iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 2 **Issue:** XII **Month of publication:** December 2014

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Efficient Filteration of Unwanted Messages in Social Networking Sites

Gunduboina Penchalaiah[1], Mr. P Jagadeeswara Rao[2]
[1, 2] Computer Science and Engineering

Abstract: ***Social networking sites that facilitate communication of information between users allow users to post messages as an important function. Unnecessary posts could spam a user's wall, which is the page where posts are displayed, thus disabling the user from viewing relevant messages. TES ensures the "reputation" of reporters by tracking how often the larger recipient community agrees with their assessment of a message. In addition, Trust Evaluation System (TES) uses an automated system of highly-proficient, fingerprinting algorithms. Advanced Message Fingerprinting maintains the privacy of the content and reduces the amount of data to be analyzed. Once a message fingerprint is cataloged as spam, all future messages matching that fingerprint are automatically filtered. Because a reputation-based collaborative system does not draw blanket conclusions about terms, hosts, or people, it has proven to increase accuracy, particularly as it relates to false positives and critical false positives, while simultaneously decreasing administration costs.***

Keywords: Online Social networking, Contention modeling, Trust Evaluation System, False positives.

I. INTRODUCTION

Today, there is a continued rise of social networking on the Web. Social networking accounts for 1 of every 6 minutes spent online and as MySpace declines, LinkedIn, Twitter and Tumblr have grown at impressive rates [1]. Social media are becoming increasingly important to recruiters and jobseekers alike. In Online Social Networks (OSNs), there is the possibility of posting or commenting unwanted messages on particular public/private areas, called in general "walls". Unnecessary posts could spam a user's wall, thus disabling the user from viewing relevant messages. Information filtering can therefore be used to give users the ability to automatically control the messages written on their own walls, by filtering out unwanted messages.

OSNs today do not provide much support to prevent unwanted messages on user walls. For example, Facebook allows users to state who is allowed to insert messages in their walls (i.e., friends, friends of friends, or defined groups of friends). Existing Filters as browser extensions and add-ons are: Spoiler Shield app to filter feeds from both Facebook and twitter, for iPhone and Open Tweet Filter which is a filter for twitter on Chrome. These filters only take keywords and filter out messages that contain the specific words. Applications that are trying to solve this issue through machine learning techniques are either in the beta phase or do not perform efficiently due to poor learning curves used to analyze the messages.

Most of the work related to text filtering by Machine Learning has been applied for long-form text. Wall messages are constituted by short text for which traditional classification methods have serious limitations since short texts do not provide sufficient word occurrences. Thus, a suitable text representation method is proposed in this paper along with a neural-network based classification algorithm [3] to classify each message as neutral or non-neutral, based on its content.

Besides classification facilities, Filtering Rules (FRs) can support a variety of different filtering criteria that can be combined and customized according to the user needs. More precisely, FRs exploit user profiles, user relationships as well as the output of the classification process to state the filtering criteria to be enforced. In addition, the system provides the support for user-defined BlackLists (BLs), that is, lists of users that are temporarily prevented to post any kind of messages on a user wall.

This system is intended to be a software application, that is, an add-on, for any social networking service that allows users to post messages. The social networking application is itself developed first with minimalistic features, the most important being posting short-text messages. This is to emulate the behavior of existing OSNs like Facebook and Twitter. Also, a database of users and relationships between the users is maintained. Thus, Post Filter acts as an intelligent software on top of existing OSNs, providing users with a view of clean (non-vulgar and non-offensive) or relevant posts.

II. LITERATURE REVIEW

The main contribution of this paper is the design of a system providing customizable content-based message filtering for OSNs, based on ML techniques. As we have pointed out in the introduction, to the best of our knowledge we are the first proposing such kind of application for OSNs. However, our work has relationships both with the state of the art in content-based filtering, as well as with the field of policy-based personalization for OSNs and, more in general, web contents. Therefore, in what follows, we survey the literature in both these fields.

