doi:10.20944/preprints202007.0648.v1

A Method Based on NLP for Twitter Spam detection

Ratul Chowdhury[1], Kumar Gourav Das[2], Banani Saha[3], and Samir Kumar Bandyopadhyay[4]

[1] Department of Computer Science and Engineering, Future Institute of Engineering and Management, Kolkata, India, ratul.chowdhury@teamfuture.in

[2]Department of Computer Science and Engineering, Future Institute of Engineering and Management, Kolkata, India, kumargouravdas18@gmail.com

[3]Department of Computer Science and Engineering, University of Calcutta, Kolkata, India, bsaha 29@yahoo.com

[4] Department of Computer Science and Engineering, University of Calcutta, Kolkata, India, 1954samir@gmail.com

Abstract

Social networking applications such as Twitter have increasingly gained significance in terms of socio-economic, political, and religious as well as entertainment sectors. This in turn, has witnessed a wide gamut of information explosion in the social networking realm that can tend to be both useful as well as misleading at the same point of time. Spam detection is one such solution that caters to this problem through identification of irrelevant users and their data. However, existing research has so far laid primary focus on user profile information through activity detection and relevant techniques that may underperform when these profiles exhibit characteristics of temporal dependency, poor reflection of generated content from the user profile, etc. This is the primary motivation for this paper that addresses the aforementioned problem of user profiles by focusing on both profile information and content-based spam detection. To this end, this work delivers three significant contributions. Firstly, exhaustive use of Natural language processing (NLP) techniques has been rendered towards creation of a new comprehensive dataset with a wide range of content-based features. Secondly, this dataset has been fed into a customized state-of-art hybrid machine learning model that has been exclusively built using a combination of both machine learning and deep learning techniques. Extensive simulation based analysis not only records over 98% accuracy but also establishes the practical applicability of this proposal by proving that modeling based on the mixed profile and content-generated data is more capable of spam detection in contrast to each of these standalone approaches. Finally, a novel methodology based on logistic regression is proposed and supported by analytical formulations. This paves the way for the custom-built dataset to be analyzed and corresponding probabilities to be obtained that differentiate legitimate users from spammers. The obtained mathematical outcome can henceforth be used for future prediction of user categories through appropriate parameter tuning for any given dataset. This makes our method a truly generic one capable of identifying and classifying different user categories.

Key-words: Twitter, Social Media, NLP, Tweet, User Categorizations and Mathematical Frame Work

Introduction

The impact of social media has brought drastic changes in the past few years in terms of socio-economic as well as organizational development. Facebook, twitter, LinkedIn are the most leading social media platform that enable users to interact with each other by both sharing, consuming information and building meaningful connection with people. As twitter data is freely available and has huge content so its derived features are very much effective for the researchers to work on various domains like spam detection, Personality identification, sarcasm detection, event detection, etc. This huge micro blogging platform has more than 313 million monthly active users tweeting around 350,000 tweets per minutes which is around 500 million tweets per day [1] Furthermore, Twitter is also infected by spammers for their private or organizational gain. Recent reports from Twitter depicts that around 9.9 million spammy twitter accounts were identified per week. Twitter spam [2] also known as unsolicited tweets contains malicious information and links. Various unfair means are used by spammers to spread their spams such as using abusive and bold languages as a reply to the users to seek attention, posting some hostile links, creating redundant profiles which can be created either by using automated tools or by manually, posting identical updates and trolling latest links to catch attention. A few spam and non spam accounts along with their respective tweets have been listed in the table1.Spam tweets often consist of Uniform resource locators (URLs) having links which are either adult contemporary or out of the context content. With the help of those URLs, spammers redirect the users to those mischievous sites which contain viruses. By using spoofing, they get personal information of the users. Spammers are adopting technologies like 'bot' [3-4] which automatically follows massive number of readers per day.

USER ID	USERNAME	TWEET	CATEGORY
6301	@msjacbowie	Come to "The Ruby Revue (Sydney)" 07 February from 18:00 to 22:00. Book now at http://www.therubyrevue.com VOTED... http://bit.ly/6bdHth	SPAM
10836	@rex_huang	@old_cat ??????? http://bit.ly/8Hfp5	SPAM
817045	@bizopinsiderr	Promoting Your SMB With Facebook And Twitter http://www.business-opportunities.biz/2009/12/07/promoting-your-smb-with-facebook-and-twitter/	SPAM
614	@dcharrisonon	@holman dang. Just read your preceding tweet, and now my quip isn't so much funny as moronic. Reverse-chron 1, Christian Harrison 0.	NON SPAM
9375	@santhoshj	Anyone in #Chennai wish to go out for a #weekend #trek? CTC is going #Tada. It will be a moderate trek (Average difficulty)!	NON SPAM

Table 1: Description of spam non spam tweets

The researchers and anti-spam team of twitter are collaboratively trying to restrict the spammers on user level as well as on tweet level. Twitter has applied a branch of restrictions in recent past, it has suspended the accounts which behaves abnormally and trimmed the number of accounts that a user can follow in a day. As per the reports published in 2019 [5], a verified account can follow up to 1000 accounts per day and in case of unverified account, the count reduces to 400. Moreover, an active twitter user can follow up to 5000 accounts. For additional following, it must receive followers to a certain threshold. Researchers on the other hand are applying various content based, URL based, graph based and account based methods for spam detection [6-7]. The above methods have been further categorized into clustering, classification and hybrid problems. The various graph based, account based, content based and URL based features are listed in figure 1. Broadly, spam detection has been performed either at user level or at tweet level. Although tweet-level detection can identify spam tweets in real-time, it is increasingly difficult to capture user-level characteristics from a single tweet. On the contrary, user-level detection facilitates in providing more distinctive information of a particular user from the profile. Thus user-level detection makes it more suited to uniquely identify a legitimate user from a spammer and is therefore adopted in this paper.

