DOI: 10.21917/ijsc.2022.0393

SOCIAL MEDIA SPAM DETECTION USING DIFFERENT TEXT FEATURE SELECTION TECHNIQUE AND MACHINE LEARNING

Anubha Sharma and Manoj Ramaiya
Department of Computer Science and Engineering, Sage University, India

Abstract

The messaging systems and social media is popular and has essential contributions to our social and professional life. Similarly, Spam is a part of the messaging system and social media. In social media, spam is found in various places (i.e. in posts, in comments, in reviews, and in chatting). Social media Spam is aimed to influence the user's decision, point of view, and credibility of the service or brand. Therefore, social spam detection is essential. However, using the social media data a number of contributions are available in literature, but a fewer amount of work is available for social media spam detection. In this paper, we proposed a social media spam detection technique using machine learning and text feature extraction techniques. In this context first, a review on social media spam detection techniques has been carried out. Using this review, we extract the different machine learning techniques used, techniques of text feature selection, and experimental datasets used. In this review, we found that the spam messages with the URLs are more critical and harmful. Next step, we design a theoretical model for social media spam detection, which includes text feature selection techniques (i.e. TF-IDF, POS, and Information Gain) and their combinations (POS+TF-IDF and POS+IG). These features are used with Support Vector Machine (SVM), Artificial Neural Network, and Naïve Bayes classifier for training. Experimental analysis with dataset available in Kaggle we found that hybrid features is more effective for accurate classification as compared to individual features. Additionally, we found for classification the SVM and ANN are more accurate as compared to the Bayes classifier.

Keywords:

Social Media Spam, Experimental Analysis, Text Feature Selection, Classification, Social Spam Filtering

1. INTRODUCTION

Social media becomes a part of our life, every age group people are enjoying social media services. It is the low cost, easy to use and efficient way to reach new people and information. Due to this a significant amount of content promoters, advertisers and spammers are using the social media to distribute content and gain the attention of audiences [1]. But some of the content promoters are utilizing the false, misleading and catchy contents to gain traffic. Additionally some of them are also utilizing these platforms for their toxic intensions [2]. Therefore identification and elimination of spam from social media is essential task. In this paper we carried out study for providing a brief overview of the some essential concepts of social media spam filtering.

The spam filtering may include the text mining techniques, and text mining can be specified by a different meaning [4]. Thus the models based on text mining can formulate the problem of social media spam filtering as Information Extraction, classification and categorization. Secondly, we need to understand what the spam messages are. In this context, we can say when the message apps are used to send malicious message is called Spamming. The malicious contents may be in form of short messaging service (SMS), social media post, social media chat, what's app chatting and other. Therefore, spamming is the practice of distributing commercial and unwanted messages. In some cases online advertising is also termed as Spam. The automated bots are also used for publishing such contents.

These messages may also contents external links, with the dual goal, the links are used for increasing visibility of an advertised product or service, secondly in some cases these links are used for phishing, to spread computer viruses, Trojan horses, or others [5]. It is not possible to prevent spammers from sending such messages but the amount of spam can be reduced by anti-spam application. There is different spam filtering techniques available [6]:

- **List-Based Filters** are categorizing sender as spammers or trusted and by blocking the message, tried to stop spam.
- **Content-Based Filters** evaluate words or phrases to determine a message is spam or legitimate.
- **Other Filtering** utilizes content and list-based filtering techniques to prevent spam.

However the spam classification is the task of text processing but now in these days Natural Language Processing (NLP) is also used for this task. The NLP is a technique of understanding, analyzing, manipulation, and generation of language using computational algorithms. The NLP applications are able to convert text into a computer friendly structure. The aim is to provide simpler representations of the text [7]. Using the NLP techniques can also organize and structure knowledge to perform summarization, relationship extraction, translation, sentiment analysis, entity recognition, speech recognition, topic segmentation and more [8]. In this study we tried to use the concept of NLP also for detecting the social media spam. This section provides the overview of the spam and the different concepts which can be used for designing the spam detection techniques. The next section reports the recent contributions in social media spam detection.

2. LITERATURE REVIEW

This section includes recent efforts and techniques contributed by researchers for spam filtering.

Ishtiaq et al. [9], the use of Graph centrality metrics for spam detection. Degree, eccentricity and closeness are used for classification. Graphs for each class are classified using the centrality scores. Results show the high precision and recall using degree centrality.

Additionally, in search of more accurate method H. Raj et al. [12] proposed a method based on LSTMs. The Word2Vec has been used to convert text into a vector. Results prove that the method outperformed.

Similarly, K. Zainal et al. [14] focus on feature extraction in spam. The objective is to control the features and considering its information or influence in spam classification.

Pinandito et al. [10] design an Android app to allow developers to build their own spam detection using this library. K-Nearest Neighbor and Naïve Bayes are implemented to identify spam. Twitter spam detection is a problem due to high variability in the language and short texts.