A. *Content-based filtering*

Information filtering systems are designed to classify a stream of dynamically generated information dispatched asynchronously by an information producer and present to the user those information that are likely to satisfy his/her requirements [3]. In content-based filtering each user is assumed to operate independently. As a result, a content-based filtering system selects information items based on the correlation between the content of the items and the user preferences as opposed to a collaborative filtering system that chooses items based on the correlation between people with similar preferences. Documents processed in content-based filtering are mostly textual in nature and this makes content-based filtering close to text classification. The activity of filtering can be modeled, in fact, as a case of single label, binary classification, partitioning incoming documents into relevant and non-relevant categories [4]. More complex filtering systems include multi-label text categorization automatically labeling messages into partial thematic categories. Content-based filtering is mainly based on the use of the ML paradigm according to which a classifier is automatically induced by learning from a set of pre-classified examples. A remarkable variety of related work has recently appeared, which differ for the adopted feature extraction methods, model learning, and collection of samples [5], [6], [7], [8],[9]. The feature extraction procedure maps text into a compact representation of its content and is uniformly applied to training and generalization phases. The application of content-based filtering on messages posted on OSN user walls poses additional challenges given the short length of these messages other than the wide range of topics that can be discussed. Short text classification has received up to now few attentions in the scientific community. Recent work highlights difficulties in defining robust features, essentially due to the fact that the description of the short text is concise, with many misspellings, nonstandard terms and noise. Focusing on the OSN domain, interest in access control and privacy protection is quite recent. As far as privacy is concerned, current work is mainly focusing on privacy-preserving data mining techniques, that is, protecting information related to the network, i.e., relationships/nodes, while performing social network analysis [5]. Works more related to our proposals are those in the field of access control. In this field, many different access control models and related mechanisms have been proposed so far (e.g., [6,2,10]), which mainly differ on the expressivity of the access control policy language and on the way access control is enforced (e.g., centralized vs. decentralized). Most of these models express access control requirements in terms of relationships that the requestor should have with the resource owner. We use a similar idea to identify the users to which a filtering rule applies. However, the overall goal of our proposal is completely different, since we mainly deal with filtering of unwanted contents rather than with access control. As such, one of the key ingredients of our system is the availability of a description for the message contents to be exploited by the filtering mechanism as well as by the language to express filtering rules. In contrast, no one of the access control models previously cited exploit the content of the resources to enforce access control. We believe that this is a fundamental difference. Moreover, the notion of black-lists and their management are not considered by any of these access control models

B. *Policy-based personalization of OSN contents*

Recently, there have been some proposals exploiting classification mechanisms for personalizing access in OSNs. For instance, in [11] a classification method has been proposed to categorize short text messages in order to avoid overwhelming users of microblogging services by raw data. The system described in [11] focuses on Twitter2 and associates a set of categories with each tweet describing its content. The user can then view only certain types of tweets based on his/her interests. In contrast, Golbeck and Kuter [12] propose an application, called Film Trust that exploits OSN trust relationships and provenance information to personalize access to the website. However, such systems do not provide a filtering policy layer by which the user can exploit the result of the classification process to decide how and to which extent filtering out unwanted information [15]. In contrast, our filtering policy language allows the setting of FRs according to a variety of criteria that do not consider only the results of the classification process but also the relationships of the wall owner with other OSN users as well as information on the user profile. Moreover, our system is complemented by a flexible mechanism for BL management that provides a further opportunity of customization to the filtering procedure. The only social networking service we are aware of providing filtering abilities to its users is MyWOT, a social networking service which gives its subscribers the ability to: 1) rate resources with respect to four criteria: trustworthiness, vendor

