Spam, a Digital Pollution and Ways to Eradicate It

Chinthapanti Bharath Sai Reddy, Shaurya Chaudhary, Saravana Kumar Kandasamy

Abstract- Due to the growing popularity of the microblogging and networking sites like twitter, Gmail, Facebook etc., there has been an increase in the number of spammers. Spammers on Twitter seem to be more dangerous than the mail spammers as they exploit the limitation on the characters of Twitter for their own purposes. Spammers have also become creative in framing their content to cleverly escape the classifiers. This survey is thus mainly used to discuss and analyze the recent research that had been put forth regarding the spam detection in social media sites such as Twitter. This survey analyses the papers that tackled various problems faced on Twitter and the problems faced by the methods that have already been presented before. We then compared all the methods present in the papers to see which method or combination of methods could give the best result in detecting spam.

Index Terms- Bayes methods, Classification algorithms, Clustering algorithms, Feature extraction and Machine learning algorithms.

I. INTRODUCTION

In this technology connected world, every human is connected to the internet all the time, there are a lot of ways people communicate with each other over the internet. Starting from Instant messaging, email, forums, tweets and websites and lot more. A lot of data is obtained by mining the data from social media, this data is being utilized for spamming and targeting people [20]. But these are also misused by some unethical people for delivering disturbing content, advertisements with the help of target ads. Spam takes place in all the platforms. Humans tend to get affected by any news very easily, if any object comes for a low price than the normal price humans tend to show interest in buying them. So, spamming mostly consists of offers for a huge discount and many other. Spamming has become one of the major inconveniences faced by every internet user. General meaning of spam is sending or submitting the same message to a large number of people in an attempt to force the message on to the people who would otherwise choose not to receive this message. Despite how many new spam filters and spam detection algorithms are being used the spammers find a way to pass through them. Few researches say that around fourteen billion spam messages are globally sent per day. Approximately there are 45% of all emails which are spam sent in a day [16]. USA stands in number one for spam generation. Spam in twitter is spreading of fake news with rigorous spamming. Also, some companies try to spam, it works as a marketing strategy as well with the help of links [17]. Email spam, also known as junk, is basically mail that is sent a huge number of times. The presence of spam has been increasing rapidly from the 1990s and is a major issue faced by most of the email users. The people who received spam often have had their email addresses obtained by spambots, which are automated programs that crawl the internet looking for email addresses. Attackers use spambots to create email distribution lists. Spammers are learning from old methods and updating their techniques for better targeting [18]. An attacker will send an email to millions of users with the estimation that only a few clicks or interact with the message. Also, nowadays a lot of defense mechanisms such as review spam detection frameworks and software at the server side for ensuring the genuity [22].

Platforms, such as Facebook and Twitter, are more powerful in making the internet connectivity. Approximately 1/3 of the world-wide population are estimated to be now connected and within 3 year one-half of the global population will be connected. As we know a lot of research work is being done with the help of twitter data, so because of the spam a lot of research is being done on the fake data resulting in the unexpected and false results which will stale the research progress [19] . One study suggested that over 15% of the active twitter users are automated spam bots. These Spammers make the fake news as it is genuine and make neutral people turn towards one side of any argument resulting in the manipulation of users free will of choice. But there have been many spam detecting methods on twitter in recent years. Some methods use hashtags as a way to detect spammers. The spammers use the trending and popular tweets for their means. After conducting their research for 2 months on 14 million tweets, they created a dataset called HSpam14 which could be used for hashtag-oriented spam research [23]. Some methods also use the spammer's behavior to detect spam. There is a study conducted for seven months and found 36, 000 spammers on twitter. They analyzed the behavior of these spammers by checking the follower-followee relationship, link payloads etc. The behavior of spam accounts is way too different than the normal users [24]. Some methods also use statistical analysis to determine if they are spam or not. The analysis sees only the content and not the user's details in detecting spam. The analysis was taken on the dataset created through machine learning techniques on the popular hashtags [25].

Revised Manuscript Received on December 15, 2019.

Chinthapanti Bharath Sai Reddy, Department of CSE, Vellore Institute of Technology, Vellore (Tamil Nadu) India.
E-mail: chi.bharathsai@gmail.com

Shaurya Chaudhary, Department of CSE, Vellore Institute of Technology, Vellore (Tamil Nadu) India.
E-mail: shauryachoudhary.2009@gmail.com

Saravanakumar Kandasamy, Department of Computer Science, Vellore Institute of Technology, Vellore (Tamil Nadu) India.
E-mail: ksaravanakumar@vit.ac.in

Some papers have been proposed that also see an importance to the URLs used by the spammers. As spammers mostly use URLs to redirect the users and get their personal details. It is more dangerous than the spam mail. There is a proposed method called Warning Bird for the URL detection in twitter which focuses on the URL redirects [26]. But as time evolves, spammers are finding clever ways to escape these detecting methods. Therefore, there is a focus on finding new and improved methods for spam detection.