Thus, Jain et al. [11] propose a deep learning architecture based on CNN and LSTM. The model is supported by semantic information using WordNet and ConceptNet. This improves performance of vector representation. Results show the effectiveness of approach.

Dutse et al. [13] present an approach for distinguishing spam vs. non-spam social media posts and offers insight to the behavior of spam users. The features related to the users, their accounts and engagement with others are used. They show efficacy and robustness of approach and compare it to typical features for spam detection.

Ho et al. [15] allow accessing various user and content-based features, evaluate and compare performance and new feature implementation has demonstrated.

Cornelissen et al. [16] propose the use of feature, Socio-Informatics, in combination with existing methods for bot detection.

For sentiment analysis the resources are described by Itani et al. [17]. For informal Arabic does not conform grammar or spelling, a feature of corpora and lexicons is developed. Also provide useful NLP datasets to understanding the approach.

Janabi et al. [18] describe a model to detect malicious content. Multisource features have been used to detect posts that contain malicious URLs. With the feature combination, a random forest model was used. A recall of 0.89 without feature selection was produced. After parameter tuning, and feature selection it is perform to 0.92.

Mawass et al. [19] use similarity between users to correct evasion-induced errors. A Markov Random Field model on the similarity graph used to link similar accounts. A graphical model with a supervised classifier is used and tested. P. Tehlan et al. [20] proposed a method to detect spam using fuzzy logic and analyze through neural network multilayer Perceptron. The aim is to learn and apply the ML algorithms to overcome the limitations of supervised learning.

Zhang et al. [21] employ semantic analysis to build self-extensible dictionary which updates and extends automatically. The semantic analysis brings extra features that help in text classification. Achieve an average detection accuracy of 93.6%.

In the same way an architecture that utilizes SVM and Neural Network is proposed by Dhawan et al. [22]. They describe a hybrid clustering NEUROSVM.

Shehnepoor et al. [23] propose NetSpam, which utilizes spam features for modeling heterogeneous information networks to classify spam. The results show that NetSpam outperforms.

Tingxuan et al. [24] is design and evaluate several classification methods. Experiments on the Yelp dataset reveal that state-of-the-art classification algorithms can achieve the performance of 72.5%. The experiments also show that deep learning techniques can be used for classification.

S. Chancellor et al. [25] develop a deep learning classifier that jointly models textual and visual characteristics. Using a million Tumblr photos, the classifier discovers deviant content with high recall (85%).

T. Green et al. [26] address the problem of spam users on Wikipedia. They use a binary classification and propose a set of features. They tested the system on a dataset built of 4.2K users. The approach reaches 80.8% accuracy and 0.88 mean precision.

T. Wu et al. [27] investigate current methods have achieved accuracy around 80%. Due to spam drift and information fabrication, ML methods cannot efficiently detect spam. They proposed a deep learning method. The syntax of each tweet will be learned through Word Vector and binary classifier.

Y. A. Amrani et al. [28] focus on the Sentiment analysis from the messages using search. Messages can be classified as positive or negative based on certain aspects.

3. REVIEW AND CONCLUSION

Now in these days social media has used for various different purpose such as advertisement, promotions, business analytics, brand building, and more. All these activities require content development and distribution. On the other hand, the social media healthy environment is also necessary. The spam in social media may impact the creditability and environment. Thus, we need an effective social media Spam filtering technique. In this context, recently a significant effort has been carried out for designing social media spam filtering techniques. The social media spam detection techniques involve text mining and ML techniques. These techniques are developed to recognize the spam in different form of social media contents such as WhatsApp messages, chat applications, social post, SMS, emails, reviews, and comments. The detection is become more crucial when the messages with the URLs, because such spam content are become threat of security, finance, and privacy. Additionally, the detection of such malicious messages is very complex due to limited size of data and limited features to be extracted.

However, not only the message of financial advertises can be a spam sometimes the messages which are not under receiver's interest are also termed as spam. Additionally, sometimes the spam may influence the users' decision. In this context, we have collected a total of 50 research articles; among 20 of them which are most relevant has been summarized in table 1. According to the collected literature there are fewer methods which are developed using list; but most of the recent techniques are developed using ML and NLP. In addition, to develop the spam filters mostly the supervised learning algorithms are used such as SVM, ANN, and CNN. Additionally some of the methods are also developed based on fuzzy logic, Naïve Bayes and other statistical analysis techniques. The ML based spam filtering methods includes three main phases, i.e. preprocessing, feature selection and classification. The aim of preprocessing is to reduce the noise, and make content clean to utilize with ML algorithms. Similarly, the feature selection is used for find the optimal keywords and phrases to identify the spam contents. Finally, the ML classifier is applied to learn the features and recognize the spam.