reliability, privacy, and child safety; 2) specify preferences determining whether the browser should block access to a given resource, or should simply return a warning message on the basis of the specified rating. Despite the existence of some similarities, the approach adopted by MyWOT is quite different from ours. In particular, it supports filtering criteria which are far less flexible than the ones of Filtered Wall since they are only based on the four above-mentioned criteria. Moreover, no automatic classification mechanism is provided to the end user. Our work is also inspired by the many access control models and related policy languages and enforcement mechanisms that have been proposed so far for OSNs, since filtering shares several similarities with access control. Actually, content filtering can be considered as an extension of access control, since it can be used both to protect objects from unauthorized subjects, and subjects from inappropriate objects. In the field of OSNs, the majority of access control models proposed so far enforce topology-based access control, according to which access control requirements are expressed in terms of relationships that the requester should have with the resource owner. We use a similar idea to identify the users to which a FR applies. However, our filtering policy language extends the languages proposed for access control policy specification in OSNs to cope with the extended requirements of the filtering domain. Indeed, since we are dealing with filtering of unwanted contents rather than with access control, one of the key ingredients of our system is the availability of a description for the message contents to be exploited by the filtering mechanism.

III. TRUST EVALUATION SYSTEM (TES)

TES is the reputation metric, or trust system, that evaluates every new piece of feedback submitted to the Nomination servers. The primary function of TeS is to assign a "confidence" to fingerprints—a value between Cmn (legitimate) and Cmx (spam), based on the "reputation" or "trust level" of the individual reporting the fingerprint. The trust level, t, is a finite numeric value attached to every community reporter. The value t is, in turn, computed from the corroborated historical confidence of the fingerprints nominated by the reporter. The circular assignment effectively turns the classifier into a stable closed-loop control system.

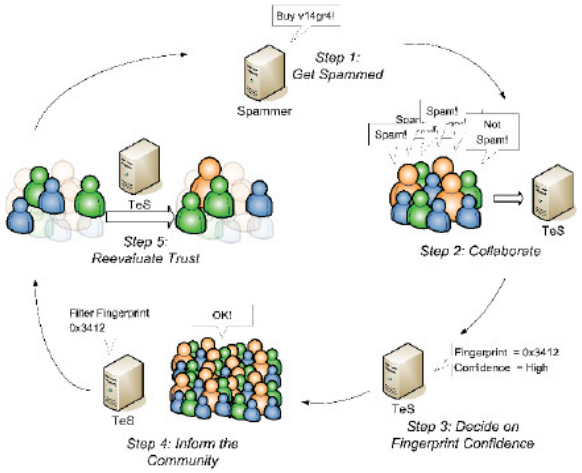


Figure 1: Process Flow of the Trust Evaluation System

TES determines both the confidence the community has in the disposition of a fingerprint and the trust that the system places in the decisions made by members of the community. In a continuous process, members of the community receive new spam (1) and report their feelings ("spam" or "not spam") about the message to the Nomination server, which in turn reports it to TES (2). Based upon the trust associated with each individual reporter, TES assigns confidence to the fingerprint (3) and reports it to the service for distribution to the community (4). TES then reevaluates the community trust values to determine who should gain and lose trust as a result of their individual assessments of the message.

Just as in the real world, trust is earned slowly and is difficult to attain. New recipients start with a trust level of zero. In the very beginning (at the launch of the classifier), there were only a few hand-picked recipients with a high trust level. As zero-valued, untrusted community members provide feedback, TES rewards reporters whose feedback agrees with those of highly-trusted, highly-reputable members of the community. In other words, TES assigns trust points to recipients when their reports are

corroborated by other highly trusted recipients[18]. In practice, for every fingerprint that achieves a high confidence meaning that the fingerprint was reported and corroborated by highly trusted recipients.TES gives one of first reporters of the fingerprint a small trust reward.

Untrusted recipients who report often and report correctly eventually accrue enough trust rewards to become trusted recipients themselves. Once trusted, they implicitly begin to participate in the process of selecting newer trusted recipients. In this manner, TES selectively inducts a community of "highly-reputable," "highly-trusted" members—reporters who routinely make decisions that are honored by the rest of the community. TES also penalizes recipients who disagree with the trusted majority. Penalties are harsher than rewards, so while gaining trust is hard, losing it is rather easy.

The second aspect of TES's responsibility is to assign confidence to fingerprints. Fingerprint confidence is a function of the reporter's trust level and the disposition (block/unblock) of their reports. TES updates confidence in real-time with every report. Once the confidence reaches a threshold, known as average spam confidence, it is promoted to Catalog servers. If a promoted fingerprint is unblocked by trusted recipients, its confidence can drop below the average spam confidence, which results in its immediate removal from the catalog servers. The real-time nature of confidence assignments results in an extremely responsive system that can self correct within seconds.