II. DEFINITIONS

Spam: Spam refers to the unwanted messages which are usually sent in bulk to a large group of people through electronic medium or on social media.
Spammers: The group of people or automated bots who spread these spam *contents* all over the internet in a regular fashion, are known as spammers.
Spam Detection: The method of classifying an input message as spam or not spam through a model is known as spam detection. It is *implemented* is tons of different methods involving approaches of machine learning, neuro computing, etc.
Information Quality (IQ): It is the quality or the fitness of the information provided by a certain source. It determines the feasibility of whether the data can be used for research purposes or not. Increase in spam content drastically deteriorates the IQ.
Dataset: Dataset is a collection of huge data of similar type, specially structured for specific purposes. The datasets are used as input to train models for spam detection.
Features: A feature is a characteristic of the observed domain [15]. In the case of spam detection, we need to choose features correctly in order to ensure they are effective and different from the other features.

- **UPF** - User Profile Features (Username, screen name, location, bio)
- **AIF** - Account Info Feat. (account age, verification)
- **EwF** - Engage with features - users can influence these features - (friend count, status count, type of tweet, time taken to create the tweet, the no. of tweets sent per a particular amount of time)
- **EbF** - Engage by feat. - can't be influenced by users directly - (Retweets count, followers count, favorites count)

Clusters: Cluster is a group of similar data points in the proximity. The data points in the same group have the same properties and features but vary extensively from that of other clusters.
Ontology: It shows the set of concepts in the focused domain along with the *relationship* between them.

III. PROBLEM DESCRIPTION

In this technology driven world we need a device in our possession all the time, social media has become an inseparable part of our lives. According to a research conducted in 2017, an average person spends 2 and a half hours every day on social media websites. With the years passed, the time is only increasing. Social media sites like Facebook and Twitter are really bringing the whole world together by enabling more than 3 billion people to connect with each other. Social media users generate and use information which leads to huge amounts of data. These data are in the form of personal as well as social information. Every second, tons of data is flowing through the internet.
We are now living in a world where the price of data has surpassed the cost of oil in the international market. Social media has now become one of the costliest things in the world. But this data is being compromised by Spams. Spammers are flooding social media sites with spam emails and tweets at an extreme rate. It has been found that on an average of 200 messages, there's one spam message while one spam tweet in every 21 tweets. The data extracted from social media sites is widely used in numerous researches, but this rapid increase in spam content is making the data biased which raises a suspicion for researchers whether they are using legitimate or accurate information.

Online spamming comes in various forms such as malware dissemination, abusive content, fake news, and generating fake product reviews. This makes it difficult to check the legitimacy of the content being posted. So, use of social media data for research will give us false results or unreliable results. To take care of this issue, many researches have been done and numerous are still going on. Hundreds of models have been proposed to detect spam content over social media platforms but still we are far from eradicating this issue. As the research evolves in the domain of Spam Filtering, spammers also get smarter over time.

Spammers have been using several techniques to target social media in order to gain maximum profits. They usually take trending topics for their benefit to target large set of users in a short time. With time, spammers keep on altering their strategies by changing the characteristics of the spam messages/tweets to fool the spam filters. This change is techniques by the spammers is known as *spam drift.*

Spamming on twitter is heavily based on the trending hashtags. They also use mentions in the tweets to target a particular set of users. Twitter also provides APIs to developers to integrate into third-party apps or websites. But these APIs are exploited by spammers in order to automate the spamming process. There are a countless number of spam bots present at the moment which keeps on spreading spam contents in a regular manner.

Most of the spam detecting methods use behavioral and statistical methods to detect spam, but they all have their limitations. For instance, a spam account detection model looks for several attributes like follower-followee ratio, tweet frequency, interaction with other users, age of the account, etc. before labelling the account. One of the major issues faced in classifier models is that the dataset used to train them gets old, so the model fails to detect new spam strategies enforced by spam drift.

The complete information about the user account can make the whole spam detection process much easier, but websites like Twitter and Facebook do not allow access to these restricted data due to privacy concerns.

Nonetheless, with advanced spam filtering models to perform efficiently on the dataset – relevant feature selection has become a necessity. It's a challenging task to eliminate the irrelevant features which occur least frequently in the training dataset.

These features not only contribute in negative results but also increase the processing time and cost.
We will be further discussing about some proposed methods in this domain of Social Media Spam Detection to overcome the above challenges.

Also, spam will cause problems such as fake data to propagate and will be able to manipulate people's mind with the help of fake data. Which will result in large chaos will may also lead to crisis and many other major threats.

