International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307

OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/IJSRCSEIT

Developing the Framework Using Deep Neural Network for Detection of Spam and Fake Spam Messages in Twitter

[1]N. Anil Kumar, [2]Thatha Anusha, [3]Nagamalla Durga Prasad, [4]Maddala Bala Manikanta, [5]Boddu Swathi

[1]Associate Professor, [2,3,4,5] UG Students

Department of CSE, Sri Vasavi Institute of Engineering & Technology, Nandamuru, Andhra Pradesh, India

ARTICLEINFO

Article History:

Accepted: 10 April 2024

Published: 22 April 2024

Publication Issue

Volume 10, Issue 2

March-April-2024

Page Number

661-668

ABSTRACT

Social media plays vital role among the user communities for social gathering, entertainment, communication, sharing knowledge so on. Twitter is one such network to connect millions of users to share information. Nowadays, there are humpteen numbers of users using social media for social engagements. Due to the fact that wide publicity of individuals and products get viral in social media, everyone wish to use social media as a platform to promote their product. Furthermore, large number of people relies on social media contents to take decisions. Twitter is one of the social media platforms to post the media contents by the user. Spammers are illegal users intrude the twitter account and send the duplicate messages to promote advertisement, phishing, scam and personal blogs etc. In this paper, a novel spam detection mechanism is introduced to detect the suspicious users on twitter. The system has been designed such a way that it initially set with semi-supervised at the tweet level and update into supervised level for learning the input tweets to detect the spammers. The proposed system will also identify the type of spammers and will also remove duplicate tweets. We have applied with multi-classifier algorithms like naïve Bayesian, K-Nearest neighbor and Random forest into twitter data set and the performance is compared. The experimental result shows very promising results.

Keywords: Twitter Spam, Multi-classifier, Classification, Random forest, Bayesian, K-Nearest neighbor

I. INTRODUCTION

Social media is one of the platforms used by large number of users for learning, entertainment, promoting advertisement and social engagement. Through social media one can share the messages or information to millions of users at a time. Survey shows that individuals spent as an average of 144 minutes per day on social media since 2014 to 2019. In 2019, a total of 4.4 billion internet users worldwide and every minutes there are 4,79000 tweets generated. With the advent of internet and advancement of technology IoT

technology integrated technology with the smart devices which generates large volume of data. In 2020, it is expected that 44 zetta byte of data will be generated by various sources connected with internet. Nowadays social media is an important source for individuals and corporate for taking business decisions, report, and opinion and so on. Data is an important asset and with the help of the available technology, one can extract the essence of the data. Due to fact that a sizable amount of gain and wide publicity can be done with the social media, social spammers making use of the network and spread fake and false information to the media users. Twitter is the one of the most wanted media network used by its users. In this paper, we introduced a framework which used to process thousands of tweets per minute and able to detect the spammers. Furthermore the system deletes the spam messages. The proposed system equipped with semisupervise framework for spam detection and multi-classifier algorithm for differentiating the spam messages. The multi classifier algorithms which includes Naive Bayesian classifier, K-Nearest, Random Forest and Decision tree is applied with the data set and the accuracy is compared. The multiple classifier algorithms are efficiently detecting the spam messages. Furthermore, the system identifies duplicate or redundant spam messages. The similarities of the tweets are identified and categorized it namely leg data, spam data and total data. The legdata describes the particular person and the spam data describes how many of those posts are spams of that particular person. The total data is used to count the overall twitter data tweeted by everyone. The proposed spam detection system used to learn the tweets and the activities, and accurately classifying the new data inputs.

II.RELATED WORK

Title: Drifted Twitter Spam Classification using Multiscale Detection Test on K-L Divergence

Twitter spam classification is a tough challenge for social media platforms and cyber security companies. Twitter spam with illegal links may evolve over time in order to deceive filtering models, causing disastrous loss to both users and the whole network. We define this distributional evolution as a concept drift scenario. To build an effective model, we adopt K-L divergence to represent spam distribution and use a Multi scale Drift Detection Test (MDDT) to localize possible drifts therein. A base classifier is then retrained based on the detection result to gain performance improvement. Comprehensive experiments show that K-L divergence has highly consistent change patterns between features when a drift occurs. Also, MDDT is proved to be effective in improving final classification result in both accuracy, recall and f-measure. INDEX TERMS Concept drift, drift detection test, twitter spam classification, K-L divergence.