In more formal terms, a given set of fingerprint reporters, R, who each have a trust level tr , send in reports that have a disposition dr , where dr = -1 if the fingerprint misclassifies a legitimate message as spam and tr = 1 if the message is spam. After a number of fingerprints are collected, it is possible to compute a fingerprint confidence using the following equation: However, it is important to note that TeS uses a variation of the above algorithm to reduce its attack vulnerability.

A. *Emergent Properties of TES*

TES has several desirable, even surprising, emergent properties when deployed on a large scale. These properties are critical to the effectiveness of the system and typical of well-designed reputation metrics. We discuss some of these properties in this section and contrast them with related properties of other anti-spam approaches.

$$confidence = \min\left\{c_{mc}, \max\left\{c_{mv}, \sum_{r \in R} d_r \max\{t_r, 0\}\right\}\right\}$$

1) *Responsiveness:* TES's reward selection metric prefers those recipients who report correctly and early. This means that over time TES can identify all such reporters whose initial reports have a high likelihood of being accepted as spam by the rest of the community. As the group of trusted recipients becomes larger, the first few reports are extremely reliable predictors of a fingerprint's final disposition. As a result, TES can respond extremely quickly to new spam attacks. Anti-spam methods that either require expert supervision, or that are inherently unable to train on individual samples, have significantly longer response latencies. These systems are unable to stop short-lived attacks that are not already addressable by their existing filtering hypotheses.

2) *Self-Correction:* The ability to make negative assertions ("Message is not spam"), combined with the dynamic nature of the confidence assignment algorithm, permits speedy self-correction when the initial prediction is incompatible with the consensus view. Since confidence and trust assignments are intertwined, community disagreement results in immediate correction of the confidence of fingerprints, as well as a trust reduction for reporting the fingerprints as spam. This results in a historical trend toward accuracy because only the reporters who consistently make decisions aligned with the consensus retain their trusted status. From a learning perspective, the reporter's reputation or trust values represent the entire history of good decisions and mistakes made by the classifier.

3) *Modeling Disagreement:* One of things we learned almost immediately after the launch of TES was that certain fingerprints would wildly flip-flop across the average spam confidence level. These fingerprints usually represented newsletters and mass

mailings that were considered desirable by some and undesirable by others. The community of trusted recipients disagreed on the disposition of these fingerprints because there was no "real" community consensus on whether or not the message was spam. By modeling the pattern of disagreement, we taught TES to identify this kind of disagreement and flag such fingerprints as contested. When agents query contested fingerprints, they are informed of the contention status so they can classify the source emails based on out-of-band criteria, which can be defined subjectively for all recipients. Contention modeling is extremely important for a collaborative classifier because it scopes the precision of the system. If the limitations of the classifier are known, other classification methods can be invoked as required. In the TES contention logic is also a catch-all defense against fingerprint collision. If a set of spam and legitimate email happen to generate the same fingerprint, the fingerprint is flagged as contested, which excludes its disposition from the classification decision. Historically aggregated contention rates in the service are an indicator of the level of disagreement in the trusted community. The level of disagreement in the service is very low, which implies that the trust model can successfully represent the collective wisdom of the community. Most machine learning systems, including statistical text classifiers like Naive Bayes, are unable to automatically identify contested documents. This is why statistical classifiers tend to work better in single-user environments where recipient preferences are consistent over time