Table.1. Review summary

Ref. no.	Research Domain	Algorithm	Features	Dataset	Results
[9]	spam SMS detection	Graph centrality	Closeness	Labeled 5574 SMS (4827 ham, 747 spam)	94.4% accuracy
[10]	Twitter spam	K-Nearest Neighbor and Naïve Bayes	stemming	Twitter API	Getting accuracy up to 88%
[11]	Twitter spam	CNN and LSTM	WordNet and ConceptNet	SMS spam and Twitter dataset	SMS Spam (99.01) and tweeter (94.70)
[12]	spam SMS detection	LSTM	Word2Vec	Spam SMS datasets	97.5% accuracy
[13]	Spam social media posts	Maximum Entropy, Random Forest, Extremely Randomized Trees, C-SV Classification, Gradient Boosting and MLP.	users, their accounts and pair wise engagement with others	Honeypot annotated spam-posts dataset and manually annotated spam-posts dataset	Getting accuracy up to 99%
[14]	SMS spam	Survey	NA	NA	NA
[15]	Spam social media posts	Evaluate and compare the performance of existing systems	user and content-based features	NA	higher true positives and lower false positives
[16]	Spam bot detection	Ensemble Learning	Socio-informatic Feature	Publicly available data	F1-score 0.899
[17]	sentiment analysis	Dialectal Arabic	corpora and lexicons	NLP data sets	100%
[18]	malicious content/URL in social media post	random forest	Multisource features to detect posts with malicious URLs	Twitter streaming API	recall (0.89) without feature selection, after feature selection (0.92)
[19]	Social spam detection	Markov Random Field and SVM	similarity graph	Twitter	Accuracy (0.952)
[20]	detect spam	fuzzy logic and neural network multilayer Perceptron	NA	tweets	FUZZY (58%) and ANN (73%)
[21]	detect spam comments in social media	Chi-Square	Duplicate comments, Noun Proportion, Hyperlink Amount, Emotional Score	Sina Weibo API	accuracy of 86.8%
[22]	community of spam data	NEURO-SVM	weight matrix	Tweeter, Facebook and Google +	Up to 85% accuracy
[23]	spam reviews	NA	review-behavioral, user-behavioral, review-linguistic, user-linguistic	NA	NA
[24]	social opinion spam	collective classification, deep learning	NA	Yelp dataset	72.5% accuracy
[25]	deviant content	deep learning	textual and visual	Tumblr photos	Recall (85%)
[26]	Spam users on Wikipedia	Support Vector Machine (SVM)7, Logistic Regression, KNearest Neighbor, Random Forest, and XGBoost.	Average size of edits, Standard deviation of edit sizes, Variance significance, Mean time between edits, Standard deviation of time between edits	4.2K users and 75.6K edits	80.8% accuracy and 0.88 mean average precision
[27]	spam drift and information fabrication	deep learning	WordVector	Twitter	Accuracy 99.35%
[28]	Sentiment analysis from the messages	PART, SVM, Decision Tree, Naive Bayes, Logistic Regression	query based on terms	SMS, Facebook, Twitter	NA

4. PROPOSED MODEL

The main aim of this study is to study the different text feature selection technique and measure the influence of feature selection on classifier's performance to detect the social media spam. In this context, an experimental model has developed and demonstrated Fig.1. The model consists of five main parts, which are explained in this section.

4.1 TWITTER SPAM DATASET

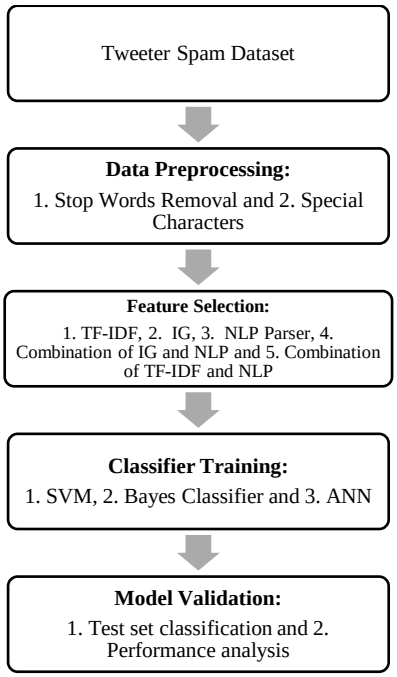

Fig.1. Proposed Model for Performance Assessment

The social media spam detection requires a suitable dataset, but there are very limited datasets are available for this task. In literature most of the authors are utilizing their owned datasets. But we have found a dataset from the Kaggle [29]. This dataset is known as "UtkML's Twitter Spam Detection Competition" dataset. The dataset contains seven attributes as:

Table.2. Dataset Attributes and description

Attributes	Description
Tweet	This is the text that was tweeted
Following	The number of people the account that tweeted is following
Followers	The number of people following the account that tweeted
Actions	The total number of favorites, replies, and retweets of said tweet
is_retweet	Binary [0,1] value: If 0 its not a retweet, if 1 it is a retweet
Location	The self-written location provided by the user on their profile, May not exist, be "Unkown", and is NOT standardized! ex. could be ("NY", "New York", "Upper East Side", Etc!)
Type	Either Quality or Spam

In this experiment we have considered this dataset as the benchmark dataset.