Title- A Constant Time Complexity Spam Detection Algorithm for Boosting Throughput on Rule based Filtering Systems

Along with the barbarous growth of spams, anti-spam technologies including rule-based approaches and machine-learning thrive rapidly as well. In anti-spam industry, the rule-based systems (RBS) becomes the most prominent methods for fighting spam due to its capability to enrich and update rules remotely. However, the anti-spam filtering throughput is always a great challenge of RBS. Especially, the explosively spreading of obfuscated words leads to frequent rule update and extensive rule vocabulary expansion. These incremental obfuscated words make the filtering speed slow down and the throughput decrease. This paper addresses the challenging throughput issue and proposes a constant time complexity rule-based spam detection algorithm. The algorithm has a constant processing speed, which is independent of rule and its vocabulary size. A new special data structure, namely, Hash Forest, and a rule encoding method are developed to make constant time complexity possible. Instead of traversing each spam term in rules, the proposed algorithm manages to detect spam terms by checking a

very small portion of all terms. The experiment results show effectiveness of proposed algorithm.

Title- A Framework for Real-Time Spam Detection in Twitter

With the increased popularity of online social networks, spammers find these platforms easily accessible to trap users in malicious activities by posting spam messages. In this work, we have taken Twitter platform and performed spam tweets detection. To stop spammers, Google SafeBrowsing and Twitter's BotMaker tools detect and block spam tweets. These tools can block malicious links, however they cannot protect the user in real-time as early as possible. Thus, industries and researchers have applied different approaches to make spam free social network platform. Some of them are only based on user-based features while others are based on tweet based features only. However, there is no comprehensive solution that can consolidate tweet's text information along with the user based features. To solve this issue, we propose a framework which takes the user and tweet based features along with the tweet text feature to classify the tweets. The benefit of using tweet text feature is that we can identify the spam tweets even if the spammer creates a new account which was not possible only with the user and tweet based features. We have evaluated our solution with four different machine learning algorithms namely - Support Vector Machine, Neural Network, Random Forest and Gradient Boosting. With Neural Network, we are able to achieve an accuracy of 91.65% and surpassed the existing solution [1] by approximately 18%.

Title- Detection of Social Network Spam Based on Improved Extreme learning Machine

With the rapid advancement of the online social network, social media like Twitter has been increasingly critical to real life and become the prime objective of spammers. Twitter spam detection refers to a complex task for the involvement of a range of characteristics, and spam and non-spam have caused unbalanced data distribution in Twitter. To solve the mentioned problems, Twitter spam characteristics are analyzed as the user attribute, content, activity and relationship in this study, and a novel spam detection algorithm is designed based on regularized extreme learning machine, called the Improved Incremental Fuzzy-kernel-regularized Extreme Learning Machine (I2FELM), which is used to detect the Twitter spam accurately. As revealed from the experience validation results, the proposed I2FELM can efficiently identify the balanced and unbalanced dataset. Moreover, with few characteristics taken, the I2FELM can more effectively detect spam, which proves the effectiveness of the algorithm. INDEX TERMS Social network, spam detection, spam features, machine learning.

Title- A Hybrid Approach for Detecting Automated Spammers in Twitter

Twitter is one of the most popular microblogging services, which is generally used to share news and updates through short messages restricted to 280 characters. However, its open nature and large user base are frequently exploited by automated spammers, content polluters, and other ill-intended users to commit various cyber crimes, such as cyberbullying, trolling, rumor dissemination, and stalking. Accordingly, a number of approaches have been proposed by researchers to address these problems. However, most of these approaches are based on user characterization and completely disregarding mutual interactions. In this study, we present a hybrid approach for detecting automated spammers by amalgamating communitybased features with other feature categories, namely metadata- , content-, and interaction-based features. The novelty of the proposed approach lies in the characterization of users based on their interactions with their followers given that a user can evade features that are related to his/her own activities, but evading those based on the followers is difficult. Nineteen different features, including six newly defined features and two redefined features, are identified for learning three classifiers, namely, random forest, decision tree, and Bayesian network, on a real dataset that comprises benign users and spammers. The discrimination power of different

feature categories is also analyzed, and interaction- and community-based features are determined to be the most effective for spam detection, whereas metadata-based features are proven to be the least effective.