4) *Resistance to Attack:* An open, user feedback-driven system is an attractive attack target for spammers. There are essentially two ways of attacking the service. One is through a technique called hash busting, which attacks fingerprinting algorithms by forcing them to generate different fingerprints for each mutation of a spam message.. The second vector for attack is through incorrect feedback. Most commonly, attackers attempt to unblock their mailings before broadcasting them to the general population. However, an attacker must first be considered trusted in order to affect the disposition of a fingerprint. In order to gain trust, the attacker must provide useful feedback over a long period of time, which requires blocking spam that others consider to be spam. In other words, spammers must behave like good recipients for an extended period of time to get even a single identity to be considered trusted. If they do spend the effort building up a trusted identity, the amount of damage they can do with one or few trusted identities is negligible because the disagreement from the majority of the trust community will result in harsh trust penalties for the spammer identities. As the pool of trusted users grows, it gets harder to gain trust and easier to lose it. Participation is proportional to the strength of attack resistance. Expert-supervised systems are resistant to such attacks by definition but are unable to scale to a large number of experts. Similarly, statistical text classification systems must go through supervised training to avoid corpus pollution. Supervision limits the amount of training data that can be considered. One real-world limitation, for example, is in a supervised classification system's inability to adequately detect "foreign language" spam—that is, spam in languages not understood by supervisors.