IV. SYSTEM ARCHITECTURE

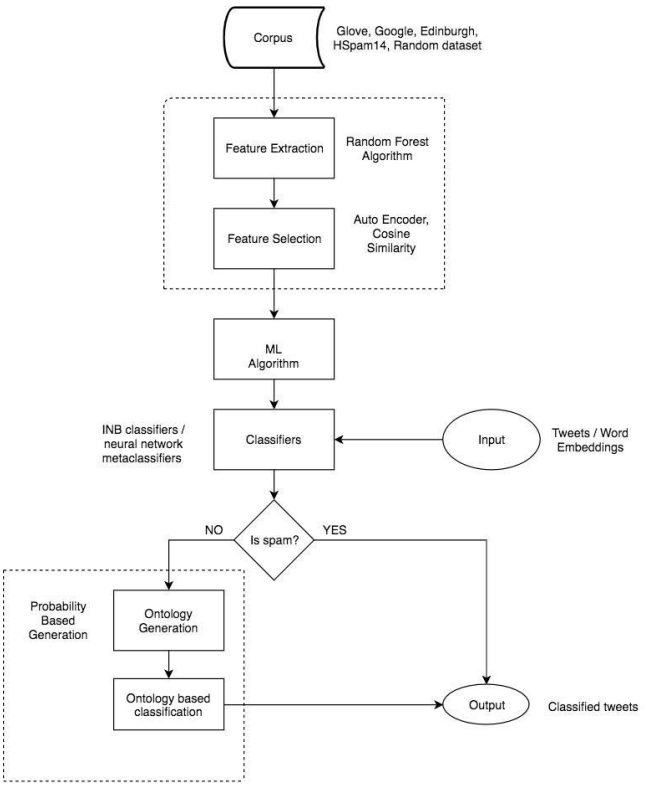

Figure 1: A Generic system architecture for Spam filtering based on the proposed papers

V. PROPOSED METHODS FOR SPAM

A. Ontology based approach

There have been many ontology-based approaches in detecting spam in the recent years. Ontologies talk about the specific topics present in a determined field and about how they relate to each other. Many spam detection methods have been introduced to detect spam. Like an ontology-based approach in detecting the spam mails. There has been research towards using two types of ontology-based spam filters. They are called global ontology filter and customized ontology filter [9].

1. Spam-tweet detection

Twitter has become unreliable for the researchers as there has been an increase of spammers. There are many spam detection methods today but they all have limitations. They mostly use behavioral and statistical methods to detect spam. But this would mean we need to have the details like public information and the relationship information such as follower-followee ratio. Spammer's are also learning about the different features to escape from the spam detection. There is also a restriction in using user's information without their consent. There is a restricted access to the Twitter APIs and the metadata that the process becomes so expensive. This led to the need for a novel ontology approach that overcomes these limitations.

The proposed method, that paper talks about, only focuses on the content of the tweets and not on the user's details. They proposed a method where the tweet messages are checked with the ontologies to classify them into spam or non-spam tweets. They followed seven unidirectional ways to create the ontology. They used the data driven discovery algorithm as it is closely related to the context of the tweets. If any changes are made to these contexts, the change could be seen all through the ontology making it really flexible and accurate. The dataset used was raw and unstructured data from the date 05-2013 to 08-2013. They found these tweets from an online archive. They prepared the data to be tested from this raw data by themselves. They clustered the cleaned data into groups based on the hashtags. It is an ideal dataset to test for handling big data. They also contain data of different types of topics with different hashtags and different time zones. It is ideal for this sort of ontology study. They fed this data to create the ontologies and make groups. They used only three main themes: sports, technology and politics. They have then conducted experiments on these different groups which have different values for token similarity threshold. They compared the ontologies with a random set of data which contain varied values of token similarity threshold. They have also found the accuracy and efficiency of the spam detection by comparing them with a random set of data which contain varied values of token similarity threshold. This method gave an idea about the false positives and false negatives in the spam detection.

They proved that their probabilistic ontology generation method outperformed message to message models such as NLTK model, Cosine vector similarity and Co-occurrence model. There is also a lot of reduction in false positives and false negatives. But it was seen that different types of ontologies produce different results of accuracy. The lower the token similarity threshold, the larger the number of false positives is. But all in all, the few are good as many approaches will work accurately [3].

2. Spam-account detection

A report by a research article showed that on an in every 200 social media posts, there will definitely be one spam post and that about 15% of users using twitter are automated bots used for spamming. The immense amount of such spam content and the use of faulty data from the bots will lead to negative or false results for the researchers.

This research showed that normal spammers and social media bots have similar behavior. There are many factors that determine whether the account is spam or not. The conventional methods used keyword-based identification to check for spam. But a real-time spam detection method is used in this paper that performs far more efficient than existing methods and models. The factors that mainly affect are follower-followee relationship, frequency of tweets, user active time and interaction with the other accounts [6].
This method suggests a better, efficient and optimized features which are disjoint and independent of the tweets previously tweeted,

which will be available for a short period of time on the social media platform, Twitter. Here in this method features related to account and user engagement with their respective followers and frequency of the tweets and display picture and banner of the account also gives some information, Also the regularity in which password change takes place will be a major factor. A research gave out that on an average 12 tweets will be tweeted by a spam account daily with in a particular time, frequency and location.