III. PROPOSED SYSTEM

The proposed system equipped with semi-supervise framework for spam detection and multi-classifier algorithm for differentiating the spam messages. The multi classifier algorithms which includes Naive Bayesian classifier, K-Nearest, Random Forest and Decision tree is applied with the data set and the accuracy is compared. The multiple classifier algorithms are efficiently detecting the spam messages. Furthermore, the system identifies duplicate or redundant spam messages. The similarities of the tweets are identified and categorized it namely leg data, spam data and total data. The legdata describes the particular person and the spam data describes how many of those posts are spams of that particular person. The total data is used to count the overall twitter data tweeted by everyone. The proposed spam detection system used to learn the tweets and the activities, and accurately classifying the new data inputs.

The proposed spam and fake spam message detection framework contains four main detector modules such as hashtag-based features, content-based features, user-based features, and domain-based features. The architecture of the proposed system is shown in figure 1. The system trained with black listed spamming domain and tested accordingly. The four detectors successfully classify the spam and non-spam tweets in the tweet window. Using the semi supervised method the required information is updated periodically based on the previous tweet window experience.

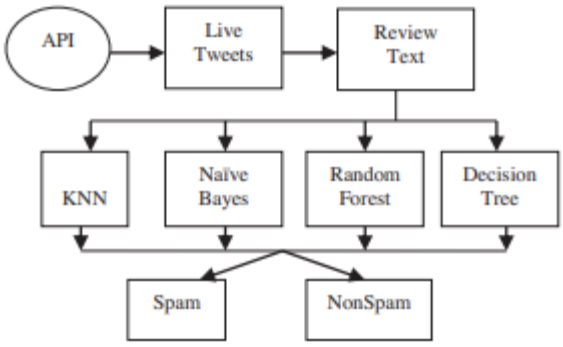

Fig 1: System Architecture

Advantages:

1. Proposed system focused on knn-algarithm, naviebayes algorithm, random forest algorithm, decision tree algorithm to find the spam or non spam accuracy's.
2. Spam or non spam accuracy's comparison to other algorithms Random forest algorithm accuracy is very high accuracy.

IV. RESULTS AND DISCUSSION

In [104]:

```
import pandas as pd
import numpy as np

from sklearn.feature_extraction.text import CountVectorizer, TfidfTransformer
from sklearn.linear_model import SGDClassifier, LogisticRegression
from sklearn.naive_bayes import MultinomialNB
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import RandomizedSearchCV
from sklearn.metrics import accuracy_score, classification_report

from scipy.stats import randint
```

Fig 2. Results screenshot

```
In [105]:
logreg = LogisticRegression()
logreg.fit(X_train_df, Y_train)

Out[105]: LogisticRegression()

In [106]:
# Generate predictions
prediction["Logistic"] = logreg.predict(X_test_df)

# Score predictions
accuracy_score(Y_test, prediction['Logistic'])

Out[106]: 0.812888198757764
```

Fig 3. Results screenshot

```
In [107]:
# Generate another classification report
print(classification_report(Y_test, prediction['Logistic'], target_names = ['Ham', 'Spam']), sep='\n')

              precision    recall  f1-score   support

         Ham       0.78      0.48      0.59       738
        Spam       0.82      0.95      0.88      1838

    accuracy                           0.81      2576
   macro avg       0.80      0.71      0.74      2576
weighted avg       0.81      0.81      0.80      2576
```

Fig 4. Results screenshot

```
In [109]:
# Generate predictions
prediction["SVM"] = svm.predict(X_test_df)

# Score predictions
accuracy_score(Y_test, prediction['SVM'])

Out[109]: 0.8012422360248447

In [110]:
print(classification_report(Y_test, prediction['SVM'], target_names = ['Ham', 'Spam']), sep='\n')

              precision    recall  f1-score   support

         Ham       0.90      0.34      0.50       738
        Spam       0.79      0.99      0.88      1838

    accuracy                           0.80      2576
   macro avg       0.85      0.66      0.69      2576
weighted avg       0.82      0.80      0.77      2576
```