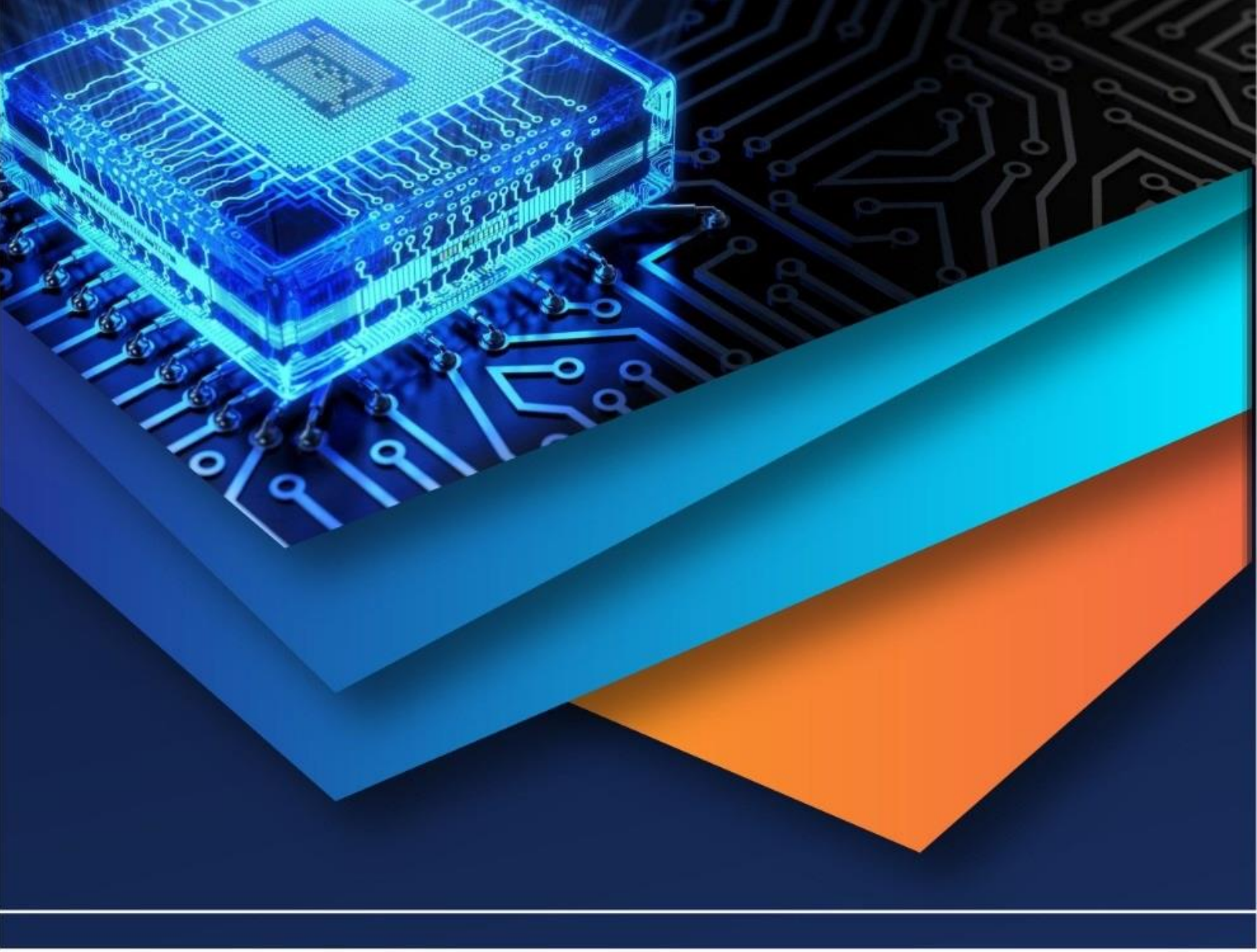
IV. CONCLUSION

In this paper, we have proposed an approach TES ensures the "reputation" of reporters by tracking how often the larger recipient community agrees with their assessment of a message. In addition, Trust Evaluation System(TES) uses an automated system of highly-proficient, fingerprinting algorithms. Advanced Message Fingerprinting maintain the privacy of the content and reduce the amount of data to be analyzed.we improve the global performance of the TES model to filter out unwanted messages from Online Social Networking (OSN).

REFERENCES

[1] Mr.Md.Amanatulla, J. Alcalá-Fdez, F. Herrera, J. Otero, Genetic learning of accurate and compact fuzzy rule based systems based on the 2-tuples linguistic representation, International Journal of Approximate Reasoning 44 (2007) 4564.
[2] A. Asuncion, D. Newman, 2007. UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences. URL: <http://www.ics.uci.edu/~mlearn/MLRepository.html>.
[3] R. Barandela, J.S. Sánchez, V. García, E. Rangel, Strategies for learning in class imbalance problems, Pattern Recognition 36 (3) (2003) 849–851.
[4] G.E.A.P.A. Batista, R.C. Prati, M.C. Monard, A study of the behaviour of several methods for balancing machine learning training data, SIGKDD Explorations 6 (1) (2004) 20–29.
[5] P. Campadelli, E. Casiraghi, G. Valentini, Support vector machines for candidate nodules classification, Letters on Neurocomputing 68 (2005) 281–288.
[6] J.R. Cano, F. Herrera, M. Lozano, Using evolutionary algorithms as instance selection for data reduction in kdd: an experimental study, IEEE Transactions on Evolutionary Computation 7 (6) (2003) 561–575.

[7] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority over-sampling technique, Journal of Artificial Intelligent Research 16 (2002) 321–357.
[8] N.V. Chawla, N. Japkowicz, A. Kolcz, Editorial: special issue on learning from imbalanced data-sets, SIGKDD Explorations 6 (1) (2004) 1–6.
[9] Z. Chi, H. Yan, T. Pham, Fuzzy algorithms with applications to image processing and pattern recognition, World Scientific, 1996.
[10] J.-N. Choi, S.-K. Oh, W. Pedrycz, Structural and parametric design of fuzzy inference systems using hierarchical fair competition-based parallel genetic algorithms and information granulation, International Journal of Approximate Reasoning 49 (3) (2008) 631–648.
[11] D. Nelms, "Social networking growth stats and patterns", http://socialmediatoday.com/amzini/306252/social-networking-growth-stats-and-patterns
[12] M. Vanetti, E. Binaghi, E. Ferrari, B. Carminati, M. Carullo, "A System to Filter Unwanted Messages from OSN User Walls", IEEE Transactions on Knowledge and Data Engineering, Vol:25, June 2013
[13] F. Sebastiani, "Machine learning in automated text categorization," ACM Computing Surveys, vol. 34, no. 1, pp. 1–47, 2002
[14] C. D. Manning, P. Raghavan, and H. Schutze, Introduction to Information Retrieval. Cambridge, UK: Cambridge UniversityPress, 2008.
[15] Wikipedia page on "Multilayer Perceptron", http://en.wikipedia.org/wiki/Multilayer_perceptron
[16] Wikipedia page about "Deep Learning", http://en.wikipedia.org/wiki/Deep_learning
[17] Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", Second Edition
[18] Wikipedia page on "Precision and recall", http://en.wikipedia.org/wiki/Precision_and_recall.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY