on some features from these research papers. In this paper, I have collected data set for detecting the spam on Facebook, Twitter and Email based on some features.

B. Data Collection

One of the major parts of this research was data collection. I have collected three different datasets from three different sources. They are...

1. *Twitter Dataset:* The source of this dataset is kaggle. There is an online competition in Kaggle named UtkMl's Twitter Spam Detection Competition, 2019 [7]. This dataset has 8 feature including id, tweet, following, followers, action, is_tweet, location and Type. There are 11968 tweets. *'following'* and *'follower'* indicate how many people follow the tweet and how many follow the account that post the tweet. *'action'* indicate the number of retweets and favorites of the tweet. *'is_retweet'* indicate if a tweet is retweet or not. *'location'* represent the person's location (user profile location). *'Type'* is the class which indicates the tweet is either *'Quality'* (not spam) or *'Spam'*. In the dataset, 49% of tweets are spam tweet. There is no missing value. There are 11787 unique tweets out of 11968 tweets. Most repeated text is *'[HAPPY BIRTHDAY TAEYANG] noriginally posted by'* and the frequency of this text is 10. It belongs to 'Quality' class which is not spam.

2. *Facebook Dataset*: The source of Facebook dataset is the national science foundation OCI-1144061 [8]. There are total 1000 texts. This dataset is so small. The dataset has binary class, positive and negative. Positive indicates the total number of positive texts and negative indicates the total number of negative texts. 64% of texts are positive. There are 979 unique texts out of 1000 texts. Most repeated text is *'I love mine!'*. The frequency of this text is 3.

3. *Email Dataset:* The source of Email dataset is kaggle [9]. This dataset is larger than the Facebook dataset. There are 5728 Emails as well as there is no missing value. There are 5695 unique texts out of 5728 texts. Most repeated text is *'Subject: re : interviews vince , no problem ...'* and the frequency of the text is 2. This dataset has binary class, true and false. 76% of texts are from false class (non spam) and 24% of texts are from true class (spam).

4. *Combined Dataset:* I have combined the 3 datasets into a dataset based on the following features.

   Twitter (Tweet, Type)
   Facebook (Text, Type)
   Email (text, spam)

   I have extracted the 2 important features from each of the datasets and made the combined dataset which also contains features named Input and Result. Input is a combination of tweet, Text (Facebook) and text (Email) data where Result is a combination of Type (Twitter), Type (Facebook) and spam (Email). Result feature has two classes named 'Spam' and 'Not Spam'. The combined dataset consists of 11896 input texts, in which 42% of the texts are based on spam.

III. Approaches

A. Vectorization

The vectorization approach helps to convert all textual data into numbers. We can import *CountVectorizer* from *sklearn.feature_extraction.text.* CountVectorizer makes vocabulary from all unique words by taking all words from each sentence. A simple vectorization example is following.

Intput:
```
text = ['I am student of UdS', 'I am not a
student of UdS.']
vectorizer = CountVectorizer()
vectorizer.fit(text)
vectorizer.vocabulary_
```

Output: `{'am': 0, 'student': 3, 'of': 2,
'uds': 4, 'not': 1}`

B. Classifiers:

As mention at the beginning of the report, I went through five traditional classifiers and two neural network techniques. They are...

1. *k-Nearest Neighbor:* It is a non-parametric method. k-NN classifier is used for classification and regression problem. This classifier is sensitive to irrelevant features. The primary goal before applying k-NN is feature selection. K=3 indicate it selects three data points nearest to the given class [10]. Then the algorithm chose the most frequent class among those three data points and finally, it returns the predicted class. The choice of the K value is very important to predict the class.

2. *Logistic Regression:* A Logistic model is a common approach for binary classification such that yes or no, 1 or 0, true or false, etc. The parameter of the logistic model is estimated by logistic regression [10].

3. *Random Forest:* Another popular traditional classification technique is a random forest. It consists of decision trees. Each subset of tree splits out a label or class prediction [10]. It measuring via Gini index or entropy.