4.2 DATA PRE-PROCESSING

However, the dataset consists of sever essential attributes but in this work we just considering only the "Tweet" attribute to measure the influence of short text features with classification performance. Thus, we apply the following steps for data preprocessing:

Extraction of Tweets from dataset: the data has processed first for extracting the Tweet attribute from. The process of extracting the twits is described in table 3. According to the defined process the dataset is first read, and the rows and columns are calculated. After that each row of Tweet is captured separately into a list variable T. if the twit contains the URL then we eliminate the URL from the text and then store it to the list variable T. According to the literature we found that the messages or post with URLs may be much harmful, but due to limited coverage of the proposed study we are not considering the URLs. Thus we eliminated the URLs and only text data will store in the variable T.

Table. 3. Data Pre-processing

Input: Dataset D
Output: Filtered Tweets T

Process:

```
[row,col]=readDataset(D)
  for(i=1;i<row;i++)
    for(j=1;j<col;j++)
      if(Dj =="Tweet" )
        if(Dj.contains(URL))
          Dj.Remove(URL)
          T.Add(Dj )
        else
          T.Add(Dj)
        end if
      end if
    end for
  end for
```

Eliminating special characters from Tweets: The tweets may contain the significant amount of special characters. In social media most of the users are utilizing the special characters in their post and comments. Therefore for ease in data processing we have eliminated the special characters from the twits.

Eliminating stop words from Tweets: Similarly the Tweets also consist of unwanted words which are not much essential in distinguishing the orientation or subject of Tweets thus we also remove them from the tweets. The table 4 consists of the process used for eliminating the stop words and special characters from the twits.

Table. 4. Eliminating Stop words and Special Characters

Input: List of stop words S, List of Special Characters C, List of Tweets T
Output: Preprocessed Data P

Process:
for(*i*=1;*i*<*T.length*;i++)
  temp=T_i
  for(*j*=1;*j*<*S*.length;*j*++)
    temp=temp.FindReplace(S_j, " ")
  end for
  for(*k*=1;*k*<*C.length*;*k*++)
    temp=temp.FindReplace(C_k, " ")
  end for
  P_i.add(*temp*)
End for
Return *P*

4.3 FEATURE SELECTION

After dataset preprocessing we have found the cleaned tweets, for feature extraction. Here we want to implement and test the effectiveness of the different text feature selection techniques therefore we have implemented the following approaches for feature selection:

TF*IDF: The TF-IDF is also termed as term frequency and inverted document frequency. That is an improvement on bag of words; sometimes words may not provide much information. Due to this domain specific words do not have larger score. Thus, the frequency is rescaled by considering frequent words occur in entire documents. This way of scoring is known as TF–IDF [30].

- TF is frequency of the word in current document.
- IDF is score of the words among all the documents.

These scores can highlight the words that are unique and represent useful information.

NLP Parser: The process of assigning the parts of speech (POS) to a word is called Parts Of Speech (POS) tagging. It includes nouns, verbs, adverbs, adjectives, pronouns, conjunction and their sub-categories [31].

For Example:

Word: Paper, ***Tag:*** Noun

Word: Go, ***Tag:*** Verb

Word: Famous, ***Tag:*** Adjective

Note that some words can have more than one tag. For example, chair can be noun or verb depending on the context. Taggers use information like: dictionaries, lexicons, rules, and other. Dictionaries have categories of words. That is a word may belong to one or more category. Taggers use probabilistic information [31]. There are two types of taggers: rule-based and stochastic. Rule-based taggers use hand-written rules. Stochastic taggers are HMM based, choosing the tag sequence which maximizes the product of word likelihood and sequence probability. Ideally a tagger should be robust, efficient, accurate, tunable and reusable. In reality taggers either identify the tag or make the guess.

Information Gain (IG): IG scores show the contribution of the presence or absence of a term to correct classification of text documents. IG assigns maximum value to a term if it is good for assigning the document to any class. IG is a global feature selection metric producing only one score for a term t and calculated as [32]:

$$IG(t) = -\sum_{i=1}^{M} P(c_i)\log P(c_i) + P(t)\sum_{i=1}^{M} P(c_i|t)\log P(c_i|t) + P(\bar{t})\sum_{i=1}^{M} P(c_i|\bar{t})\log P(c_i|\bar{t})$$

where, M is the number of classes, $P(C_i)$ is the probability of class C_i, $P(t)$ and $P(\bar{t})$ are the probabilities of presence and absence of term t, $P(c_i|t)$ and $P(c_i|\bar{t})$ are the conditional probabilities

Combination of NLP parser and TF*IDF: In this method we combine the POS tagged features and the TF-IDF based computed features for preparing a new set of features.

Combination of IG and NLP parser: As similar to the last feature set here we combine the features of POS tagger and information gain.