These features when employed the important features are grouped as related to twitter account. Feature elimination has been deployed to verify the robustness each feature. When compared with the earlier study this method came out as more efficient and more reliable for spam account detection using ontology

B. Feature Based detection

1. Context based approach

The problem answered in this paper is addressing about the data transformation that is happening, before the usage of machine learning classifiers. Feature portrayal that keeps class differentiability with lower level space for identifying spam is being executed or proposed. More number of features will generate negative performance on the learning classifier also Computational time for data processing during the training process will be drastically increased due to presence of more number of features in the data. Pre-processing is one of the major steps nowadays in any training processes. Data pre-processing stage also sums up the speed of computation and also plays a key role in improving the classification accuracy. So, without pre-processing the negative impacts will cause more wrong results to the project. The significant bit of leeway in regards to the proposed highlight portrayal is its strength that empowers classifiers like Random Forest, Support Vector Machines, and the decision tree C4.5 to distinguish an approaching email as spam or non-spam where the element size is exceptionally little with a decent speculation independent of the information source.

As Spammers are careful with the words they use in the mail and the way of writing mails they avoid more commonly used words for spam this will cause classifiers to work with poor performance. The paper proposes utilization of thick feature representation which catches the sentence structure and semantic meaning inside a record joined with cosine similarity and Autoencoder for feature learning will prompt a decent order with better consistency when contrasted with the condition of-craftsmanship feature representation approaches in spam filtering task [4].

2 Semantic based approach

Nowadays, words are being chosen by feature selection methods, these are being used to generate feature vectors for training different approaches. Also, this research teaches a new method of selection of features which is going to take the advantage of semantic ontology to categorize words into topics and utilize them to build vectors.

(i) Information Gain, it is the most popular feature selection method in the domain of spam-filtering and classifying the spam accounts from a set of features.

(ii) Latent Dirichlet Allocation, it is a probabilistic model which allows data set of observations to explain unobserved groups which tells us why few parts of data are similar.

(iii) Semantic-based feature selection, this paper proposal results have shown the efficiency and reliability and more advantages of topic-driven methods to develop an efficient model and deploy high-performing spam filters.

This work is engaged in the portrayal of email messages utilizing subjects as highlights to filter spam with ML algorithms. In spite of the fact that they are removed from words, point highlights represent the topic of the rather than crude terms. One of the significant downsides of maintaining a strategic distance from FS lies in the repetition of highlights [5].

C. Neural networks-based approach

1. 5 CNN+ 1 Feature based model

There has been spam for years at tweet level in twitter. They are really dangerous than the spam mail. Although, the twitter users can report spammers and spamming accounts, the spammers can continue their activity by creating new accounts. This where we need a tweet level spam detection.

They proposed a solution by creating a neural network algorithm, a feature-based model and an ensemble. They used words as the feature representation of tweets. The CNNs use word embeddings in their proposed method. The feature-based model used user-based, content-based, and n-gram features. This paper proposed a combination of five CNNs model and one feature-based model via neural networks regarding this problem. Here, neural networks work as a meta-classifier. They have used two datasets to check their proposed method. One of them is a subset of HSpam14 data set. They took 1 million tweets in the starting of this dataset. This subset is then split into a ratio of 2:1 and then classified either as spam or non-spam. The second dataset 1KS10KN and it was not a balanced dataset. They have used the dense representation of the word vectors because of low computational speed and generalization power. It is also preferred as there is a correlation between the features. There are five layers present in the CNN architecture proposed in this paper. They are input layer, convolution layer, pooling layer, hidden layer, and an output layer. The training of the word embeddings was done using the skip-gram method. The ensemble method combined the five CNNs and one feature-based model. They took word embeddings like Twitter Glove, Google news corpus word2vec, Edinburgh Twitter corpus word2vec, HSpam14 Twitter corpus and random embeddings which had different dimensions and compared it with every CNN. The random forest algorithm is then applied to the features in this ensemble method to detect spam.

The proposed method seemed to take more execution time than the other methods. It seemed to show better performance for a smaller and unbalanced dataset like 1KS10KN rather than a big and balanced dataset like HSpam14. But all in all, it shows better performance than all the base methods and has great robustness so that no spammer can escape this detection. This proposed method could be more efficient if there is a better feature representation. The input taken for the machine learning techniques used by them was raw tweets. The performance of the deep learning techniques could be better with additional information [2].

2. Auto-GA-RWN

Email communication is now being used more than ever. It's prevalent and indispensable nowadays. Every second, tons of data are flowing through it. This huge database of information makes it vulnerable as Cyber criminals get lured into it. The threat of spamming is getting more serious. A survey revealed that 40% of emails were spam in 2006 and recently it has reached as high as 70%. The spam drift is making the problem even severe. Spammers are using different features for their spam messages and are evolving over time. Spamming is usually done with similar content and in large quantities. This makes filtering comparatively easy for the most part. The spam messages squander the significant assets, including capacity, transmission capacity, and profitability.

The spammers tend to change their techniques over time, which breaks the pattern and make the email unpredictable. This produces a requirement for a model that could naturally distinguish the features and enhance the detection process.