4. *Gradient Boosting:* Gradient boosting algorithm is also used for regression and classification problem. It is also a tree-based algorithm. There are three elements in gradient boosting such as loss function, weak learner and additive model. Loss function depends on the dataset. Weak learner help to make a prediction. The additive model minimizes the loss of function by adding weak learner [11]. It cannot work well with a small dataset.

5. *Support Vector Machine:* Another common supervised technique is the support vector machine. It is known as a discriminative classifier. It is defined by a separating hyperplane [10]. By escalating the input to the higher dimensionality space, it solves the classification problem [4].

6. *MLP Classifers:* It is class of feed forward ANN. It makes efficient classifiers, which can offer superior performance compared to other classifiers, but are also criticized for the number of free parameters. Parameters are most usually set using either the validation package or the cross-validation techniques [12].

7. *Convolutional Neural Network:* Nowadays, Neural network is the most popular technique. Most of the researchers use neural network approaches in their research. CNN or convolutional neural network is used for image recognition and classification, text classification, etc. It consists of neurons. Neurons have learnable weights and biases where each of the neurons receives input. There is a loss function in the last layer [13].

C. Evaluation Matrices

The models are evaluated based on the following evaluation metrics.

- Accuracy.
- Precision.
- Recall.
- AUC-ROC Value.

IV. Results and findings

A. Twitter Datasets

I have applied all the traditional classifiers that are mentioned in the approaches section. From table 1, we can see the results for twitter datasets.

Datasets	**kNN**	**LR**	**RM**	**GB**	**SVM**
Accuracy	0.841	0.921	0.914	0.887	0.960
AUC-Value	0.806	0.914	0.923	0.872	0.934
Precision (Non-spam)	0.83	0.92	0.91	0.85	0.91
Precision (Spam)	0.86	0.93	0.94	0.97	0.92
Recall (Non-spam)	0.92	0.95	0.96	0.97	0.95
Recall (Spam)	0.73	0.88	0.87	0.76	0.90

Table 1: Classification results from Twitter dataset.

Support vector machine provides the highest accuracy (96%) and AUC (93.4%) value where k-Nearest neighbor classifier provides the lowest accuracy (84.1%) and AUC value (80.6%). The accuracy and AUC value for random forest and gradient boosting are almost similar. All classifiers except k-Nearest Neighbors and gradient boosting have accuracy and AUC over 90%.

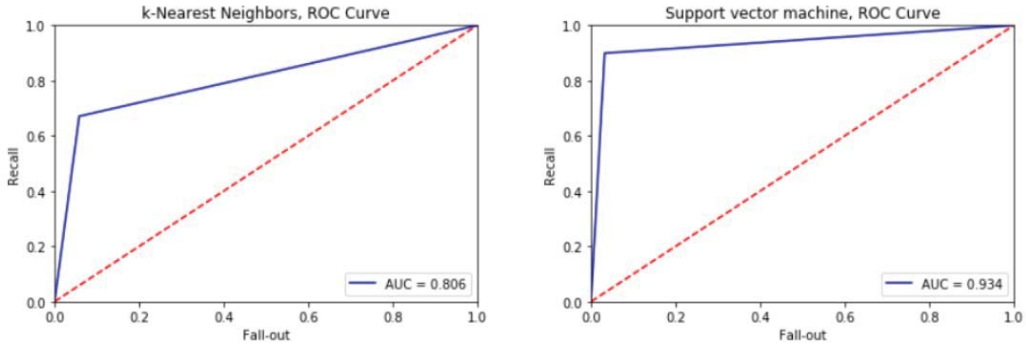

Figure 1. ROC curve of k-NN and SVM for Twitter Dataset.

ROC graph is a graphical plot that explains the power of classifier. It is constructed by plotting TP (True Positive) rate against FP (False Positive) rate. Since the AUC value for k-Nearest Neighbor is the lowest, it is not closest to (0, 1) in ROC graph.

Since support vector machine is the best classifier, we can make a prediction with this classifier in the test data. The sample of prediction results is shown below.