4.4 CLASSIFIER TRAINING

In this phase we have used the extracted features and their combinations to train the ML classifiers. Additionally further use them to test the performance and measure influence of feature extraction techniques in classifier's performance. In this experiment we have used the following classifiers:

Support vector machine (SVM): SVM is one of the most popular ML techniques for data classification. It can be used for classify the linear as well as nonlinear data. The goal is to separate the two classes using a function prepared using training data. The SVM is maximizing the margin during classification. It is used to solve the binary classification problems. The classifier is finding the hyper-plane with the largest margin; training data are not always linearly separable [33]. Thus to handle the nonlinearly some slack variables have been used to tolerate training errors. This variable is referred to as a soft-margin. The SVM classifier creates one or multiple hyper planes for classification and regression.

Bayes Classifier: The Naive Bayes classification is a probabilistic classifier. This can derive by using Bayes' theorem. Based on the nature, we train the Naive Bayes algorithm as a supervised learning. There are two types of probabilities are used [34]:

- Posterior Probability: [$P(H/X)$]
- Prior Probability: [$P(H)$]

where, X is data and H is assumption. Thus, Bayes Theorem stated as:

$$P\left(\frac{H}{X}\right) = \frac{P\left(\frac{X}{H}\right)P(H)}{P(X)}$$

Artificial Neural Network (ANN): The neural network is defined in two phases' training and prediction: training method utilizes data and designs the model. By this data model prediction is performed [35]. We need here two arrays; one is input and hidden unit and the second is output layer. First array is a two-dimensional array W_{ij} and output is a one dimensional array Y_i.

The initial weights are random assigned, and then output is calculated as:

$$x_j = \sum_{i=0} y_i W_{ij}$$

where, y_i is the input or output of previous layer and W_{ij} is the weight of the connection.

Next, we can use different activation functions for producing the outcome for next layers such as sigmoidal function. Described below:

$$y_i = \left[\frac{e^x - e^{-x}}{e^x + e^{-x}} \right]$$

When event of all output units is determined, the network calculates the error (E) given in equation.

$$E = \frac{1}{2} \sum_{i=0} \left(y_i - d_i \right)^2$$

where, y_i is output predicted by model and d_i is the actual output.

4.5 CLASSIFICATION AND PERFORMANCE

The implemented classifiers are trained with the extracted features from five different techniques. The training has been done with 70% of extracted features and 30% features are used for testing. The implemented ML algorithms are classifying the test features into two class labels i.e. legitimate and spam. Based on the classification outcomes the performance of the models is compared in next section.

5. RESULT ANALYSIS

5.1 EXPERIMENTAL SCENARIOS

The proposed work is aimed to investigate the performance influence of text feature selection techniques over the classification approaches for social media spam detection. Thus, we perform two sets of experiments in these scenarios.

Table.5. Experimental Analysis of Feature selection Approach

Dataset	Accuracy (%)					F-Score (%)				
	TF-IDF	IG	POS	POS+ TF-IDF	POS + IG	TF-IDF	IG	POS	POS+ TF-IDF	POS + IG
1000	86.3	88.9	87.8	90.5	89.7	78.4	79.1	78.9	80.1	80.8
2000	84.7	87.5	89.4	91.7	92.7	78.8	80.4	79.2	82.3	81.6
3000	87.5	89.7	90.2	93.1	94.5	80.1	81.6	81.4	83.6	82.7
5000	88.9	92.4	92.7	94.4	95.8	81.6	82.7	80.1	84.3	85.3
8000	86.7	91.6	93.1	96.5	96.7	82.5	83.2	84.3	85.9	85.6
10000	89.2	91.4	94.5	96.9	97.1	84.1	84.9	85.7	86.7	87.3
11968	88.4	93.7	95.9	97.2	98.3	86.7	87.5	88.2	88.4	90.1

Table.6. Performance of classifiers

Dataset	Accuracy (%)						F-Score (%)					
	Bayes		ANN		SVM		Bayes		ANN		SVM	
	POS+ TF-IDF	POS + IG	POS+ TF-IDF	POS + IG	POS+ TF-IDF	POS + IG	POS+ TF-IDF	POS + IG	POS+ TF-IDF	POS + IG	POS+ TF-IDF	POS + IG
1000	80.6	82.8	91.5	93.6	90.5	89.7	73.2	74.3	84.3	86.7	80.1	80.8
2000	82.4	83.4	92.6	93.9	91.7	92.7	74.8	75.1	85.1	88.4	82.3	81.6
3000	85.2	85.7	93.4	94.8	93.1	94.5	76.4	77.3	86.9	89.1	83.6	82.7
5000	86.8	87.6	94.6	96.3	94.4	95.8	77.9	78.7	88.6	90.5	84.3	85.3
8000	88.4	89.3	96.2	97.5	96.5	96.7	78.4	79.5	90.3	91.4	85.9	85.6
10000	89.7	90.7	97.6	98.2	96.9	97.1	79.5	80.3	91.5	92.6	86.7	87.3
11968	90.4	91.6	98.6	99.2	97.2	98.3	82.8	82.1	92.4	94.3	88.4	90.1

