iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** X **Month of publication:** October 2017

DOI: http://doi.org/10.22214/ijraset.2017.10268

www.ijraset.com

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

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue X, October 2017- Available at www.ijraset.com

Extraction and Analysis of Social Network Data Using Text Mining Techniques

Hayath TM.[1],Naveen Kumar G[1]
[1, 2] *Ballari Institute of Technology and Management,Ballary, Karnataka, India*

Abstract: ***Social network websites provide a means of communication between people who are located in different locations and this can be done by establishing a network in which the information such as text messages, pictures, audio and videos can be shared. The information that is stored in these websites will be in unstructured manner and hence it may lead to ambiguity such as lexical, semantic, syntax of data. Moreover the data set that is generated from the information pattern is more complex and difficult to analyze. Another problem with these kinds of websites is that there may be a chance of fraudulent activities carried out, such as creating fake profiles. The proposed technique implements pre-processing (tokenization) which removes irrelevant or redundant data which are stored in unstructured manner and then the hybrid classifier is used to detect the fraudulent activities carried out in social networking websites which in turn increases the performance of the system.***
Keyword: Text Mining, Classification, Social Network. Extraction

I. INTRODUCTION

Social networking websites provides an easy way for two way communication. Face book is one type of such network that is widely used by people for communication which is rich in texts that enable the user to create various text contents in the form of comments, wall posts, social media, and blogs. Due to the ubiquitous use of social networks in recent years, an enormous amount of data is available over the web. Text mining is used by most of the application in social networking websites that provides appropriate result for person-to-person interaction. Moreover, text mining techniques in conjunction with social networks can be used for finding a general opinion about any specific subject such as human thinking patterns and group identification in large-scale systems. In most of the social networking websites people use formal language for communication. Since informal communication is been carried out between different people and also it differs from person to person so there is a chance of different kinds of ambiguities are occurred such as lexical, syntactic, and semantic [1] of data. Therefore, forming a standard acceptable way for convenient communication is a critical task and can be solved by text mining.Text mining is an interactive technique that enhances computational intelligence which comprises of multidisciplinary fields. In information retrieval, text analysis, natural language processing the information is classification based on logical and non-trivial patterns from large data sets and many authors defines that text mining is also part of data mining technique. Data mining techniques are mainly used for the extraction of logical patterns from the structured database. Text mining techniques are very complex than data mining due to unstructured and fuzzy nature of natural language text. Most of researchers till today use decision trees and hierarchical clustering for group recommendations in facebook where the user can join the group based on similar atterns in user profiles.

II. RELATED WORK

Social networking websites provides a very powerful medium for communication among individuals in order to share valuable knowledge. Ex: Facebook, LinkedIn and MySpace, etc .In these websites generally people use unstructured or semi structured language for communication and this leads to ambiguities such as lexical, syntactic and semantics in the data[1]. In order to extract logical pattern with accurate information in very difficult from unstructured format, this is a problem in critical tasks. The solution to this problem is text mining. Text mining is a knowledge discovery technique that provides high quality information from text mining pattern. Text mining can be classified as

A. Machine Learning Based Text Classification

Machine learning techniques can be classifies as supervised, unsupervised and semi-supervised method. This section mainly focuses on supervised classification technique.

1) *Rocchio's Algorithm:* This algorithm is implemented based on feedback method which provides the synonymy for different words with similar meanings in natural language [2]. However, the problem with this technique is that the user must have sufficient knowledge in order to indicate relevance feedback and this may not work efficiently when the user spells a word in a

different wayThis algorithm uses a vector space method for document routing or filtering in informational retrieval, to build prototype vector for each class using a training set of documents and to calculate the similarity between test documents, prototype vectors which is used to assign test document for relevant and non relevant documents to the class with maximum similarity.

$$C_i = \alpha * centroid_{C_i} - \beta * centroid_{\overline{C_i}}$$

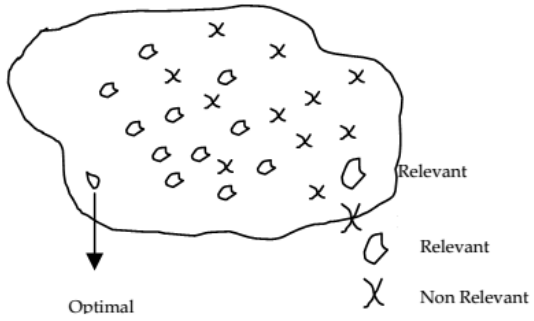


Figure 2.1 Rocchio Optimal query for separating relevant and non relevant documents

2) *Instant Based Learning Algorithm:* This algorithm works based on the comparison between new problem instances and already stored instances during training [3]. Once the new instances is been arrived and then the related sets of instances are retrieved from the memory for further processing such that new instances can be classified accordingly. An example of instant based learning is case based reasoning (CBR) and k-nearest neighbour (K-NN) algorithm. K-NN identifies the closest feature space by measuring the angle between the two feature vectors (or) by calculating the distance between the vectors.CBR uses TCBR to deal with textual knowledge source in making decision. However this approach extracts similar cases and represents the knowledge without losing the key concepts with low knowledge, which is a drawback of this method. It calculates the similarity between test document and each neighbour and assigns the test document to the class which contains most of the neighbours which is depicted in below figure.

