A Survey on Support Vector Machine Using Spam Filtering Techniques

Shiva Sharma[1], U.Datta[2]
[1]Dept of CSE/IT
[2]Dean Academic
[1, 2] Maharana Pratap College of Technology, Gwalior, India

Abstract- *Spam is unsolicited, junk email with variety of shapes and forms. To filter out spam, numerous strategies are used. Techniques like Naïve Bayesian Classifier, Support Vector Machine (SVM) are regularly used. Also, some of tools for spam filtering either paid or unfastened are to be had. Amongst all techniques SVM is often used. This paper surveys exclusive spam filtering strategies. SVM is the popular machine learning techniques in spam filtering due to the fact it can handle data with large number of attributes.*

Keywords- SVM, Spam filtering, Collaborative spam filtering, Image spam etc.

I. INTRODUCTION

SVM is one of high-quality machine learning algorithms, which was proposed in 1990's and utilized essentially for design reputation. This has also been executed to many example classification issues together with image recognition, speech recognition, face detection and faulty card detection, and so forth. Pattern popularity goals to classify statistics based totally on both a priori information or statistical statistics extracted from raw data, which is a powerful tool in data partition in many controls. SVM is a directed type of machine learning calculation wherein, given an arrangement of preparing illustrations, each set apart as having a place with one of the numerous classifications, a SVM training algorithm constructs a form that predicts the class of the new example. SVM has the more ability to generalize the problem, that's the goal in statistical learning. The factual learning idea gives a characterize for studying the bother of gaining knowledge, making predictions, making choices from a hard and fast of data. In measurable learning thought the bother of supervised learning is detailed as takes after. We are given an arrangement of training data (x1, y1) (xn, yn) in Rn x R sampled in keeping with unknown opportunity probability P(x, y), and a loss characteristic V(y, f(x)) that measures the mistake, for a given x, f(x) is " predicted" instead of the real cost y. The problem consists in locating a characteristic f that minimizes the expectancy of the error on new data i.e. finding a trademark f that limits the expected errors: ∫ V(y, f(x)) P(x, y) dx dy . Early machine learning algorithms planned to learn portrayals of simple features. Henceforth, the goal of learning of turned into to output a hypothesis that finished the appropriate order of the training data and early learning calculations were composed to discover such an correct in shape to the data. The ability of a hypothesis to correctly classify data no longer inside the education set is known as its generalization. SVM performs better in time period of now not over generalization when the neural networks might come to be over generalizing without difficulty [1].

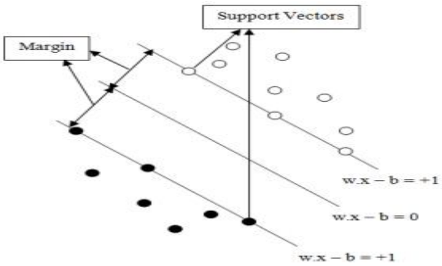

Fig 1. SVM Model

II. BASICS OF SVM

Early machine learning algorithms aimed to research representations of easy functions. The capacity of a theory to appropriately arrange data not inside the training set is called its speculation. SVM operates better in term of now not over generalization whilst the neural networks may grow to be over generalizing easily. The SVM identifies with design characterization meaning the algorithm is utilized for grouping the extraordinary types of patterns. There are two sorts of styles i.e. Linear and non- linear. Linear patterns are designs that are effectively discernable or can be effortlessly isolated in low measurement, while non-direct outlines are plans that are not effectively recognizable or can't be effortlessly isolated and subsequently these sorts of examples should be additionally controlled with the goal that they can be easily separated. SVM can likewise be reached out to learn non-direct decision capacities by first projecting the input data onto a high dimensional feature space utilizing bit works and detailing a straight characterization issue in that feature space.