Tweet	Type
Obama Criminal Enterprise Collapsing https://...	Spam
I only learned to dream in sound #love	Quality
Cause I ain't trying to out here thinking you ...	Quality
When will they get that it's about #Liberty ? ...	Spam
GM UAW workers to receive profit-sharing up to...	Spam
RIP list of black people killed by police. Thi...	Spam
Heres a list of media companies who have donat...	Spam

Table 2. Prediction ’type’ against ’Tweet’ (‘Quality’ represents ‘Not Spam’).

B. Comparative Analysis with Facebook, Twitter and Email Data

Like Twitter dataset, Support vector machine provides the highest accuracy for both Facebook and Email datasets. In Facebook datasets, logistic regression estimates the highest AUC where support vector machine provides the highest AUC for both Twitter and Email datasets. From table 3, we can see that k-Nearest Neighbors measure the lowest accuracy and AUC for both Twitter and Email datasets. For Facebook, we have received the lowest accuracy the random forest. In overall, Logistic regression and support vector machine provide accuracy above 90% where k-Nearest Neighbor classifier never provides more than 87% accuracy for any datasets. For Facebook data, random forest measures the lowest AUC value. Gradient boosting also provide lower AUC value for Facebook datasets.

Classifiers	Dataset	Accuracy	AUC Value
k-Nearest Neighbors	Twitter	0.841	0.806
	Facebook	0.86	0.768
	Email	0.87	0.825
Logistic Regression	Twitter	0.921	0.914
	Facebook	0.917	0.833
	Email	0.949	0.843
Random Forest	Twitter	0.954	0.923
	Facebook	0.775	0.588
	Email	0.937	0.926
Gradient Boosting	Twitter	0.887	0.872
	Facebook	0.794	0.626
	Email	0.934	0.913
Support Vector Machine	Twitter	0.960	0.934
	Facebook	0.954	0.817
	Email	0.962	0.976

Table 3. Classification results comparison.

From table 3, we can define that logistic regression and support vector are the best classifier for all of the datasets. Support vector machine classifier is slightly better than logistic regression.

C. Neural Network (MLP Classifier)

I have applied simple MLP classifier technique to all of the datasets. Since we know that MLP Classifier consists of at least three layers of node, my MLP classifier architecture is following.

```
MLPClassifier(activation='tanh', alpha=0.0001, batch_size='auto', beta_1=0.9,
       beta_2=0.999, early_stopping=False, epsilon=1e-08,
       hidden_layer_sizes=(6, 2), learning_rate='adaptive',
       learning_rate_init=0.001, max_iter=400, momentum=0.9,
       n_iter_no_change=10, nesterovs_momentum=True, power_t=0.5,
       random_state=None, shuffle=True, solver='adam', tol=0.0001,
       validation_fraction=0.1, verbose=False, warm_start=False)
```

Figure 2: MLP classifier architecture

By applying the MLP Classifier, I have received the highest accuracy for the Twitter dataset and the lowest accuracy for Facebook dataset. The accuracy of Twitter dataset is 88.7% and the accuracy of Facebook dataset is 62.3%.

Dataset	Accuracy
Twitter	0.887
Facebook	0.623
Email	0.712

Table 4. MLP classifier results.

By measuring the traditional classifier and MLP classifier, I can say that Twitter dataset is the most organized dataset as it provides the highest accuracy all classifier. From traditional classifier, the support vector machine provides the best accuracy and AUC for Twitter.