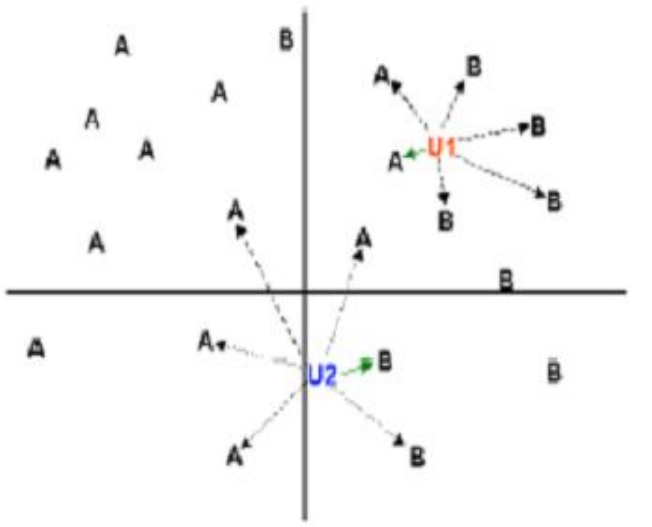


Figure 2.2 k-Nearest Neighbour Algorithm

3) *Decision Rules Classification Method:* This algorithm does text classification which depends on large number of sufficient and insufficient number of relevant features in decision tree which may lead to poor performance in text classification [4].

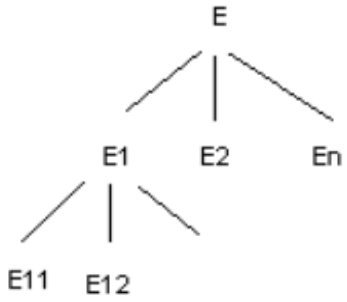


Figure 2.3 Decision Tree

The above figure represents decision tree which shows different entities and their level. It is impossible to assign a document to a category exclusively due to the rules from different rule a set that is applicable to each other and it's a major drawback. Besides, the learning and updating of decision rule methods need extensive involvement of human experts to construct or update the rule sets. The decision rules method does not work well when the number of distinguishing features is large.

4) *Genetic Algorithm:* This algorithm is used to make feature selection and termed weight method for assigning weights to each concept in the document based on the relevant topics. This algorithm provides better results and is widely used to solve optimization problems. As the process is recursive, an end function needs to be specified based on monitoring the improvement of results in the consecutive generations [5]

5) *Support Vector Machine (SVM):* A SVM algorithm use both negative and positive training data sets in order to constructs a hyper plane that separates the positive and negative data. Here the document that is closest to decision surface is called support vector and it is based on structure risk minimization principle which is used to find a hypothesis to guarantee the lowest true errors. SVM needs both positive and negative training set which are uncommon for other classification methods. This method is used to seek the decision surface which separates the positive data set from the negative data set in the 'n' dimensions space and it is called as hyper plane. The document which is closest to the decision surface are called as support vector and the performance of SVM classification remains unchanged even if the document does not belongs to the support vectors which are moved from the set of training data[6] as shown in figure 2.4

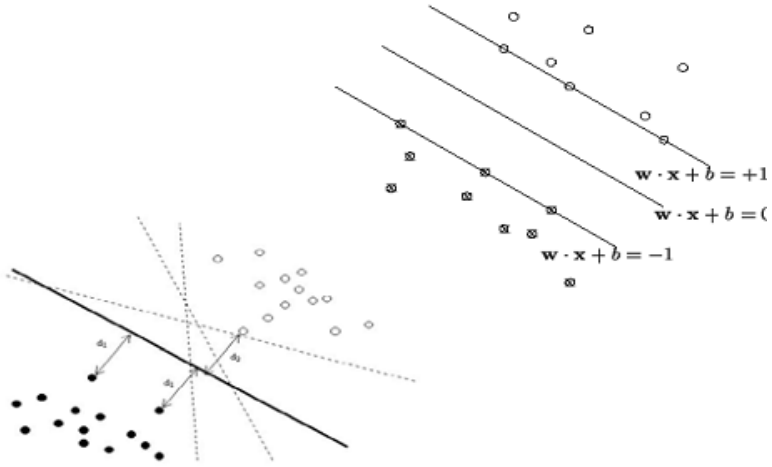


Figure 2.4 Illustration of optimal separating hyper plane, hyper planes and support vectors