In this paper along with the account based feature we have proposed other four new content based features namely stylistic, embedded, topic word based and hashtag based features. The applicability of the proposed features have been verified by various machine learning, deep learning and hybrid frameworks. So the main contributions of the paper are:

1. Creation of a more upright dataset contains 1200 legitimate and 800 spam accounts.
2. In tune with the previous contribution the next focus of this paper is the application of various NLP methods for feature extraction. Different novel techniques like Latent Dirichlet allocation (LDA), Latent Semantic allocation(LSA) have been used for creation of various derived features.
3. Exhaustive investigation on the applicability of the derived feature is carried out through various machine learning, deep learning and hybrid methods. This establishes the feasibility and part superiority of the proposed methodology over and above existing methods in literature.
4. A novel methodology based on logistic regression is further proposed and implemented on the given dataset. Subsequent analytical framework delivers corresponding probabilities for legitimate and spam users and can further be used for user categorization on any given dataset.

The rest of the paper is structured as follows, section 2 reflects the literature survey portion, section 3 describes dataset creation and preprocessing, section 4 shows the feature extraction, section 5 represents the proposed model, section 6 analyzes the results and performs various comparisons. Furthermore, section7 introduces and implements the proposed methodology for user categorization using a mathematical framework based on logistic regression. Finally, the paper is concluded in section 8.

Literature survey

In twitter environment, to broadcast an event, it is necessary to identify the relevant group of users related to that event. Twitter spam is a dispensable information or event for a particular group of user. In a report, Nexgate [8] mentioned that on an average within every 200 social media post, there is at least one spam tweet and according to [9] approximately 15% of the tweeter users are automatic bots. The demand of various social media platforms are increasing rapidly. According to the social media statistics [10] within 2020 approximately one third of the global population will be connected to social media. So spam detection and identification is an ongoing social threat.

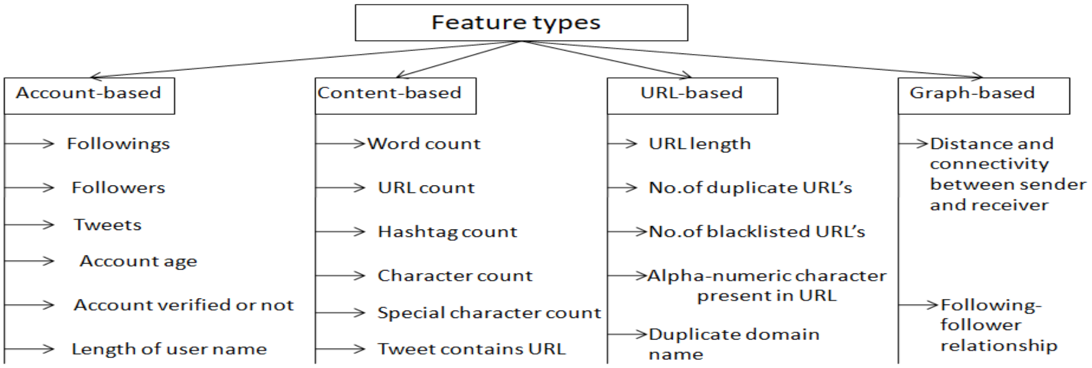


Figure 1: Various account-based, content-based, URL-based and graph-based feature description

This section describes different research works performed by the researchers based on various account based, graph based, content based and URL based technique in user level as well as tweet level. In [11-15] the researchers have used various account based features described in figure 1. But in recent scenario, these account based feature can be easily fabricated. Among various graph based methods, in [16] the authors have estimated a relationship between sender and receiver, the connectivity represents the strength of the connection. Their experimental result shows that most of the spam comes from account that has less relation with its receivers. In another work, Alex hai wang [17] proposed a directed social graph model to identify the follower and friend relationship. They have used 4 graph based features namely follower, friend, mutual friend and stranger. Their proposed model records above 90%

accuracy. Amleshwaram et al [18] presented a CATS system based on some Bait-oriented, behavioral entropy oriented, URL based and content entropy based features. Their work provides low latency and fast detection rate. In the past few years, authors have used various machine learning based approach [19-23] in spam detection. Bayesian, KNN, SVM, RF are popularly used with various derived features. Wu et al [24] presented a deep learning approach. According to the authors due to the problem of information fabrication and spam drift concept the real life spammer are not efficiently detected using traditional machine learning method.In this work the Word2vec output of the Ground-Truth twitter dataset is fed into MLP for classification. The proposed model shows above 90% accuracy. In another work, Madisetty et al [25] proposed a neural network based ensemble approach using deep CNN. Here CNN is combined with Twitter Glove, Google News, Edinburgh and random embedding. Among two different dataset HSPAM14 result is not up to the mark. In [26-27] the authors have also focused on various lightweight lexical features of URL like number of URL, length of URL, number of alphanumeric character present in each URL and others [28-29] have used various real time features of URL like HTTP Header, DNS, java script event, html content, etc. But lee et al [30] shows that the above mentioned scheme can't catch suspicious URL that repeats long time interval. They proposed a sliding window based suspicious URL detection technique called WARNINGBIRD which can identify suspicious URL in long and short time interval. In tweet level Shedhai et al [31] proposed a semi supervised approach where tweets are categorized into four groups:1. Blacklisted domain detector 2. Near duplicate detector 3. Reliable hap tweet detector 4.Multicast based detector but the number of listed features considered in this work are extremely large. In a very recent studies Inuwa-Dutse et al[32] shows that most of the studies relay on large number of count based historical feature which not only creates an extra overhead on detection but also easily fabricated by the influence of bots.