Fig 5. Results screenshot

```
Out[111]: DecisionTreeClassifier()

In [112]:
# Generate predictions
prediction["Tree"] = tree.predict(X_test_df)

# Score predictions
accuracy_score(Y_test, prediction['Tree'])

Out[112]: 0.7666925465838509

In [113]:
print(classification_report(Y_test, prediction['Tree'], target_names = ['Ham', 'Spam']), sep='\n')

              precision    recall  f1-score   support

         Ham       0.60      0.57      0.58       738
        Spam       0.83      0.85      0.84      1838

    accuracy                           0.77      2576
   macro avg       0.71      0.71      0.71      2576
weighted avg       0.76      0.77      0.76      2576
```

Fig 6. Results screenshot

```
In [117]:
print("Tuned Tree Parameters: {}".format(tree_cv.best_params_))
print("Best score is {}".format(tree_cv.best_score_))

Tuned Tree Parameters: {'criterion': 'gini', 'max_depth': None, 'max_features': 4, 'min_samples_leaf': 2}
Best score is 0.7249902661425038

In [118]:
# Reusing the same param_grid we used for the logistic regression, since both classifiers take a C value
svm_cv = GridSearchCV(svm, param_grid, cv = 5)
svm_cv.fit(X_train_df, Y_train)

print("Tuned SVM Parameters: {}".format(svm_cv.best_params_))
print("Best score is {}".format(svm_cv.best_score_))

Tuned SVM Parameters: {'C': 3.727593720314938}
Best score is 0.8040632679804235
```

Fig 7. Results screenshot

```
In [119]:
alpha_values = np.arange(0.1, 4, 0.1)
alpha_grid = {'alpha': alpha_values}

nb_cv = GridSearchCV(nb, alpha_grid, cv = 5)
nb_cv.fit(X_train_df, Y_train)

print("Tuned NB Parameters: {}".format(nb_cv.best_params_))
print("Best score is {}".format(nb_cv.best_score_))

Tuned NB Parameters: {'alpha': 2.8000000000000003}
Best score is 0.8039346554633106
```

Fig 8. Results screenshot

II. CONCLUSION

In this paper, a novel framework for detecting spam tweets has been presented. The proposed system extracts the live tweets successfully through TwitterAPI. The system successfully extracts the data from the lists of review text. Multi-classifiers algorithms such as KNN, Naïve Bayes, Random Forest and Decision tree algorithms are applied with the dataset and the performance is compared. The multiclassifier algorithms successfully classify the tweets into spam and not spam tweets respectively. Among the algorithm with this dataset, Random classifier algorithm showcased with highest accuracy. The prediction mechanism successfully identifies with the binary classification of spam messages and no spam messages. Furthermore, the identified spam messages can be deleted.

VI. FUTURE WORK

We used more than 10,000 tweets obtained through the dataset . The tweets are labeled as spam and ham in all tweets and the remaining portion of tweet is labeled as unknown. The tweets which are labeled as unknown due to the fact that they are not able determine their labels using manual observation. The twitter data is is classified using the four machine learning algorithms such as KNN, Random Forest, Bayesian and Decision Tree algorithms. The performance of the algorithm is compared with the metrics precision, recall, F-measure and accuracy. The proposed system considers the spam class as positive and non-spam class as negative. The system incorporates with Top-30 words of user-based features in the tweet text so that the system predicts the spam contents. This is helps to fine tune the system to identify spam contents and mitigate the loss due to spam. This property contributes to detect spam messages in real time by assigning 50% for training set and testing set.

III. REFERENCES

[1]. Wu, Tingmin, et al., "Twitter spam detection: Survey of new approaches and comparative study". Computers & Security. 76.10.1016/j.cose.2017.11.013.

[2]. M. Jiang, et al., "Suspicious behavior detection: Current trends and future directions," IEEE Intelligent Systems, vol. 31, pp. 31-39,2016.

[3]. J. Tanha, et al., "Semi-supervised self-training for decision tree classifiers," International Journal of Machine Learning and Cybernetics,vol. 8, pp. 355-370, 2017.