Table.7. Mean performance of feature selection techniques and classifiers

	Feature selection Techniques					Classifier performance					
						Bayes		ANN		SVM	
	TF-IDF	IG	POS	POS+ TF-IDF	POS + IG	POS+ TF-IDF	POS + IG	POS+ TF-IDF	POS + IG	POS+ TF-IDF	POS + IG
Accuracy (%)	87.38	90.74	91.94	94.32	94.97	86.21	87.3	94.92	96.21	94.32	94.97
F-score (%)	81.74	82.77	82.54	84.47	84.77	77.57	78.18	88.44	90.42	84.47	84.77

- **Influence of feature selection over classification:** in this experimental we fixed the classification algorithm to SVM as binary classifier, and the implemented five different features are used. The aim is to know how the individual feature and their combinations are influencing the performance of SVM classifier.
- **Measuring effect of Classification technique used:** in this phase we have identified the top two feature selection approaches and use them with the implemented classification techniques. The aim is to know how the selection of classification algorithm will improve the performance of social media spam detection.

This section describes the experimental scenarios; next section describes the conducted experiments and their consequences in detail.

5.2 INFLUENCE OF FEATURE SELECTION OVER CLASSIFICATION

In order to measure the impact of both the experiment scenarios, we have involve two performance parameters namely accuracy (%) and F-score (%).

The accuracy is describing how accurately an algorithm will learn on the given training samples and then recognize them. The accuracy of an algorithm can be defined using the following formula:

$$accuracy=(total\ correctly\ recognized)/(total\ samples\ to\ recognize)\times 100$$

The next parameter is F-score. The F-Score is measured using the two factors namely precision and recall. That is the harmonic mean of both the parameters. The following formula can be used for measuring f-score.

$$F\text{-}score = 2*(precision*recall)/(precision+recall)\times 100$$

In this experiment the dataset has a total of 11968 instances of data. The entire dataset has subdivided in 8 different sets for performing the experiments. The used size of data samples are demonstrated in Table.5. The aim of the preparation of these sets is to measure the performance with increasing amount of data. Additionally we have made use a common ML classifier for training and testing. Thus we have used here the SVM classifier for measuring the influence of feature extraction techniques. The performance of the classifier with the different extracted features is demonstrated in Table.5 in terms of accuracy and F-score. The obtained accuracy of the SVM with different feature extraction techniques is given in Table.5.

The Table.5 demonstrates the classification accuracy of three individual features and their two combinations. Similarly the F-Score of the models has also been measured and reported in Table.5. In both the diagrams X axis shows dataset size used in experiment and Y axis shows the obtained performance in terms of accuracy and f1-score respectively. According to the experimental results we made the following conclusion about the results for feature selection techniques:

- In order to classify the social media spam the sentiment features POS tag are providing higher accuracy as compared to simple text classification features i.e. TF-IDF and IG.
- The combination of features is improving the accuracy as compared to individual use of features section techniques
- The feature combination POS tag and IG is more effective than the combination of POS and TF-IDF based feature representation
- The performance of classifiers is also influencing with the size of training samples used for providing training to the classification algorithms.

According to the experimental results we found the hybrid approaches are performing more accurately as compared to individual features. Thus, in next experiment for selection we utilize the hybrid text features i.e. the POS and IG, and POS and TF-IDF.

5.3 MEASURING EFFECT OF CLASSIFICATION TECHNIQUE USED

The aim of this phase of experiment is to know how the classification techniques used can affect the performance of the social media spam detection. In this experimental investigation we have used the similar performance parameters i.e. accuracy and f-score. Additionally, to train and test the classification algorithms the similar size of dataset has been used. Additionally unlike the previous experiment here we have used three popular classification algorithms namely SVM, ANN and Naïve Bayes. The performance of classifiers in terms of accuracy (%) and the f-score measured has given in Table.6.

The X axis of this diagram is including the size of data samples used in experiment and Y axis shows the accuracy in percentage and f-score respectively. Additionally the observed experimental consequences have given in Table.6 for all the three classifiers and their feature combinations. According to the obtained performance in terms of accuracy and f1-score we found similar trends. However, we can see the classification accuracy of the classifiers with the feature selection technique POS and IG demonstrate the higher values as compared to POS and TF-IDF based feature selection in all the implemented scenarios.

In addition, the performance of Naïve Bayes classifier has increasing with the size of dataset used in increasing manner. The second highest performance we have found with the SVM classifier. Additionally, the ANN based classifier demonstrates the highest performance as compared to other two implemented algorithms. In order to compare and get more clearly different in performance variation we also measured the mean performance in both the scenarios.

6. MEAN PERFORMANCE SUMMARY

The mean performance of both the experimental scenarios are measured and reported in Table.7. The Table.7 demonstrates the performance for comparing the influence of feature selection techniques with the classifier in terms of mean accuracy and f-score. According to the obtained results the hybrid feature selection techniques are performing better than the individual feature selection techniques. Additionally, we found that the hybrid feature selection techniques i.e. POS tagger and information gain based, and POS tagger and TF-IDF based techniques are performing much accurately. Thus, extended the experiments with these two hybrid features and three classification techniques in terms of mean performance is demonstrated in Table.5. In this experiment we found the SVM

and ANN has higher accuracy as compared to the Bayes classifier. additionally, ANN and the combination of features POS and IG provides more superior performance then other implemented combinations of classifiers and feature selection techniques. Thus, in near future for extension the ANN classifiers and POS and IG based features has been used.

7. CONCLUSIONS

The aim of this paper is to investigate the social media spamming and the recent work around the detection of social media spam. Thus first we have done a review to investigate about the social media spamming, the techniques and tools available, datasets and the requirements of feature selection methods. Next we have developed a generic ML model for finding influence of different feature selection techniques over their detection accuracy. Additionally we have also make effort for finding the impact of classification algorithm used for social media spam detection. The experiments on the publically available dataset we have measured the performance of designed model additionally we found the following facts as the conclusion.

- The feature selection techniques have significant influence on the spam detection performance; the better feature representation can increase the detection accuracy up to 5-8%.
- The size of training sample and quality of training samples may help in better learning of classifier. The increasing size of training sample may improve the accuracy up to 3-5%.
- The use of suitable classifier for feature learning may improve the classification performance up to 8-10%.
- The individual features are less effective as compared to hybrid and combinations of the feature selection techniques.

Based on the conducted experiments and obtained consequences we have planned the following future extensions:

- The spam messages with URL are more harmful thus need to involve the URL classification techniques to detect harmful spam messages also

The large data analysis with the classical machine learning techniques are computationally expensive thus in near future we need to apply the deep learning models.

REFERENCES

[1] G. Appel, L. Grewal, R. Hadi and A.T. Stephen, "The Future of Social Media in Marketing", *Journal of the Academy of Marketing Science*, Vol. 48, pp. 79-95, 2020.

[2] S.R. Srivastava, S. Dube, G. Shrivastava and K. Sharma, "Smartphone Triggered Security Challenges - Issues, Case Studies and Prevention", *Cybersecurity in Parallel and Distributed Computing*, Vol. 78, pp. 1-14, 2018.

[3] A. Sharma and M. Ramaiya, ""SPAM" In Social Media: A Review", *Wesleyan Journal of Research*, Vol. 14, No 1, pp. 1-12, 2018.

[4] S Umajancy and A.S. Thanamani, "An Analysis on Text Mining Text Retrieval and Text Extraction", *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 2, No. 3, pp. 1-14, 2013.

[5] M.W. Mosing, "The Ups and Downs in the History of EU-Spam-Regulations and Their Practical Impact", Available at https://www.it-law.at/wp-content/uploads/2014/09/spamsymposium -eu-mosing.pdf, Accessed at 2020.

[6] B. Satterfield, "Ten Spam-Filtering Methods Explained", Available at https://www.techsoupcanada.ca/en/learning_center/10_sfm_explained, Accessed at 2021.

[7] A. Copestake, "Natural Language Processing", Available at https://www.cl.cam.ac.uk/teaching/2002/NatLangProc/revised.pdf, Accessed at 2004.

[8] R. Collobert and J. Weston, "Natural Language Processing (Almost) from Scratch", *Journal of Machine Learning Research*, Vol. 12, pp. 2493-2537, 2011.

[9] A. Ishtiaq, M.A. Islam, M.A. Iqbal, M. Aleem and U. Ahmed, "Graph Centrality Based Spam SMS Detection", *Proceedings of International Bhurban Conference on Applied Sciences and Technology*, pp. 1-8, 2019.

[10] A. Pinandito, R.S. Perdana, M.C. Saputra and H.M. Az-zahra, "Spam Detection Framework for Android Twitter Application Using Naive Bayes and K-Nearest Neighbor Classifiers", *Proceedings of International Conference on Software and Computer Applications*, pp. 77-82, 2017.

[11] G. Jain, M. Sharma and B. Agarwal, "Spam Detection in Social Media using Convolutional and Long Short Term Memory Neural Network", *Proceedings of International Conference on Annals of Mathematics and Artificial Intelligence*, pp. 1-9, 2019.

[12] H. Raj, Y. Weihong, S.K. Banbhrani and S.P. Dino, "LSTM Based Short Message Service (SMS) Modeling for Spam Classification", *Proceedings of International Conference on Computing Machinery*, pp. 19-21, 2018.

[13] I.I. Dutse, M. Liptrott and I. Korkontzelos, "Detection of Spam-Posting Accounts on Twitter", *Neurocomputing*, Vol. 315, pp. 496-511, 2018.

[14] K. Zainal and M.Z. Jali, "A Review of Feature Extraction Optimization in SMS Spam Messages Classification", *Proceedings of International Conference on Software Engineering*, pp.158-170, 2016.

[15] K. Ho, V. Liesaputra, S. Yongchareon and M. Mohaghegh, "Evaluating Social Spammer Detection Systems", *Proceedings of International Conference on Computing Machinery*, pp 1-6, 2018.

[16] L.A. Cornelissen, P. Schoonwinkel and R.J Barnett, "A Socio-Informatic Approach to Automated Account Classification on Social Media", *Proceedings of International Conference on Computing Machinery*, pp. 19-21, 2019.

[17] M. Itani, C. Roast and S.A. Khayatt, "Developing Resources for Sentiment Analysis of Informal Arabic Text in Social Media", *Procedia Computer Science*, Vol. 117, pp. 129-136, 2017.

[18] M.A. Janabi, E.D. Quincey and P. Andras, "Using Supervised Machine Learning Algorithms to Detect Suspicious URLs in Online Social Networks", *Proceedings of International Conference on Advances in Social Networks Analysis and Mining*, pp. 1-13, 2017.

[19] N.E. Mawass, P. Honeine and L. Vercouter, "Supervised Classification of Social Spammers using a Similarity-based Markov Random Field Approach", *Proceedings of*

International Conference on Computing Machinery, pp. 15-19, 2018.

[20] P. Tehlan, R. Madaan and K.K. Bhatia, "A Spam Detection Mechanism in Social Media using Soft Computing", *Proceedings of the 13th INDIACom*, pp. 1-13, 2018.

[21] Q. Zhang, C. Liu, S. Zhong and K. Lei, "Spam Comments Detection with Self-Extensible Dictionary and Text-Based Features", *Proceedings of International Conference on Computers and Communications*, pp. 1-13, 2017.

[22] S. Dhawan and Simran, "An Enhanced Mechanism of Spam and Category Detection using Neuro-SVM", *Procedia Computer Science*, Vol. 132, pp. 429-436, 2018.

[23] S. Shehnepoor, M. Salehi, R. Farahbakhsh and N. Crespi, "NetSpam: a Network-based Spam Detection Framework for Reviews in Online Social Media", *Proceedings of International Conference on Computers and Communications*, pp. 1-15, 2017.

[24] S. Tingxuan and R.Y.K. Lau, "Collective Classification for Social Opinion Spam Detection", *Proceedings of International Conference on Computers and Communications*, pp. 19-21, 2019

[25] S. Chancellor and Y. Kalantidis, "Multimodal Classification of Moderated Online Pro-Eating Disorder Content", *Proceedings of International Conference on Computers and Technology*, pp. 6-11, 2017.

[26] T. Green and F. Spezzano, "Spam Users Identification in Wikipedia via Editing Behavior", *Proceedings of 11th International AAAI Conference on Web and Social Media*, pp. 1-13, 2017.

[27] T. Wu, S. Liu, J. Zhang and Y. Xiang, "Twitter Spam Detection based on Deep Learning", *Proceedings of International Conference on Computers and Communications*, pp. 111-123, 2017.

[28] Y.A. Amrani, M. Lazaar and K.E. Elkadiri, "Sentiment Analysis using Supervised Classification Algorithms", *Proceedings of International Conference on Computing Machinery*, pp. 321-335, 2017.

[29] Twitter Spam, Available at https://www.kaggle.com/c/twitter-spam/overview, Accessed at 2021.

[30] P. Bafna, D. Pramod, A. Vaidya, "Document Clustering: TF-IDF Approach", Proceedings of International Conference on Electrical, Electronics, and Optimization Techniques, pp. 1-14, 2016.

[31] S.M. Mohammad, S. Kiritchenko and X. Zhu, "NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets", *Proceedings of International Conference on Semantic Evaluation Exercises*, pp. 321-327, 2013.

[32] T. Kenter and M. De Rijke, "Short Text Similarity with Word Embeddings", *Proceedings of International Conference on Electrical and Electronics*, pp. 19-23, 2015.

[33] S. Liu and H. Shen, "Adaptive Cotraining SVM for Sentiment Classification on Tweets", *Proceedings of International Conference on Information and Knowledge Management*, pp. 2079-2088, 2013.

[34] C. Wan and A.A. Freitas, "An Empirical Evaluation of Hierarchical Feature Selection Methods for Classification in Bioinformatics Datasets with Gene Ontology-Based Features", *Artificial Intelligence Review*, Vol. 78, pp. 1-13, 2017.

[35] A. Ghosh, "Comparative Study of Financial Time Series Prediction by Artificial Neural Network with Gradient Descent Learning", *International Journal Of Scientific and Engineering Research*, Vol. 3, No. 1, pp. 1-14, 2012.