Recent studies in literature have primarily focused on count based historical features for spam detection. Considering the problem of their dynamic variation, it can be concluded that there is a lot of room for improvement. The modification can be effectively performed by embedding new features, analyzing the contents of maximum number of tweets of a particular user and by applying various hybrid and ensemble approaches.

Dataset creation and preprocessing

This section describes the details of various data collection and preprocessing procedures. Two different datasets have been considered for experimental purpose. The first one is social honey pod(SHP) [33] dataset, collected in between Dec 30, 2009 to Aug 2, 2010 that contains 22,223 and 19,276 number of polluters and legitimate users respectively. Another dataset is a custom dataset which is a manually created dataset containing 1200 and 800 number of legitimate and spam users respectively. As mentioned in previous section, the recent scenario spammers are more intelligent, may pretend as legitimate users and after gaining acceptance from other users they post spam tweets [31]. So to find the applicability of the derived features in recent scenario, the custom dataset was prepared in 2019. For legitimate class, the tweets of different verified users have been collected via tweepy API. More than 2000 tweets per user have been considered, where the users belong to different categories like sports, business, politics, education etc. For polluted class, a spammy keyword dictionary has been applied. The dictionary was created by Snovio in 2019 that contains more than 550 catchy spam trigger words. In custom dataset, the unverified twitter accounts containing more than two spammy keywords are marked as polluted. The details of the custom dataset preparation are shown in figure 2.

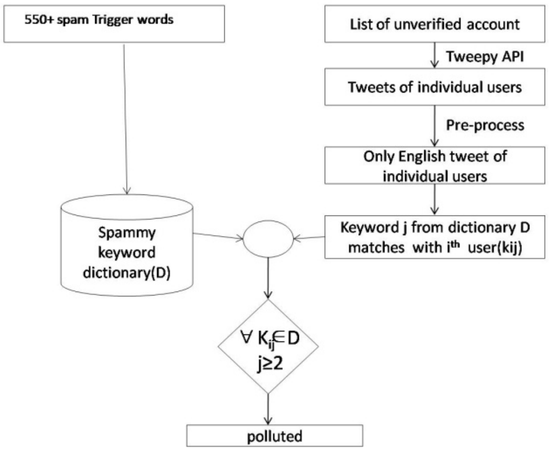

550+ spam Trigger words
List of unverified account
Tweepy API
Tweets of individual users
Pre-process
Only English tweet of individual users
Keyword j from dictionary D matches with ith user(kij)
Spammy keyword dictionary(D)
∀ Kij∈D
j≥2
polluted

Figure 2: Custom dataset preparation framework

Feature extraction

In order to perform the experiment,5 different types of lightweight features like account based, stylistic, hashtag oriented, embedded and topic word based features are used here. The description of the individual features and its purposes are shown in table 2.

Account based feature

Account based features are the outer information about the profile. Five different types of account based features such as profile age, reputation score, tweet frequency, length of screen name and length of profile descriptions have been considered in this category.

Feature Name	Description
Profile age	Present data-date of account creation
Reputation score	$\frac{\text{Follower}}{\text{Following} + \text{Follower}}$
Tweet frequency	$\frac{\text{Tweet count}}{\text{Profile age}}$
Screen Name	Length of the screen name of a particular user
Profile Description	Length of the profile description of a particular user

Table 2: Description of various account based features

The profile age is selected here as the spammers generally create new accounts to replace their suspended accounts within a short interval of time. So the age of the spam account is generally less as compared to genuine account. Similarly, the reputation score creates a significant distinction among spam user, celebrity user and genuine user based on their following follower relationship. For genuine user the value of the reputation score is less than one, for celebrity user the value is nearly equal to one and for spam user the value is nearly equal to zero. Tweet frequency is considered as a next feature because a spam user sends a bulk number of tweets with in a short period of time. So the tweet frequency count is very high in case of spammers. The next two significant features are length of screen name and length of profile description. Since the spammers frequently generate new account so they do not provide proper information into their profile.

Stylistic feature

Stylistic features employ a significant role to identify the symmetric variations of natural languages. 8 different types of stylistic features that are used here are shown in table 3. For the number of happy emotions, sad emotions, emoji and slangs individual dictionary has been prepared under each category. Stop of words has been collected from NLTK corpus and the rest of the features are extracted programmatically.

Feature Name	Description
Happy emoticons	Total number of happy emoticons present in a user tweet
Sad emoticons	Total number of sad emoticons present in a user tweet
Emoji	Total number of emojis present in a user tweet
Slang words	Total number of slang words present in a user tweet
Stop-words	Total number of stop-words words present in a user tweet
Hashtag	Total number of hash tag words present in a user tweet
Punctuations	Total number of punctuations present in a user tweet
URL frequency	$\frac{\text{URL count}}{\text{Tweet count}}$

Table 3: Description of various stylistic features

The spammers generate enormous number of tweets with repeated URLs for promotional purpose that contains very fewer number of emoticons, stop words, hash tags and punctuation as compared to the genuine user. In case pornographic user the slang word counts are very high.

Hashtag based feature

Hashtags are the type of metadata tag used in various social media platform which allows users to apply dynamic, user-generated tagging facilities to make it possible for others to easily find messages on a specific topic or phenomenon. The hashtag based features are the information containing in the hashtag. Latent semantic analysis is a popular natural language processing technique used here to extract the vector representation of the hashtags, The LSA has been applied on each and every hashtags of a particular user and a 25bits vector has been prepared. LSA works in 3 steps.

1) Initially, a separate technique called term frequency-inverse document frequency (tf-idf) is used to find the frequency of the word in each document.
2) In next step, singular value decomposition (SVD) is applied on tf-idf for dimensionality reduction.
3) Finally, LSA is used as vector.

Word embedding based feature

Word embedding is the vector representation of a particular text where the words with the similar meaning have similar representation. In present work, Word embedding based feature are the vector representation of similar words present in a tweet. GloVe and word2vec model are two popular vector representation technique used in various NLP applications. By considering the various advantages of GloVe over traditional word2vec, the vector has been prepared using GloVe.

Topic word based feature

Topic words are the important keyword present in a particular document. In this work, to identify the topic word present in a tweet, Latent Dirichlet Allocation (LDA) topic modeling technique is applied.

Proposed model

The proposed model as illustrated in Figure 3 comprises of 4 primary sections – i) Tweet extraction; ii) preprocessing; iii) Feature extraction; iv) Train-Test split. Both SHP and custom dataset initially contains the raw tweets. In preprocessing phase, all the non English tweets have been removed. Whereas in feature extraction phase, a 338 bits feature vector has been prepared for training. The feature vector 1-8 containing the stylistic feature, vector 9-33 represents a 25 bits hashtag feature. The next 200 bits are the embedded feature containing the Glove output of the tweets. A 100 bits topic word based feature has been added next by applying LDA followed by LSA on the tweet respectively. Finally, 5 account based features are considered.

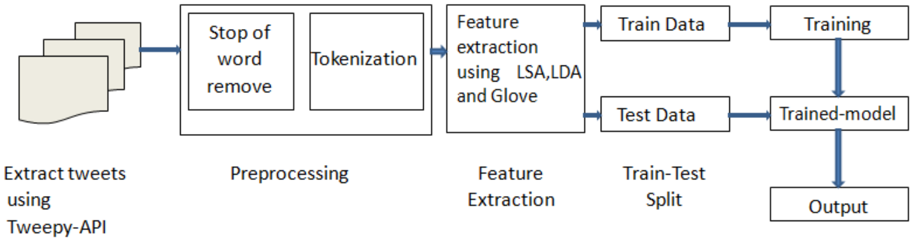


Figure 3: Complete framework for spam detection

In next phase, these 338 bits feature vector has been fed into various machine learning classifier for training. For deep learning framework a 20*20 feature vector has been prepared. The intermediate structure of the CNN model is shown in figure 4. For CNN-SVM, the dense layer output of the CNN is transferred to SVM for classification.

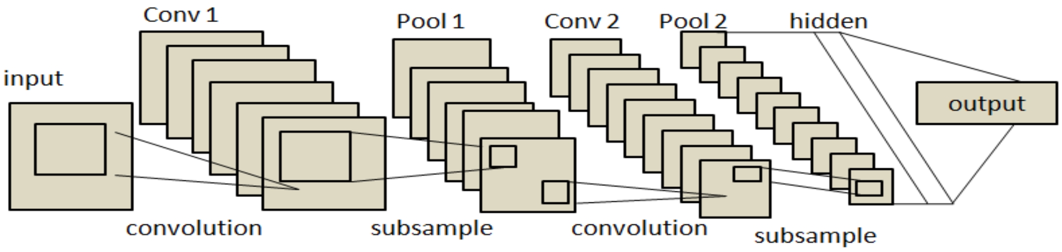


Figure 4: Detailed CNN architecture

5. EVALUATION METRICS

The proposed method used 5 important performance indicators to evaluate the performance of the model.

1. Accuracy (AC): It is the percentage ratio of correctly classified instances by the total number of instances.

doi:10.20944/preprints202007.0648.v1

$$AC= \frac{TP+TN}{TP+TN+FP+FN}$$

2. Precision: It is the ratio of relevant instances by the retrieved instances.

$$Precision= \frac{TP}{TP+FP}$$

3. Recall: It is the ratio of relevant instances that have been retrieved by the total amount of relevant instances.

$$Recall= \frac{TP}{TP+FN}$$

4. F1 Score: It is the weighted average of precision and recall.

$$F1\ Score= \frac{2*(Recall*Precision)}{Precision+Recall}$$

TP (True positive): is the number of positive instances that are corrected classified.

TN (True negative):is the number of negative instances that are correctly classified.

FP (False positive):is the number positive instances that are wrongly classified.

FN (False negative):is the number of negative instances that are wrongly classified.

Hence a good IDS always have a good detection rate and low false alarm rate.

Experimental set up and Result Analysis

For preprocessing, feature extraction and classification purpose, the method used the most promising machine learning package Keras 2.1.6 in python 3.6.4 environment. All the experiments have been conducted on a personal notebook with Intel core i7-2760M CPU @ 2.70GHz configuration and 8GB memory. All the experiments have been performed for two dataset SHP and custom. The applicability of the individual and combined features has been investigated through four different machine learning, two deep learning and one hybrid classifiers. Finally, a critical comparison has been done between two datasets with the above mentioned parameters.

The directive of the first experiment is to evaluate the effectiveness of the proposed features for SHP dataset. Table 4 shows the accuracy achieved through various machine learning, deep learning and hybrid frameworks. The table clearly shows that the MLP records the highest accuracy (93%) for combined feature while profile based features also record higher detection rates as compared to others. Similarly, table 5 shows the accuracy of Custom dataset. Here the stylistic, embedded and combined feature sets have contributed a significant role where specifically stylistic feature itself records above 98% accuracy for RF and CNN-SVM classifiers.

Classifier	Stylistic feature	Hashtag feature	Embedded feature	Topic word feature	Account feature	Combined feature
NB	62.81	59.18	62.3	74.98	63.36	89.1
J48	77.46	89.45	77.39	74.18	89.11	90.3
SMO	73.4	67.85	82.14	81.95	71.78	86.96
RF	78.85	83.95	84.5	81.56	89.34	88.6
MLP	79	48.8	85.9	81.8	90.3	93
CNN	79.8	56.1	84.2	81.6	90.5	92.5
CNN-SVM	79.16	49.95	76.55	76.05	88	88.8

Table 4: Performance evaluation of used features based on various classifiers for SHP dataset

Classifier	Stylistic feature	Hashtag feature	Embedded feature	Topic word feature	Account feature	Combined feature
NB	94.6	40.35	94.5	86.9	65.1	95.05
J48	98	89.45	94.65	86.9	91.3	97.55
SMO	94.45	67.85	95.45	88.85	60	95.05
RF	98.15	93.95	97.05	86.9	92.9	97.5
MLP	97.5	57.8	97	86.8	91.7	96.75
CNN	97.8	73	96	84	87.8	97.5
CNN-SVM	98	62.5	97	62.5	89.5	97.7

Table 5: Performance evaluation of used features based on various classifiers for custom dataset

In the next section, a graphical comparison has been performed with respect to both custom and SHP datasets in terms of three parameters: precision, recall and F1–Score. This graphical comparison has been made solely for Stylistic, embedded and combined feature sets which have acted as primary contributors for differentiation between legitimate users and spammers. Figure 5 (a) shows precision comparison. Recall comparison is shown in figure 5 (b). F-Measure comparison for combined feature is presented in figure 5 (c). Figures 6 (), 6 (b) and 6(c) show Precision Comparison, Recall comparison and F-Measure comparison for embedded features.

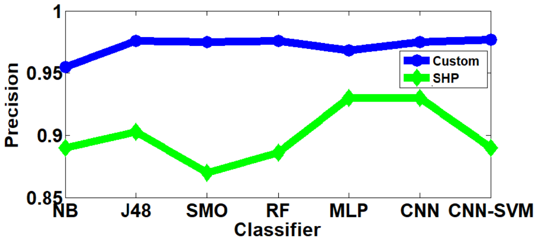


Figure 5a: Precision comparison for combined feature

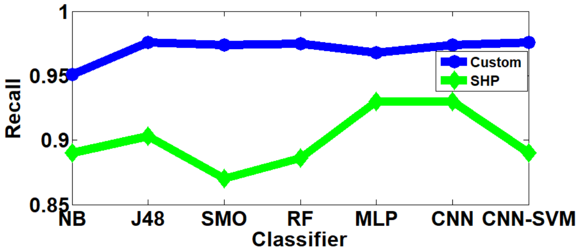


Figure 5b: Recall comparison for combined feature

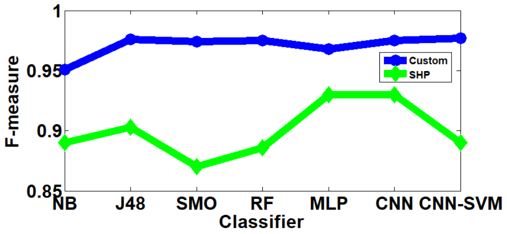


Figure 5 c: F-Measure comparison for combined feature

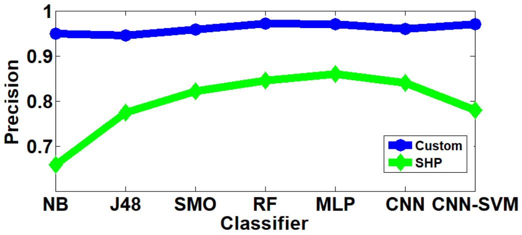
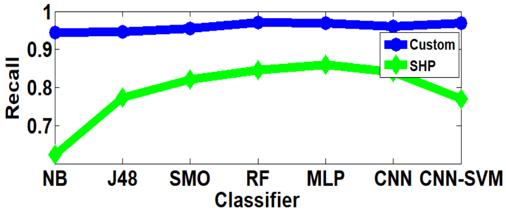


Figure 6a: Precision Comparison for embedded feature



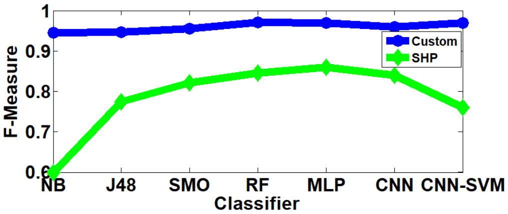
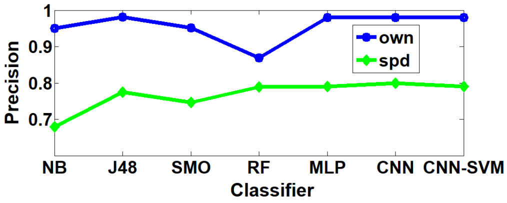
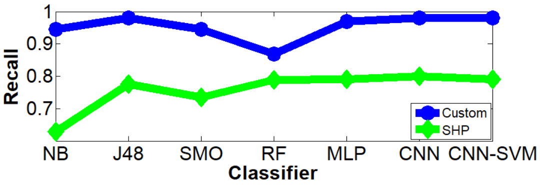


Figure 6b: Recall comparison for embedded feature

Figure 6c: F-Measure comparison for embedded feature





(a)

(b)

Figure 7a: Precision comparison for stylistic feature

Figure 7b: Recall comparison for Stylistic feature

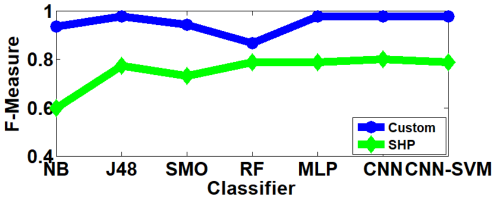


Figure 7c: F-Measure comparison for Stylistic feature

Figure 7 (a), (b) and (c) show the same for stylistic feature. Combining both the accuracy outcome as obtained from Table 4, 5 and the aforementioned illustrations are in Figures 5, 6, and 7 of the proposed features record outstanding results for Custom dataset. Further, it may be noted that although the Profile based feature plays a significant role for SHP dataset, it is outperformed by stylistic feature set when it comes to custom dataset. This is because the usage of various Stylistic features has increased manifold in modern-day tweets. Moreover, embedded features also play a crucial role as reflected in the outcome. The vector representation of Glove output creates a drastic difference between Spam and Legitimate user in Custom Dataset that naturally enhances the detection capability with respect to the embedded features.

Although 98% accuracy was observed in [33] for the SPD dataset, it has several shortcomings in comparison to our work as discussed in the following.

i) Limited no. of tweets: It is quite noticeable that this dataset was formatted on a limited number of tweets of a particular user (less than 200) which is not significant enough to evaluate their interests. This is because spammers in modern times are observed to behave as legitimate users initially. Only after being rendered as a legitimate user by others, they begin their spamming activities. Under such circumstances, limited number of tweets is insufficient to completely validate the authenticity of a particular user.

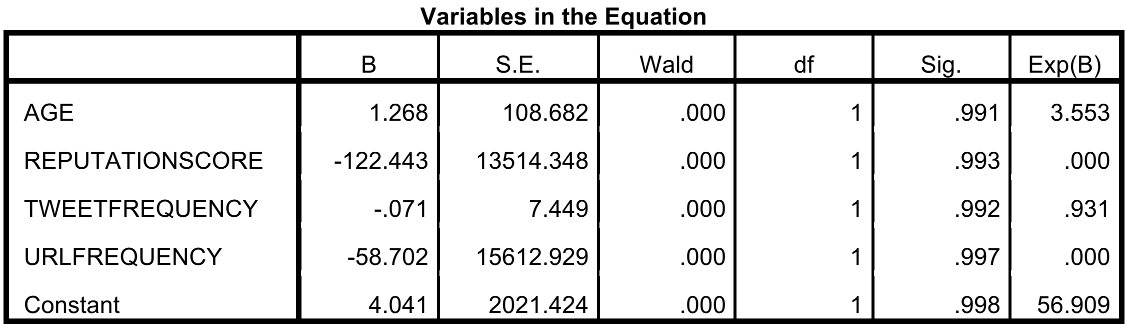
ii) Overtly dependent on account-based features: Modern-day spamming activities are highly characterized by the content of their tweets in contrast to the superficial user-related information such as followers, followings, user credentials, etc. This is because account based features can be easily fabricated, thereby rendering spam detection relying on account-based features highly ambiguous. Hence, the work in [33] falls short of providing statistically significant accurate results in today's scenario due to its over-reliance on account-based features.

The proposed work, on the contrary, addresses both these shortcomings effectively. Accordingly, in our custom dataset, the maximum number of tweets for a particular user is 2000. In addition, above 98% detection rate is achieved successfully based on only 8 stylistic features derived specifically for this purpose. As the stylistic features are based on contents of the tweet it cannot be fabricated. Therefore, the accuracy obtained by our proposed model is highly significant under the current scenario of spam detection.

Mathematical formulation for user categorization

Finally, this section delivers the proposed methodology for user categorization in real time using logistic regression based mathematical model. Accordingly, in this section, the attributes are first selected based on their significance and mathematical formulation is thereafter devised based on the working principal of Logistic regression. Finally validation of the proposed procedure is performed through graphical analysis over a randomly selected data.

To perform the user classification in real time here we have used only 4 high significant valued features Profile age, reputation score, tweet frequency and URL frequency. Logistic regression classifier has been used here because at training time it provides some B weights for each of the individual features which are further been used for predicting classification probabilities in real time. The significant features and its corresponding B values are shown in table 6.



Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
AGE	1.268	108.682	.000	1	.991	3.553
REPUTATIONSCORE	-122.443	13514.348	.000	1	.993	.000
TWEETFREQUENCY	-.071	7.449	.000	1	.992	.931
URLFREQUENCY	-58.702	15612.929	.000	1	.997	.000
Constant	4.041	2021.424	.000	1	.998	56.909

Table 6: Significant features and its corresponding B values

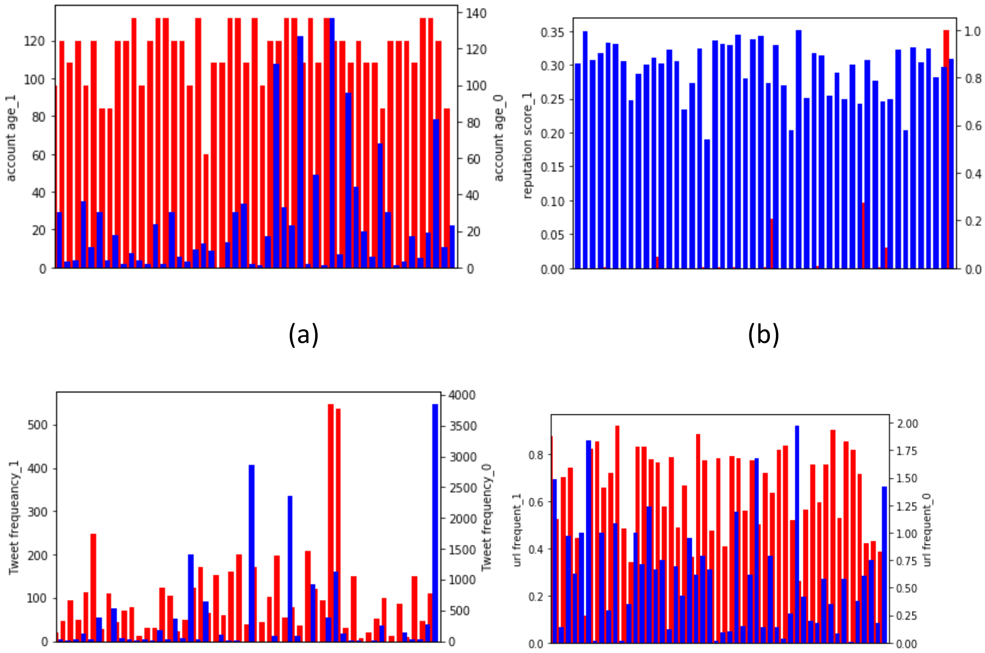
According to the working principal of Logistic regression the corresponding B values are further formulated by the following equation for providing the classification probabilities of any user.

Logit=Constant + Profile age * B1 + Reputation score * B2 + Tweet frequency * B3 + URL frequency * B4

Odds= e^{Logit}

Classification Probabilities=$\frac{Odds}{1+Odds}$

We have randomly chosen 100 accounts, 50 legitimate and 50 spammers for validation purpose. The corresponding feature values in terms of graph are shown in figure 8.



(a) (b)

(c) (d)

Figure 8 Feature values are presented

These feature values are formulated through equation 5, 6 and 7 respectively. According to the resultant classification probability 97 accounts was correctly classified and only 3 was misclassified.

Conclusions

This paper proposes and implements a novel framework for real time spam detection considering profile based as well as content-based features. Use of both profile based and content-based feature detection allows this paper to alleviate the problems in related works in literature where spam detection suffers from the problems of temporal dependency as well as standalone profile based features. Specifically, the proposed work in this paper implements a three-pronged approach towards accurate and real time spam detection. Firstly, a comprehensive and feature intensive data set is generated by deploying NLP techniques that lays the foundation for executing a novel hybrid learning framework involving both machine learning as well as deep learning algorithms in the second step. Finally, this enables the design of an analytical model based on logistic regression to differentiate legitimate users from spammers using distinguishable set of probabilities for each class of users. The experimental results demonstrate the effectiveness of the derived feature for SPD and custom dataset. Due to the limited tweet

doi:10.20944/preprints202007.0648.v1

information in SPD dataset the custom dataset has been prepared by considering more than 2000 tweets of a particular user. The proposed stylistic and embedded features are found to be superior based on limited number of users. Overall 98% accuracy is observed in classifying users as either legitimate ones or spammers in the given data set. The proposed model also provides a suitable research platform for further studies particularly related to multi label spam classification as well as its impact on demographic user profiling.

Conflict of Interest

There is no conflict of interest of any author.

References

[1]H. Tsukayama, "Twitter turns 7: Users send over 400 million tweets per day," Washingon Post, March 2013. [Online]. Available: http://articles.washingtonpost.com 2013-03- 21/business/37889387 1 twitter-jack -dorsey-twitte

[2] Alom, Zulfikar, Barbara Carminati, and Elena Ferrari. "Detecting Spam Accounts on Twitter." In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 1191-1198. IEEE, 2018.

[3] Gilani, Zafar, Liang Wang, Jon Crowcroft, Mario Almeida, and Reza Farahbakhsh. "Stweeler: A framework for twitter bot analysis." In Proceedings of the 25th International Conference Companion on World Wide Web, pp. 37-38. International World Wide Web Conferences Steering Committee, 2016.

[4] Haustein, Stefanie, Timothy D. Bowman, Kim Holmberg, Andrew Tsou, Cassidy R. Sugimoto, and Vincent Larivière. "Tweets as impact indicators: Examining the implications of automated “bot” accounts on T witter." Journal of the Association for Information Science and Technology 67, no. 1 (2016): 232-238.

[5]https://blog.twitter.com/en_us/topics/company/2018/how-twitter-is-fighting-spam-and-malicious-automation.html

[6] Wu, Tingmin, Sheng Wen, Yang Xiang, and Wanlei Zhou. "Twitter spam detection: Survey of new approaches and comparative study." Computers & Security 76 (2018): 265-284.

[7] Lalitha, L. A., Vishwanath R. Hulipalled, and K. R. Venugopal. "Spamming the mainstream: A survey on trending Twitter spam detection techniques." In 2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon), pp. 444-448. IEEE, 2017.

[8] NexGate, State of Social Media Spam Research report, Online, Accessed:18-02-2018 (2013).

[9] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, A. Flammini, Online Human-Bot Interactions: Detection, Estimation, and Characterization, in: International AAAI Conference on Web and Social Media, AAAI Press, 2017, pp. 280–289.

[10] Social Media Statistics and Facts, Online: www.statista.com/topics/1164/social-networks, Accessed: 18-02-2018.

[11]G. Stringhini, C. Kruegel, and G. Vigna, “Detecting Spammers onSocial Networks,” Proc. 26th Ann. Computer Security ApplicationsConf. (ACSAC), 2010.

[12] C. Grier, K. Thomas, V. Paxson, and M. Zhang, “@spam: TheUnderground on 140 Characters or Less,” Proc. 17th ACM Conf.Computer and Comm. Security (CCS), 2010.

[13]K. Lee, J. Caverlee, and S. Webb, “Uncovering Social Spammers:Social Honeypots þ Machine Learning,” Proc. 33rd Int’l ACMSIGIR Conf. Research and Development in Information Retrieval, 2010.

[14] Agarwal, S. and Sureka, A., 2015, February. Using knn and svm based one-class classifier for detecting online radicalization on twitter. In International Conference on Distributed Computing and Internet Technology (pp. 431-442). Springer, Cham.

[15] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida,“Detecting Spammers on Twitter,” Proc. Seventh Collaboration,Electronic Messaging, Anti-Abuse and Spam Conf. (CEAS), 2010.

[16] Song, Jonghyuk, Sangho Lee, and Jong Kim. "Spam filtering in twitter using sender-receiver relationship." In International workshop on recent advances in intrusion detection, pp. 301-317. Springer, Berlin, Heidelberg, 2011.

[17] Wang, Alex Hai. "Don't follow me: Spam detection in twitter." In 2010 international conference on security and cryptography (SECRYPT), pp. 1-10. IEEE, 2010.

[18] Amleshwaram, Amit A., Narasimha Reddy, Sandeep Yadav, Guofei Gu, and Chao Yang. "Cats: Characterizing automation of twitter spammers." In 2013 Fifth International Conference on Communication Systems and Networks (COMSNETS), pp. 1-10. IEEE, 2013.

[19] Chen, C., Zhang, J., Xie, Y., Xiang, Y., Zhou, W., Hassan, M.M., AlElaiwi, A. and Alrubaian, M., 2015. A performance evaluation of machine learning-based streaming spam tweets detection. IEEE Transactions on Computational social systems, 2(3), pp.65-76.

[20] Mccord, M. and Chuah, M., 2011, September. Spam detection on twitter using traditional classifiers. In international conference on Autonomic and trusted computing (pp. 175-186). Springer, Berlin, Heidelberg.

[21] Wang, B., Zubiaga, A., Liakata, M. and Procter, R., 2015. Making the most of tweet-inherent features for social spam detection on twitter. arXiv preprint arXiv:1503.07405.

[22] Singh, Monika, Divya Bansal, and Sanjeev Sofat. "Who is who on twitter–spammer, fake or compromised account? a tool to reveal true identity in real-time." *Cybernetics and Systems* 49, no. 1 (2018): 1-25.

[23] Wang, A.H., 2010, July. Machine learning for the detection of spam in twitter networks. In International Conference on E-Business and Telecommunications (pp. 319-333). Springer, Berlin, Heidelberg.

[24] Wu, T., Liu, S., Zhang, J. and Xiang, Y., 2017, January. Twitter spam detection based on deep learning. In Proceedings of the Australasian computer science week multi conference (p. 3). ACM.

[25] Madisetty, S. and Desarkar, M.S., 2018. A neural network-based ensemble approach for spam detection in Twitter. IEEE Transactions on Computational Social Systems, 5(4), pp.973-984.

[26] McGrath, D.K. and Gupta, M., 2008. Behind Phishing: An Examination of Phisher Modi Operandi. LEET, 8, p.4.

[27] J.-L. Lee, D.-H. Kim, and L. Chang-Hoon, “Heuristic based Approach for Phishing Site Detection Using URL Features,” CEET-2015, 2015.

[28] Ma, J., Saul, L.K., Savage, S. and Voelker, G.M., 2009, June. Beyond blacklists: learning to detect malicious web sites from suspicious URLs. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1245-1254). ACM.

[29] Thomas, K., Grier, C., Ma, J., Paxson, V. and Song, D., 2011, May. Design and evaluation of a real-time url spam filtering service. In 2011 IEEE symposium on security and privacy (pp. 447-462). IEEE.

[30] Lee, S. and Kim, J., 2013. Warningbird: A near real-time detection system for suspicious urls in twitter stream. IEEE transactions on dependable and secure computing, 10(3), pp.183-195.

[31] Sedhai, S. and Sun, A., 2017. Semi-supervised spam detection in Twitter stream. IEEE Transactions on Computational Social Systems, 5(1), pp.169-175.

[32] Inuwa-Dutse, I., Liptrott, M. and Korkontzelos, I., 2018. Detection of spam-posting accounts on Twitter. Neurocomputing, 315, pp.496-511.

[33] Lee, K., Eoff, B.D. and Caverlee, J., 2011, July. Seven months with the devils: A long-term study of content polluters on twitter. In Fifth International AAAI Conference on Weblogs and Social Media.