The resulting feature space is much larger than the size of the dataset which are not possible to store in popular computers. Investigation on this issue leads to several decomposition based algorithms. The fundamental thought of decomposition technique is to part the factors into two sections: set of free factors called as working set, which can be updated in every cycle and set of settled factors, which are settled at a specific value briefly. This method is repeated until the point that the end conditions are met. Fundamentally SVM depends on the development of the ideal hyper plane, which can be utilized for classification for directly classification. The main working strategy at means that if the margin size is larger than it more correctly classifies the patterns. One of the hyper plane is represented to by the accompanying condition:
Hyper plane, aX + bY = C (i)

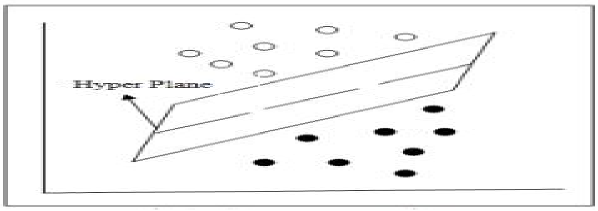

Fig 2. SVM Hyper plane

Choosing diverse kernel function is an imperative viewpoint in the SVM-based characterization, regularly utilized bit capacities incorporate LINEAR, POLY RBF, and SIGMOID.

Another important parameter in SVM is the parameter C. It is likewise known as a complexity parameter and is the sum of the distances of all factors which might be on the incorrect aspect of the hyper plane. Basically, the complexity parameter is the amount of blunders that may be omitted during the category system. But the cost of category process cannot be both too more or too small. If the cost of complexity parameter is just too greater than the performance of category is low and vice versa. The foremost principle of SVM is that given a fixed of impartial and identically distributed training pattern (xi, yi)N i=1, wherein x ϵ Rd and yi ϵ −1,1 ,denote the enter and output of the class. The intention is to find a hyper aircraft wT.X + b = zero, which separate the two distinct samples correctly. Therefore, the hassle of fixing most reliable class now interprets into fixing quadratic programming issues. Where we have to maximize the weight of the margin. It is expressed as: Min Φ (w) = ½ || w || 2 = ½ (w, w), Such that: yi (w. xi + b) >= 1 (iii) 2.2 Concepts used in SVM Concepts of SVM on which SVM is identified are given as bellows: The Separating hyper plane [2].

- The maximum margin hyper plane.
- Soft margin.
- The Kernel function
- These Concepts are explained for arrangement of the given arrangement of examples by building an ideal hyper plane. For any sort of patterns, human are thought to be an ultimate judge, who can without much of a stretch recognize the distinctive example given to them, however for a PC framework it is exceptionally hard to recognize and speak to it.

In the fig 2.3(a), there are two various types of examples and our job is to characterize these two examples. For this situation, it is anything but difficult to order outwardly with our naked eye as it can be outwardly divided. Be that as it may, with a specific end goal to represent to these patterns to have a place with two unique classes, a line can be drawn that isolates this example. The fig 3(b) demonstrates portrayal for the grouping of two unique examples utilizing a solitary line, gave that the examples are displayed in two dimensional space. The fig. 3(c) shows the similar type of two different patterns, but in one dimensional space. So, in order to separate these patterns, given in one dimension, a point can be used to separate it. At the point when the comparable sorts of examples that are exhibited in fig 3(b) is spoken to three dimensional space, at that point a plane can be utilized to speak to a line for the order of examples into two unique classes as appeared in the fig 3(d). The plane that isolates these two unique sorts of example spoke to in 3-D space is known as an isolating hyper plane that isolates designs. Additionally, to separate the previously mentioned designs there may exist numerous such planes as appeared in the fig 3(e) that isolates the examples mentioned above. The next task is to choose the plane from the arrangement of planes whose margin is augmented. The plane with the maximum margin i.e. opposite separation from the negligible line is known as ideal hyper plane or most extreme edge hyper plane as appeared in fig 3(f). The illustrations that lie on the edges of the plane are called support vectors. Amid the characterization and portrayal of examples, there may exist a few mistakes in the portrayal, as appeared in the fig 3(g), such sorts of errors is called delicate edge. Amid order of such kind of examples portrayal, the mistake can be disregarded to some limit value.