[4]. Y. Xia, et al., "A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring," Expert Systems withApplications, vol. 78, pp. 225-241, 2017.

[5]. S. Sedhai and A. Sun, "Semi-supervised spam detection in Twitter stream," IEEE Transactions on Computational Social Systems, vol.5, pp. 169-175, 2017.

[6]. S. Liu, et al., "Addressing the class imbalance problem in twitter spam detection using ensemble learning," Computers & Security, vol.69, pp. 35-49, 2017.

[7]. C. Chen, et al., "Investigating the deceptive information in Twitter spam," Future Generation Computer Systems, vol. 72, pp. 319-326,2017.

[8]. G. Lin, et al., "Statistical twitter spam detection demystified: performance, stability and scalability," IEEE Access, vol. 5, pp. 11142-11154, 2017.

[9]. C. Chen, et al., "Statistical features-based real-time detection of drifted twitter spam," IEEE Transactions on Information Forensics andSecurity, vol. 12, pp. 914-925, 2016.

[10]. A. Singh and S. Batra, "Ensemble based spam detection in social IoT using probabilistic data structures," Future Generation ComputerSystems, vol. 81, pp. 359-371, 2018.

[11]. C. Li and S. Liu, "A comparative study of the class imbalance problem in Twitter spam detection," Concurrency and Computation:Practice and Experience, vol. 30, p. e4281, 2018.

[12]. R. Aswani, et al., "Detection of spammers in twitter marketing: a hybrid approach using social media analytics and bio inspiredcomputing," Information Systems Frontiers, vol. 20, pp. 515-530, 2018.

[13]. A. T. Kabakus and R. Kara, ""TwitterSpamDetector": A Spam Detection Framework for Twitter," International Journal of Knowledgeand Systems Science (IJKSS), vol. 10, pp. 1-14, 2019.

[14]. X. Wang, et al., "Drifted Twitter Spam Classification Using Multiscale Detection Test on KL Divergence," IEEE Access, vol. 7, pp.108384-108394, 2019.

[15]. B., Mukunthan. (2019). Improved Content Based Medical Image Retrieval using PCA with SURF Features. International Journal ofInnovative Technology and Exploring Engineering. 8. 10.35940/ijitee.J1020.08810S19.

[16]. M.Arunkrishna, B.Mukunthan " Review on Classification of Anti-Spam Solutions : Approaches, Algorithms Demystified." Studies inIndian Place Names Vol. 40 No. 60 (2020): Vol-40-Issue-60-March-2020 , vol. 40, no. 60, 6 Mar. 2020, pp. 4449–4458.

[17]. Mukunthan B, Nagaveni N. Identification of unique repeated patterns, location of mutation in DNA finger printing using artificialintelligence technique. Int J Bioinform Res Appl. 2014;10(2):157-176. doi:10.1504/IJBRA.2014.059516

[18]. A, Pushpalatha & B, Mukunthan. (2010). Automation of DNA Finger Printing for Precise Pattern Identification using Neural-fuzzyMapping approach. International Journal of Computer Applications. 12. 10.5120/1761-2411.

[19]. V. Vishwarupe, et al., "Intelligent Twitter spam detection: a hybrid approach," in Smart Trends in Systems, Security and Sustainability,ed: Springer, 2018, pp. 189-197.

[20]. C.C. Wei and N.S. Hsu,"Derived operating rules for a reservoir operation system: Comparison of decision trees, neural decision treesand fuzzy decision trees,"Water resources research,vol.44, 2008.

[21]. Mukunthan, B. & Nagaveni, N.Nagaveni. (2011). "Automating Identification of Unique Patterns, Mutation in Human DNA usingArtificial Intelligence Technique". International Journal of

Computer Applications. 25. 26-34. 10.5120/3003-4038.

[22]. H. Tajalizadeh and R. Boostani, "A novel stream clustering framework for spam detection in twitter," IEEE Transactions onComputational Social Systems, vol. 6, pp. 525-534, 2019.

[23]. B., Mukunthan. (2019). Improved Content Based Medical Image Retrieval using PCA with SURF Features. International Journal ofInnovative Technology and Exploring Engineering. 8. 10.35940/ijitee.J1020.08810S19.