A 2-stage hybrid model based on combination of Random Weight Network and Genetic algorithm is proposed for Email spam detection. The 2 stages are: Feature selection, and email classification. The model is named as Auto-GA-RWN.

Generic Algorithms are a class of optimization techniques, widely used for feature selection. GAs is intensively used in various fields because of their simplicity of usage and effectiveness.

The transformative stages in GAs start with an irregular populace of candidate solutions (singular genomes, search agents or phenotypes). Every agent has a pool of chromosomes or genotypes which must be transformed and advanced during the exploration and misuse exploitation periods of this algorithm [13]. In this manner, each generation can create a posterity populace dependent on three center systems: selection, crossover, and mutation. These systems are motivated from the thought of common determination in accomplishing the best applicant quality while keeping up the decent variety to dodge youthful combination and stagnation to Local Optima (LO) during the GA-based enhancement steps [14].

Unlike most methods, this model uses Random Weight Network as base classifier in place of k-nearest neighbor. RWN is a multi-hidden-layered neural network with very fast learning speed and better generalization performance. It is an automated method and requires no human-intervention for setting parameters like learning rate [8].

Three datasets namely SpamAssassin, CSDMC2010, and LingSpam are used for the proposed model experimentation.

The proposed method can be divided into stages:

I. Feature Extraction: The three datasets are constructed based on SpamAssassin, CSDMC2010, and LingSpam. Then EMFET is used to convert these email corpuses into feature sets.

II. Feature Selection: False Spam method is executed in this step on the training dataset to get rid of features which are not relevant.

III. Evaluation and Assessment: The RWN network is tested of its predictive powers in this method. The analysis is done on the matrices like Precision, accuracy, and recall.

IV. Feature Importance Analysis: Further analysis is done to identify the most influencing features in the dataset. It helps in planning progressively exact spam channels.

Feature selection, Auto-tuning of hidden neurons, and evolution of model, all tasks take place simultaneously. The model recognizes the most applicable highlights and improves the config. of its Centre classifiers. The model at that point recognizes the spam messages dependent on its RWN.

The proposed model shows more accuracy than SVM, Naive Bayes and nearest neighbor. On the other hand, SVM delivered slightly better Recall and Precisions value with RWN as second. While RWN gave highest value for Recalls and Precision RWN model gave highest G-mean value. The performance varied slightly on different datasets but overall the Auto-GA-RWN gave the most promising results for spam email detection.

The Auto-GA-RWN method is evaluated to find out that it can hit very promising figures and it is capable of updating its own classifier over time with most relevant features. The proposed model is very capable with very few limitations and is very application oriented in detection on spam emails over the internet.

D. Clustering

Recently, there have been many machine learning techniques which are both supervised and unsupervised. Even though better results can come through the supervised techniques, they lack the flexibility and applicability. Clustering is an unsupervised machine learning technique. Clusters are data points that are more similar to each other than the other data points. They are different types of clustering methods: Centroid based, density based, connectivity based and distribution-based clustering.

In spam filtering, there has been research regarding using clustering as a way to detect spam. Some research suggests that the ham and spam emails can be divided into clusters using semi supervised clustering method [10]. But most of the conventional clustering methods have some limitations. The big micro clusters are not very accurate as they have asymmetric distribution. This may lead in-accurate results while clustering the incoming stream.

Thus, the authors of this paper proposed INB-DenStream Clustering method as a way to make past this inaccuracy in the conventional methods. It acts similar to the DenStream method but the Euclidean distance used in the online phase with a set of INB classifiers. The main reason for replacing the Euclidean distance was to consider the microclusters that do not have symmetric distribution. The methods present at the moment only considered the mean which did not give a very accurate result. Through the method they proposed, they planned to take the mean as well as the boundary of the microclusters. Their method starts off similar as that of the DenStream Clustering method by calculating the Euclidean distance and classifying them into clusters. But by the second window of data, the population is checked with a minimum value called MinC and on excession, an INB classifier was assigned to that cluster. The INB classifier took the information like the mean and variance of the clusters. As the data comes in,

the process is repeated until it exceeds a value called the SimThreshold and is assigned to the micro cluster which has the INB with higher probability. They have applied the methods on datasets-I, II, III, IV which were created for that purpose [11].

The proposed method was ensuring that the data does not take too much memory by only keeping important information. It also ensures adaptability by updating and retraining the INB classifiers as time goes by. This way the method ensures low computational complexity with low usage of memory. But there are undesirable results when the datasets are small and DenStream clustering seems to be having higher clustering than INB-DenStream method.

In conclusion, the method proposed has shown evident improvement to the other methods it was compared to. Although there are drawbacks to this method when the micro cluster is very small, but the improvements shown outshine these drawbacks. [1]

E. Collective-based framework

With time, Spammers are also getting smarter. They are adopting new strategies and tricks to exploit social media platforms by changing characteristics of the spammed tweets. This variation in the concept of spamming is known as spam drift. Moreover, spammers launch their contents in frequent manner in a very short period of time on trending topics in order to take maximum benefits out of users. They use a set of services provided by Twitter to target their attacks like URL [12], Hashtags, and mentions. To automate the spamming process, spammers make use of APIs provided by Twitter to developers.

Another system is proposed to manage this spam drift. It utilizes unsupervised ML to hold a real-time regulated tweet-level spam recognition model in bunch mode. It adaptively finds and learns the examples of new spam exercises. For the application of these methods on Machine Learning, supervised annotated datasets are required. But it costs a huge amount of time and resources to generate such an annotated dataset. Even if we go through the trouble to develop the required dataset, due to spam drift it gets outdated and require continuous adaptation to learn about the new spamming patterns and behavior. Thus, utilizing static dataset to prepare a classification model is very wasteful.

To handle this confinement, a structure of an online collective-based spam tweets characterization system is suggested that uses the extraordinary benefits of unsupervised ML techniques, to occasionally and naturally give a annotated dataset by which refreshed supervised classification models can be delivered. The model utilizes the relationship between social spammers' tweets in a brief period to foresee spamming conduct [7].

The proposed model uses ground truth dataset which comprises of tweets directly observed through different ways like manual inspection, clustering and blacklists. The framework is divided into two different modules for different needs. The first module is used for real-time tweet filtering, whereas the other module is used for periodic classification model learning to keep the dataset up-to-date. Latter is the core of the framework.

The first module uses predefined light features to prepare a feature vector for a streamed tweet. Then, the vector is passed through an already learned classifier, which predicts and assigns class label to the streamed tweets. These tweets are again saved by the second module in a database to create a new training dataset once a certain amount of new streamed tweets is stored.

After fulfilling the state of streamed tweets, another feature space is readied utilizing all clarified tweets in the capacity segment. At last, an old-style managed learning strategy (e.g., Random Forest, SVM, J48) is applied to the new labelled feature space to construct a binary classification model to supplant the present classifier model. [7]

An unsupervised clustering method is used to establish a relationship between spam accounts and their tweets. This is achieved in a 5-stage process where firstly, we extract the users of the streamed tweet and then form a cluster according to user's account age. In the third stage, a characterized number of communities is distinguished for each cluster through an improvement procedure. At that point, hand-structured features are separated for every community by utilizing just user's tweet and account data. Furthermore, in the last stage, a choice is made about every community utilizing a straightforward discriminative classification model which is based in the features.

This model names each tweet of spam communities as spam tweets.

Table-I Shows all the metrics for various algorithms.

Algorithms	Feature extraction model (unsupervised)		
	Precision (%)	Recall (%)	F - measure (%)
SVM	97.8	98.8	97.8
RF	97.6	97.1	97.1
C4.5	95	94.2	94.2

VI. EVALUATION METHODOLOGIES

Performance of any ML model or algorithm can be evaluated or put into scale with the predefined metrics such as Precision, Accuracy, F-measure and all these are calculated with the help of confusion matrix. Also, these help us identify the limitations and efficiency of the models

A. Precision:

This entity tells us about the positive identification percentage or ratio was correct.

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

B. Recall:

It tells us about how much a correct value is positively chosen correct.

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

C. F – MEASURE:

It is a weighted harmonic mean of the parameters calculated such as precision and recall of the approach.

D. ACCURACY:

It helps evaluating classification models, it is percent or fraction of value of predications that are absolutely correct to the total predictions made.

$$Accuracy = \frac{Number\ of\ correct\ prediction}{Total\ number\ of\ predictions}$$

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

E. CONFUSION MATRIX

A confusion matrix is 2X2 matrix consisting of 4 quantities namely True Positive, True Negative, False Positive, False Negative. These are used to calculate the parameters that are discussed above, it is the foundation for any of the above parameter's calculation.

Table-II Confusion matrix

	Positive	Negative
Positive	True Positive	False Positive
Negative	False Negative	True Negative

Table-III shows all the performance metrics for ontology-based approach.

Algorithms	Feature extraction model (Semantic based) - using topic guessing		
	Accuracy	FN	FP
SVM	90.8	7.4	1.8
RF	99.2	0.6	0.2
C4.5	97.6	1.2	1.2

Table IV: Comparison of evaluation parameters of Semantic based Feature Extraction model

Algorithms	Feature extraction model (Semantic based) - using topic guessing		
	Accuracy	FN	FP
SVM	90.8	7.4	1.8
RF	99.2	0.6	0.2
C4.5	97.6	1.2	1.2

Table V: Comparison of evaluation parameters of Ontology based model

	Ontology based			
	Accuracy (%)	Precision (%) (0, 1)	Recall (%) (0, 1)	F-measure (%) (0, 1)
SVM	88.13	90, 86	86, 90	88, 88
RF	94.7	93, 91	92, 93	93, 92
C4.5	-	-	-	-

Table VI: The evaluation parameters used in 5CNN1FB model.

Method	Accuracy	Precision	Recall	F-measure
CNN + Glove 200d ns	0.912	0.711	0.945	0.812
CNN + Google 300D ns				
CNN + Edinburgh 400d ns	0.952	0.869	0.895	0.822
CNN + HSpam 200d ns				
CNN + Random	0.936	0.782	0.943	0.855
	0.939	0.796	0.938	0.861
	0.922	0.785	0.839	0.811
Proposed Method (5 CNN+ 1 Feature based)	0.957	0.88	0.909	0.894

Table VII: Data sets used for respective methods.

Data Sets	Method
Hspam14 Data Set, 1ks10kn	Neural Network-Based Ensemble Approach
raw and unstructured data	An Ontology-Based Tweet Spam Detection
Ground-Truth	Collective Approach Of Unsupervised And Supervised Model
Spamassassin	Identification Of The Most Relevant Features With Random Weight Networks
Lingspam	
Csdmc2010 Corpus	
Dataset -I, -II, -III and -IV [11]	Stream Clustering Framework
Enron Data	Unsupervised Feature Learning
Imdb Data	
Trec07	

Csmining Spam Emails Datasets	Semantic-Based Feature Selection
Concept Drift In E-Mail Datasets	
Spam-Posts Detection Dataset Automated	Based On User Activity And Behaviour
Honeypot	
Spam-Posts Detection Dataset Manual	

All the data is collected from the datasets which showed best output for the approach proposed and
also, the data is rounded to its nearest digits by approximation. For Feature extraction using unsupervised model, SVM is more efficient and more suggestive due to its metrics being high in all aspects also when compared with feature extraction using semantic based topic guessing Random forests showed high values for Accuracy, FN and FP. Similarly, for ontology-based approach Random Forest showed better efficiency.
The above table shows the evaluation parameters used in 5CNN1FB model. The proposed ensemble method which took in the 5 CNNs and feature based method has accuracy, precision and F-Measure metrics has better performance. Since there are more features in the model, more execution time is taken. But the proposed method outperforms all the base methods used nowadays in case of both the balanced and the imbalanced dataset.

Table VIII: The evaluation parameters used in INB-Denstream clustering model.

Evaluation Parameter	Datasets and their respective results
F1- Measure	Dataset I (63.7), Dataset II (60.6), Dataset III(51.3), Dataset IV(49.9)
Purity	Dataset III(98.31), Dataset IV(78.15)

The F1 Measure that the table shows is taken from the full data set. It was seen that the proposed method outperformed the other clustering methods present today. The purity measures could also be seen outperforming the standard DenStream Clustering method.
All the proposed methods have all shown improvements to the preceding methods but they do have some minor shortcomings. Like INB-Denstream clustering method seems to be working better for bigger datasets than the smaller ones. But this could be overlooked as the chances of the dataset being small in a real time scenario is very slim. Similarly the disadvantages of the other methods could be ignored as they are very minute. As overall, the accuracy, efficiency and the performance of the methods surpasses the disadvantages.

VII. CONCLUSION

This paper gives a survey of different methods proposed in the domain of spam filtering. We discussed methods by dividing them into different categories such as feature based, ontology based, neural network based, cluster based and collective framework-based approaches. Although there are methods in detecting spam today, the spam content is getting clever to evade these detecting methods. Therefore, new and updated methods seem necessary. After surveying all papers which include different methods and algorithms in spam filtering, we made a generic architecture that combines the most important parts in the methods that the papers have proposed. A detailed description of the methods and our views on them was presented. We have then used different evaluation parameters such as F-measure, precision, accuracy etc to compare the results in the papers. We also gave a comparison on the algorithms and data sets followed by these papers. The results of these comparisons were thus apparent. They all performed better than the most methods used today. They have shown higher accuracy, efficiency and performance. They have minute disadvantages that could be overlooked. In the future, it would be better if there are advancements in removing even these minute disadvantages.

REFERENCES

1. Tajalizadeh, H., & Boostani, R. (2019). A Novel Stream Clustering Framework for Spam Detection in Twitter. IEEE Transactions on Computational Social Systems, 6(3), 525-534.
2. Madisetty, S., & Desarkar, M. S. (2018). A neural network-based ensemble approach for spam detection in Twitter. IEEE Transactions on Computational Social Systems, 5(4), 973-984.
3. Halawi, B., Mourad, A., Otrok, H., & Damiani, E. (2018). Few are as good as many: an Ontology-based tweet spam detection approach. IEEE Access, 6, 63890-63904.
4. Diale, M., Celik, T., & Van Der Walt, C. (2019). Unsupervised feature learning for spam email filtering. Computers & Electrical Engineering, 74, 89-104.
5. Méndez, J. R., Cotos-Yañez, T. R., & Ruano-Ordás, D. (2019). A new semantic-based feature selection method for spam filtering. Applied Soft Computing, 76, 89-104.
6. Inuwa-Dutse, I., Liptrott, M., & Korkontzelos, I. (2018). Detection of spam-posting accounts on Twitter. Neurocomputing, 315, 496-511.
7. Washha, M., Qaroush, A., Mezghani, M., & Sedes, F. (2019). Unsupervised Collective-based Framework for Dynamic Retraining of Supervised Real-Time Spam Tweets Detection Model. Expert Systems with Applications.
8. Faris, H., Ala'M, A. Z., Heidari, A. A., Aljarah, I., Mafarja, M., Hassonah, M. A., & Fujita, H. (2019). An intelligent system for spam detection and identification of the most relevant features based on evolutionary random weight networks. Information Fusion, 48, 67-83.
9. Youn, S. (2014). SPONGY (SPam ONtoloGY): Email classification using two-level dynamic ontology. The Scientific World Journal, 2014.
10. Whissell, J. S., & Clarke, C. L. (2011, September). Clustering for semi-supervised spam filtering. In Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (pp. 125-134). ACM.
11. Chen, C., Zhang, J., Xie, Y., Xiang, Y., Zhou, W., Hassan, M. M., ... & Alrubaian, M. (2015). A performance evaluation of machine learning-based streaming spam tweets detection. IEEE Transactions on Computational social systems, 2(3), 65-76.
12. Dangkesee, T., & Puntheeranurak, S. (2017, November). Adaptive Classification for Spam Detection on Twitter with Specific Data. In 2017 21st International Computer Science and Engineering Conference (ICSEC) (pp. 1-4). IEEE.
13. D.S. Weile, E. Michielssen, Genetic algorithm optimization

applied to electromagnetics: a review, IEEE Trans. Antennas Propag. 45 (3) (1997) 343–353.
14. H. Mühlenbein, D. Schlierkamp-Voosen, Predictive models for the breeder genetic algorithm i. continuous parameter optimization, Evol. Comput. 1 (1) (1993) 25–49.
15. Bishop, Christopher (2006). Pattern recognition and machine learning. Berlin: Springer. ISBN 0-387-31073-8.
16. Li, F. H., Huang, M., Yang, Y., & Zhu, X. (2011, June). Learning to identify review spam. In Twenty-second international joint conference on artificial intelligence.
17. Benczur, A. A., Csalogany, K., Sarlos, T., & Uher, M. (2005, May). Spamrank–fully automatic link spam detection work in progress. In Proceedings of the first international workshop on adversarial information retrieval on the web (pp. 1-14).
18. Carpinter, J., & Hunt, R. (2006). Tightening the net: A review of current and next generation spam filtering tools. Computers & security, 25(8), 566-578.
19. Agarwal, N., & Yiliyasi, Y. (2010, November). Information quality challenges in social media. In ICIQ.
20. Joshi, A., Finin, T., Java, A., Kale, A., & Kolari, P. (2007, October). Web 2.0 mining: Analyzing social media. In Proceedings of the NSF symposium on next generation of data mining and cyber-enabled discovery for innovation.
21. Gudivada, V. N., Baeza-Yates, R., & Raghavan, V. V. (2015). Big data: Promises and problems. Computer, (3), 20-23.
22. Shehnepoor, S., Salehi, M., Farahbakhsh, R., & Crespi, N. (2017). NetSpam: A network-based spam detection framework for reviews in online social media. IEEE Transactions on Information Forensics and Security, 12(7), 1585-1595.
23. Sedhai, S., & Sun, A. (2015, August). Hspam14: A collection of 14 million tweets for hashtag-oriented spam research. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 223-232). ACM.
24. Lee, K., Eoff, B. D., & 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.
25. Martinez-Romo, J., & Araujo, L. (2013). Detecting malicious tweets in trending topics using a statistical analysis of language. Expert Systems with Applications, 40(8), 2992-3000.
26. Lee, S., & 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), 183-195.

AUTHORS PROFILE

Chinthapanti Bharath Sai Reddy is a tech enthusiastic whose is pursuing his 3rd year B.Tech (C.S.E) in Vellore Institute of Technology, currently working in the domains of Machine Learning and Data Science . He Strongly believe that every problem in this world have a solution if we see the problem in the correct aspect. My dream is to create a world driven by data and a more efficient predictions for a large-scale problem. He is also a founder of a startup named dukhan.in which is a platform for buying groceries in rural areas.

Shaurya Choudhary is an adept learner exploring the depths of technology. He is currently pursuing his 2nd year B.Tech (CSE) from Vellore Institute of Technology, Vellore. He is a core member of Apple Developers Group as well as IEEE-PCS. He is mainly focused in the domains of Machine Learning, Blockchains, Ethical hacking, and Android development. Apart from these, he also shares a keen interest in the field of designing using Adobe Photoshop. He has worked on several projects as a part of VIT curriculum as well as outside of it.

Saravanakumar Kandasamy is working as Associate Professor in the School of Information Technology and Engineering, Vellore Institute of Technology, Vellore since June 2008. He has completed his masters MTech (CSE) in Indian Institute of Technology Guwahati and PhD in Vellore Institute of Technology. He has a total of 19+ years of experience in teaching. His research interests include Advanced DBMS concepts, Natural Language Processing, Information Retrieval and Question Answering Systems. His current research interests are Database, Information Retrieval, Natural Language Processing, and Question Answering Systems.