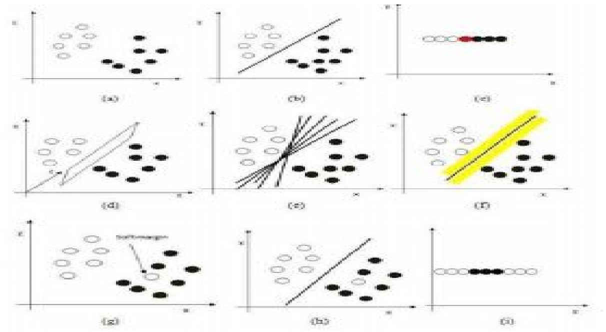

Fig.3 Classification concept using SVM

III. STRENGTHS AND WEAKNESSES OF SVM CLASSIFIERS

With regards to pattern recognition based face detection and acknowledgment, SVM are an attractive order system:

a) Non- linearity of the classifier. The kernel approach of SVMs takes into consideration self-assertive complex choice limits amongst confront and nonface cases, bringing about high accuracy rates.
b) Resistance to over-preparing. The maximum-margin ensure makes SVMs impervious to overstraining and basic overlooking, and to some degree manages exception issues.
c) Human interpretative learning process. The natural geometric understanding of the SVM as a hyper-plane isolating two classes of focuses contends about the conduct of the classifiers, while the idea of help vectors gives knowledge into what cases in the preparation set are great delegates of the object class.
d) Effective estimate techniques. The geometric idea of the SVM algorithm permits the advancement of approximation techniques that can be utilized to get extremely productive classifiers [3].

IV. SPAM FILTERING

Spam filtering in Internet email can work at levels, an individual user level or an enterprise level (see Figure 4). An person user is commonly a person operating at home and sending and receiving electronic mail via an ISP. Such a person who needs to discover and filter out unsolicited mail electronic mail installs a existing mail filtering gadget on her character PC. This system will either interface without delay with their present mail user agent (MUA) (extra normally known as the mail reader) or greater generally will act as a MUA itself with complete capability for composing and receiving email and for coping with mailboxes.

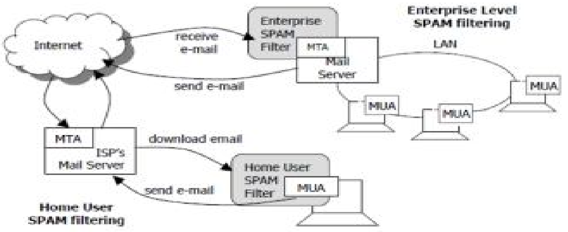

Fig.4. Alternatives for spam filtering in Internet e-mail.

Enterprise- level spam filtering filters mail as it enters the internal network of an organization. The software program is established on the mail server and interacts with the mail transfer agent (MTA) classifying messages as they may be received.

Spam email, that's recognized by way of the spam filter, spam filter, , may be categorized as a spam message for all users on that network. Spam can be filtered at an individual level on a LAN also. A networked user can select to filter spam locally as it's downloaded to their PC at the LAN by installing an appropriate system. The tremendous majority of current spam filtering structures use rule-based scoring strategies. A set of rules is applied to a message and a rating accumulates based at the policies which might be actual for the message. Systems generally encompass loads of guidelines and those rules want to be up to date regularly as spammers regulate content material and behavior to keep away from the filters. Systems also include listing-based totally strategies in which messages from identified users or domains can be automatically blocked or allowed through the filter.if the rating for an email exceeds a threshold, the email is classified as spam. Limited learning abilities are beginning to seem in structures inclusive of Mozilla and the MacOS X Mail application however these structures are nevertheless of their infancy. Naïve Bayes gives off an impression of being the strategy for decision for adding a learning ability to commercial spam filtering systems. The design of spam filtering is appeared in Fig. 5. Firstly, the model will accumulate person emails that are considered as both spam and legitimate electronic mail. After collecting the emails the preliminary transformation system will begin. This model comprises of beginning transformation, the person interface, work extraction and determination, email records classification, and analyzer segment. Machine learning algorithms are utilized finally to prepare and test whether the demanded email is spam or legitimate.

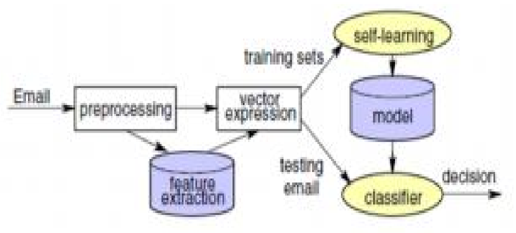

Fig 5. The process of spam filtering

V. SPAM TECHNIQUES

In case a sponsor has one database containing names, addresses, and phone quantities of arranged customers, they can pay to have their database facilitated against an outer database containing email addresses. The organization at that point has the way to send email to people who have not asked for email, which may incorporate people who have intentionally withheld their email address.

a) Image spam

Image spam is an obfuscating method in which the substance of the message is secured as a GIF or JPEG image and showed in the email. This keeps content based spam channels from detecting and blocking spam messages. Image spam was supposedly utilized as a part of the mid 2000s to advertise "pump and dump" stocks [4].Often, image spam contains nonsensical, computer created content which just bothers the reader. Notwithstanding, new innovation in a few projects endeavor to peruse the images by endeavoring to discover message in these images. They are not exceptionally precise, and in some cases sift through pure pictures of items like a crate that has words on it. A more up to date system, in any case, is to utilize an energized GIF image that does not contain clear content in its underlying edge, or to bend the states of letters in the image (as in CAPTCHA) to keep away from recognition by OCR tools.

b) Blank spam

Blank spam will be spam without a payload advertisement. Frequently the message body is missing out and out, and also the headline. In any case, it fits the meaning of spam as a result of its temperament as mass and spontaneous email. Blank spam might be started in various ways, either intentional or unexpectedly:

1) Blank spam can have been sent in a directory harvest attack, a type of word dictionary attack for social event substantial locations from an email service provider. Since the objective in such an attack is to utilize the bounces to isolate invalid locations from the substantial ones, spammers may get rid of most components of the header and the whole message body, and still achieve their objectives.
2) Blank spam may in like manner happen when a spammer ignores or for the most part fails to incorporate the payload when he or she sets up the spam run.
3) Often blank spam headers seem truncated, recommending that PC glitches may have added to this issue from inadequately composed spam programming to malfunctioning relay servers, or any issues that may truncate header lines from the message body.
4) Some spam may give off an impression of being clear when in certainty it isn't. A case of this is the VBS. Davinia. B email worm which spreads through messages that have no title and seems clear, when in truth it utilizes HTML code to download different documents.

c) Backscatter spam

Backscatter is a side-effect of email spam, viruses and worms, where email servers receiving spam and other mail send skip messages to a innocent gathering. This happens in light of the fact that the first message's envelope sender is produced to contain the email address of the casualty. A substantial extent of such email is sent with a manufactured From: header, coordinating the envelope sender. Since these messages were not requested by the beneficiaries, are generously like each other, and are conveyed in mass amounts, they qualify as spontaneous mass email or spam. All things considered, frameworks that produce email backscatter can wind up being recorded on different DNSBLs and be infringing upon internet service providers' Terms of Service.

VI. CLASSIFICATION OF SPAM FILTERING METHODS

Depending upon utilized systems spam filtering methods are by and large isolated into two classes:

- Methods to avoid spam distribution in their roots;
- Methods to avoid spam at destination point. We should consider these strategies in point by point form.

1) Methods to Avoid Spam Distribution

Legislative measures limiting spam distribution, improvement of email protocols utilizing sender authentication, blocking mail servers which distribute spam are the techniques which stay away from spam distribution in starting point. Utilizing these strategies alone doesn't give extensive outcomes. For instance, there are numerous hard authoritative limitations for spam distribution in USA; all things considered, the best measure of spam is distributed from this locale. One reason is a presence of high level broadband Internet access in USA.

There are a portion of the methodologies, offering to make spam sending economically unrewarding. One of these statements is to influence sending of every e-to mail paid. The payment for one email ought to be the amazingly inconsequential. For this spammers for the typical client it will be subtle. For spammers who send thousand and millions messages the cost of such mailing ends up plainly extensive that makes it financially unbeneficial. This kind of strategies keeping away from spam in their starting points is a subject of creator's another papers. They ought to be actualized together with the strategies depicted in the following area, which channel spam at the destination point [5].

2) Methods to Avoid Spam Receiving

Strategies which filter spam in goal point can be isolated into the accompanying classes:

- Depending on utilized theoretical methodologies: customary, learning-based and hybrid techniques;
- Depending on filtration area: server side, customer side and filteration public mail-servers.