D. Comparison with CNN

Another interesting neural network is convolutional neural network (CNN). I tried to go through the simple CNN architecture to compare my CNN results with the MLP classifier results. Since we know that CNN consists of multiple layers, my CNN architecture is made of 128 filters with the size of 5.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 100, 100)          5601800
_________________________________________________________________
conv1d_1 (Conv1D)            (None, 95, 128)           76928
_________________________________________________________________
global_max_pooling1d_1 (Glob (None, 128)               0
_________________________________________________________________
dense_1 (Dense)              (None, 10)                1290
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 11
=================================================================
Total params: 5,680,029
Trainable params: 5,680,029
Non-trainable params: 0
```

Figure 3: Simple CNN architecture

The table 5 shows the results of the CNN. Like MLP classifier, CNN offers the highest accuracy for the Twitter dataset and the lowest accuracy for the Facebook dataset. The accuracy of MLP classifier for the Twitter dataset is higher than the accuracy of CNN for the Twitter dataset.

Method	**Dataset**	**Accuracy**
MLP classifier	Twitter	0.887
	Facebook	0.623
	Email	0.712
CNN	Twitter	0.688
	Facebook	0.564
	Email	0.601

Table 5: MLP classifier vs. CNN

By comparing the statistics of both MLP classifier and CNN, I can say that MLP classifier provides relatively better results for all of our datasets. I have also found that both MLP classifier and CNN worked well on Twitter dataset and both worked poorly on the Facebook dataset. We can identify the spam based on the MLP classifier model if we concern about the neural network approach otherwise we can go with our best traditional classifier which is support vector machine

E. Results of Combined Dataset

Like other datasets, I have received the highest accuracy for support vector machines, but logistic regression yields the highest AUC value. The k-Nearest algorithm measures better accuracy and AUC value for the combined dataset. The accuracy of the kNN model is 90% for the combined dataset, where other 3 datasets provide less than 90% accuracy.

Traditional Classifiers	**Accuracy**	**AUC Value**	**Neural Networks**	**Accuracy**
k-Nearest Neighbors	0.90	0.826	MLP Classifier	0.808
Logistic Regression	0.935	0.921	CNN	0.692
Random Forest	0.934	0.912		
Gradient Boosting	0.936	0.913		
Support Vector	0.960	0.908		

Table 6: Combined dataset results.

Overall, the combined dataset measures good results for traditional classifiers. Similarly, MLP classifier gives a better result than CNN. Combined dataset overcomes the bad result for Facebook. Since the Facebook dataset is so small, the results for neural networks are not so good and no good AUC value is measured for traditional classifiers. The fusion of Twitter, Facebook and Email data improves accuracy for both traditional classifiers and neural networks.

V. Conclusion

This paper has brought an interesting technique for predicting social spam. Also, this report discusses the comparative differences between three different datasets via machine learning techniques. In this paper, analysis has been done by traditional classifiers and two neural networks. Data has been collected from three different sources where one dataset is a kaggle completion dataset. Spam has been identified for Twitter with the help of best classifier named SVM. I did not go through all the neural network approaches otherwise the research would be more interesting. Nowadays, spam is a very common term in the virtual world. Therefore, Identity spam is becoming prevalent in machine learning. The future work will be to work with live tweets and make honeypot for Facebook to collect the texts, and then apply all relevant deep learning approaches to find the best deep learning technique for the collected texts, and finally detect spam from those texts with the help of the best deep learning method.

References

[1] Statista.https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users

[2] Al-Radaideh, Q. A., & Al Nagi, E. (2012). Using data mining techniques to build a classification model for predicting employees performance. *International Journal of Advanced Computer Science and Applications*, *3*(2).

[3] Pal, A. K., & Pal, S. (2013). Evaluation of teacher's performance: a data mining approach. *International Journal of Computer Science and Mobile Computing*, *2*(12), 359-369.

[4] Zakaria, Ghada.Detecting Social Spamming on Facebook Platform. *Master's Thesis, University of Tartu,Estonia, 2018.*

[5] Ren, Y., & Zhang, Y. (2016, December). Deceptive opinion spam detection using neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 140-150).

[6] Anis, Sourial.Inferring the 2016 USA election foreign state-sponsored accounts on Twitter, *Saarland Uni-versity, Privacy Enhancing Technologies 2018*

[7] Twitter Dataset: *www.kaggle.com/c/utkml-twitter-spam/data*
[8] Facebook Dataset: *http://cucis.ece.northwestern.edu/projects/Social/sent imentdata.html*
[9] Email Dataset: *www.kaggle.com/balakishan77/spam-or-ham-email-classification/data*
[10] Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering, 160, 3-24*.
[11] Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics, 7, 21*.
[12] Windeatt, T. (2006). Accuracy/diversity and ensemble MLP classifier design. IEEE Transactions on Neural Networks, 17(5), 1194-1211.
[13] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.