VII. CLASSIFICATION OF SPAM FILTERING METHODS

Depending on Theoretical approaches as we noted above relying upon utilized theoretical methodologies spam filtering techniques are separated into customary, learning-based and hybrid strategies. In traditional techniques the order show or the data (rights, patterns, keywords, arrangements of IP locations of servers), in light of which messages are characterized, is characterized by master. The data storage gathered by experts is called as the learning base. There are additionally utilized trusted and questioned senders records, which help to choose legal mail. Actually it makes sense only creation of the "white" rundown, since spammers utilize invented email addresses. This strategy can't speak to itself as a high- grade anti-spam filter, however can diminish significantly measure of false operations, being a piece of email filtration framework in light of other characterization techniques. In learning-based strategies the grouping model is created utilizing Data Mining procedures. There are some problems from the point of view of data mining as changing of spam content with time, the proportion of spam to legitimate mail, insufficient amount of training data are characteristic for learning-based methods.

Traditional methods. Traditional techniques are separated into the accompanying classifications:

a. **Methods based on analysis of messages.** The received e-mail is analyzed for specific signs of spam on the base of:
   - formal signs;
   - Content utilizing signature in updated database;
   - content applying measurement techniques in light of Bayes theorem;
   - Content by methods for utilize SURBL (Spam URL Realtime Block Lists), when run look for found references in email and their confirmation under base of SURBL. This technique is powerful if rather than advertisement, the reference of website with commercial is located in email.

b. **Detectors of mass distribution.** Their assignment is to identify circulations of comparative messages to the greater part of clients. The accompanying strategies are utilized for the discovery:
   - users' voting (Razor/Pyzor)
   - Analysis of e-mails coming through mail framework (DCC); receipt of email to the spam "trap" and its following analyses (executed in Symantec Brightmail Anti-Spam). Independent from a method for mass location the possibility of a strategy is that for spam filtration the figured email signature (the control total) is utilized. For the techniques in view of recognition of redundancies two essential issues are trademark. The first is a spam “personification”. This means that each spam e-mail has insignificant differences at the cost of which it is hard to collect steady signatures. To take care of this issue the different steady marks are utilized. For instance, in Yandex Mail System the strategy for shingles is figured it out. The second issue is recognition of legitimate bulk mailings.

c. **Methods based on acceptance of sender as a spammer.** These strategies depends on various blackhole lists of IP and email addresses. It is conceivable to apply claim blackhole and white records or to utilize RBL services (Real-time Blackhole List) and DNSBL (DNS-based Blackhole List) for address. These strategies depends on various blackhole arrangements of IP and email addresses. It is conceivable to apply possess blackhole and white records or to utilize RBL administrations (Realtime Blackhole List) and DNSBL (DNS-based Blackhole List) for address verification. Advantage of these methods is detection of spam in early step of mail receiving process. Disadvantage is that the policy of addition and deletion of addresses is not always transparent. Often the entire subnets belonging a place with suppliers get to the Black records. For such frameworks it is really difficult to evaluate the level of false positives (the legitimate e-mail wrongly named spam) on real mail streams.

d. **Methods based on verification of sender's e-mail address and domain name.** This is the least difficult technique for filtration if DNS ask for's name is the same with the space name of sender. But spammers can use real addresses, so that current method is ineffective. In this case it may be verified with possibility of sending the message from current IP address. Firstly, the Sender ID technology can be utilized where sender's email address is protected from falsification by methods for publishing the policy of area name use in DNS. Furthermore, there can be utilized SPF (Sender Policy Framework) technology, where DNS protocol is utilized for check of sender's email address. The standard is that if domain's owner needs support SPF verification, at that point he adds special entry to DNS passage of his domain, where shows the arrival of SPF and scopes of IP addresses from where may turn into an email from clients of current space.

e. **Method based on SMTP server response emulation.** If the real mail delivery systems, which follow the SMTP protocol correctly, observe such error, they get some interval (1 - 2 hours) and repeat attempt again. Be that as it may, the dominant part of spam-bots has brief time out periods. So channels in perspective of this method back off the SMTP transaction to the point that some SPAM senders will bomb however where real mail delivery systems will at introduce continue and convey mail the larger part of spam-bots has brief time out periods successfully. All above methods are based on some data for analysis collected by experts of third-party suppliers and same for all users. With the goal that customary strategy's has the accompanying burdens:

- It is important to refresh the learning base consistently;
- There is a reliance on refresh suppliers;
- The security level is low;
- "Impersonalized" model of characterization doesn't consider singular specifics of user's correspondence; reliance on characteristic language of correspondence;
- Low level of detection in light of general models of characterization.
- **Collaborative spam filtering.** This is gathering spam reports between P2P customers or from mail server (Google Gmail). The collaborative centralized spam filtration is more monetary in correlation with individual approach, however only under state of essence of satisfactory methods of the examination of false operations and agent renaming of not accurately grouped messages. In the papers it is proposed such sort of multi- agent spam filtration and personalized cooperative spam filtering.
- **Social networking against spam.** This is a one of the most recent techniques where the data extricated from interpersonal organizations is utilized to fight spammers. For instance, P.A. Chirita et al. evaluate the rank of clients relying upon their interpersonal organization exercises and dependable senders are positioned and delegated spam or non-spam. They call this calculation as Mail Rank mapping and demonstrate that it is exceptionally safe against spammer attacks, which clearly must be viewed as ideal from the earliest starting point in such an application situation. So if there should be an occurrence of learning-based techniques client characterizes the order display himself, with the goal that the greater part inconveniences of traditional methods are solved successfully; intellectual strategies are autonomous, autonomous on external knowledge base, doesn't require standard refresh, multilingual, free of regular dialect, prepared to examine new sorts of spam user supported [5].

There is advantage as construction of personalized mail classification model, where user himself defines which mail is legal or which one is a spam. Therefore learning-based methods have higher rank in spam determination. In many spam filtration systems in light of the learning-based methods the Bayes' theorem, Marcov's chain and others are successfully associated. Learning-based strategies have additionally two or three disadvantages as over fitting, reliance on quality and compound of trainee set, resource intensivety. Use of measurement calculations with muddled mathematic counts prompted high loading of computing system's resources [6]. For the spam filtering systems processing fair amount of requests the productivity of algorithm is a prime significance, so resource-intensivety aspect is the most essential disadvantage of learning based totally strategies.

- **Hybrid methods.** One of the trendy procedures in unsolicited mail filtering is hybrid filtration gadget that is a combination of different algorithms, in particular in the event that they use unrelated functions to produce a solution. In this situation it could be carried out various filtering strategies and get the benefits of the conventional and gaining knowledge of-based totally methods [7].

VIII. LITERATURE SUVEY

M. Kepa et al. [8] In this paper, authors presented a hybrid classification model that uses k-nearest neighbor and SVM techniques. This technique is two degree method based totally on the one-vs-near scheme turned into tested on large datasets. In the first level, the KNN classifier is used to compute the class neighbor list that's learning section. The KNN figures the gap between each centroid within the shape of an ordered list which is used in second level classifier. The 2d stage SVM uses the saved neighbor listing to restrict the dataset used for training the classifier for a single category.

R. C. Barik et al. [9] In this paper, authors delivered a completely unique characteristic extraction approach and then classifies with linear SVM. This method is two level technique primarily based on the one-vs-near scheme changed into tested on massive datasets.In the principal stage, the KNN classifier is utilized to figure the class neighbor posting that is learning area. The KNN figures the space among each centroid as a requested posting which is utilized as a part of second stage classifier. The second stage SVM utilizes the spared neighbor rundown to constrain the dataset utilized for preparing the classifier for a single category.

R. C. Barik et al. [9] In this paper, authors introduced a completely unique feature extraction approach and then classifies with linear SVM. In this approach, first, time domain and frequency domain analysis is done to the original data signal by translating and scaling co-efficient using Discrete Wavelet Transform (DWT) of mother wavelet Daubechies level 1 and then level 2 decomposition respectively. After that the subsequent transformed data is attended to a statistical method as Multi-Dimensional Scaling (MDS) to find out the similarities and dissimilarities to classify the precise class. Then for classifies the data SVM is applied on the data.

R. Bruni et al. [10] In this paper, creators developed a class show grounded on Logical Analysis of Data (LAD). This paper offers with the hassle of producing a fast and unique data type, getting to know it from a probable small set of facts which can be already labeled. In this style, data should be encoded into binary frame by a discretization way called binarization. This is performed via using the training set for figuring genuine values for each field, referred to as reduce-points in the case of numerical fields that cut up every zone into binary qualities. The specific binary qualities constitute a support set, and are blended for creating logical principles known as examples. Patterns are used to classify each unclassified record, on the origin of the sign of a weighted sum of the patterns activated by that record.

A. Chaudhuri, et al. [11] In this paper, creators introduced a solitary fuzzy support vector machine (FSVM) device or a variation of FSVM called modified fuzzy SVM. This version is to categorize the credit approval problem. In FSVM, every pattern is given a fuzzy membership which denotes the mind-set of corresponding point in the direction ofone magnificence. The membership highlight which is a hyperbolic tangent kernel grasps the impreciseness in preparing tests. In MFSVM, the victory of the classification lies in proper selection of the fuzzy membership function which is a function of center and radius of each class in feature space and is represented with kernel. The kernel used in MFSVM is hyperbolic tangent kernel. This kernel permits lower computational cost and higher rate of idealistic eigen estimations of kernel matrix which facilitates a few constraints of different kernels.

E. Baralis et al. [12] In this paper, authors proposed an innovative and active approach to estimate the Bayesian probability. The Entropy-based totally Bayesian classifier, referred to as EnBay, emphases on selecting the minimal set of long and no longer overlapped patterns that exceptional conforms to a conditional-independence model, essentially in light of an entropy-based evaluator. Additionally, the possibility approximation is quite tailored to every group. This model works on intervals which may be 1) segment the trait set into a base range of huge subsets all together that their conditional dependence, given an arbitrary magnificence, is minimized, 2) pick out common item units considered by conditionally unbiased attribute sets.

IX. CONCLUSION

Social spam is an e-crime on social networking sites with contents such as comments, post, chat, etc. There are many spamming activities going through social media such a malicious links posting, insulting posts, hate speech, fake friends, deceitful reviews, etc. SVM are supervised learning models with associated learning model that analyze data and are in particular used for type cause. SVM are set of associated supervised getting to know strategies used for class and regression. SVM map enter vector to a better dimensional plane wherein a maximal setting apart hyper plane is constructed.

REFERENCES

[1] Ashis Pradhan ‘”Support vector machine-A Survey" International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, Volume 2, Issue 8, August 2012) .

[2] Ms. Snehal S. Joshi, Mr. Navnath D. Kale “Survey: Support Vector Machine and Its Deviations in Classification Techniques” Volume 4, Issue 12, December 2014 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering.

[3] Ms. Ruchida S. Sonar, Dr. P.R. Deshmukh “Support Vector Machines for Human Face Detection: A Review” International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 2 Issue: 11.

[4] Omar Saad, Ashraf Darwish and Ramadan Faraj “A survey of machine learning techniques for Spam filtering” IJCSNS International Journal of Computer Science and Network Security, VOL.12 No.2, February 2012.

[5] Saadat Nazirova “Survey on Spam Filtering Techniques” Communications and Network, 2011, 3, 153-160.

[6] N. Cristianini, B. Schoelkopf, "Support vector machines and kernel methods, the new generation of learning machines". Artificial Intelligence Magazine, 23(3):31–41, 2002.

[7] P. Cortez, C. Lopes, P. Sousa, M. Rocha and M. Rio, “Symbiotic Data Mining for Personalized Spam Filtering,” IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Milan, 15-18 September 2009, pp. 149-156.

[8] M. Kepa, J. Szymanski, "Two stage SVM and kNN text documents classifier," In: Pattern Recognition and Machine Intelligence, Kryszkiewicz M. (Ed.), Lecture Notes in Computer Science, Vol. 9124, pp. 279-289, 2015.

[9] R. C. Barik and B. Naik, "A Novel Extraction and Classification Technique for Machine Learning using Time Series and Statistical Approach," Computational Intelligence in Data Mining, vol. 3, pp. 217-228, 2015.

[10] R. Bruni and G. Bianchi, "Effective Classification Using a Small Training Set Based on Discretization and Statistical Analysis," IEEE Trans. Knowl. Data Eng., vol. 27, no. 9, pp. 2349-2361, 2015.

[11] A. Chaudhuri, "Modified fuzzy support vector machine for credit approval classification," IOS Press and Authors, vol. 27, no. 2, pp. 189-211, 2014.

[12] E. Baralis, L. Cagliero, and P. Garza, "EnBay: A novel pattern-based Bayesian classifier," Tkde, vol. 25, no. 12, pp. 2780- 2795, 2013.