6) *Artificial Neural Networks:* Artificial neural networks are constructed from a large number of elements with an input fan order of magnitudes larger than in computational elements of traditional architectures. These elements, namely artificial neuron are interconnected into group using a mathematical model for information processing based on a connectionist approach to computation. The neural networks make their neuron sensitive to store item. It can be used for distortion tolerant storing of a large number of cases represented by high dimensional vectors. Different types of neural network approaches have been implemented to document classification tasks. Some of the researches use the single-layer perceptron, which contains only an

input layer and an output layer due to its simplicity of implementing. Inputs are fed directly to the outputs via a series of weights. In this way it can be considered the simplest kind of feed-forward network. The multi-layer perceptron which is more sophisticated, which consists of an input layer, one or more hidden layers, and an output layer in its structure, also widely implemented for classification tasks.

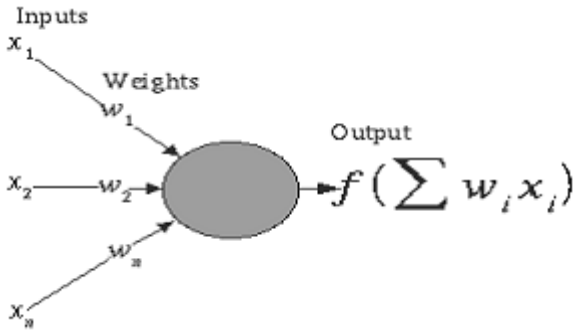


Figure 2.5 Artificial Neural Network

The main advantage of the implementation of artificial neural network in classification tasks is the ability in handling documents with high-dimensional features, and documents with noisy and contradictory data. The drawback of the artificial neural networks is their high computing cost which consumes high CPU and physical memory usage. Another disadvantage is that the artificial neural networks are extremely difficult to understand for average users. [7].

III. DATA AND METHODOLOGY

A. *Existing System*

The existing system works based on various classification algorithms have been used for text classification and analysis. Several surveys shows that the hybrid classification algorithms provides better results and increases text categorization accuracy instead of applying individual method (single method)[8] .There is no guarantee that a high level of accuracy acquired by one test set will also be obtained in another test set.

B. *Draw Backs of Existing System*

1) Single method provides poor performance
2) Pre-processing has been carried out but, accuracy is very low
3) Analysis is more complex if data is large in size
4) Generate garbage in garbage out phenomenon.

C. *Proposed System*

The proposed system is been implemented by combining pre-processing technique and hybrid classifier. Pre- processing is been done in first phase which takes an unstructured stream of text and generates tokens. In the second phase hybrid classifier is applied on the generated token to convert textual data to numerical data. Further the obtained numerical data is given as input to SVM which detects fraudulent activities in order to achieve better performance.

IV. IMPLEMENTATION DETAILS

The implementation stage is a system project involves careful planning, investigation of the current system and its constraints on implementation design of the methods to achieve change over etc. Since the package is found to be free from errors and the data, variables are sent correctly from one module to another for implemented the package. The errors in the code will be rectified during the phases of testing.

A. *About Tools*

1) *LIBSVM:* LIBSVM is a popular open source machine learning library developed by National Taiwan University. It implements SMO algorithm for SVMs, Supporting Classification and Regression. LIBSVM provides a simple interface where the user can easily link it with this program[8]. The features of LIBSVM are as follows:

a) Different SVM formulations
b) Efficient multi class classification cross validation for model selection
c) Probability estimates
d) Weighted SVM for unbalanced data

B. Major Modules

1) *Tokenization*: Tokenization is one of the best techniques which are used to extract redundant and noisy data in text mining. This technique is applied on raw text data which is extracted from social networking websites such as facebook, which generally will be in unstructured manner. This unstructured data is fed as input to the pre processing process which splits the sequence of characters into tokens by performing lexical analysis.

2) *Classifiers (ANN & SVM):* In this module classifiers are been used to detect fake profiles and fraudulent usage carried out in the network. The extracted tokens obtained after pre processing is been supplied to ANN classifier. A network model is created which consists of three different layers namely input layer, hidden layer and output layer. Later it checks for the fraud words present in the tokens and count the number of fraud words. Once the train file is been loaded three file will be automatically generated i.e. text file, model file and output file path for a specific file. SVM is been applied on these files to check the accuracy of the system.

V. RESULTS AND DISCUSSIONS

The below figure 5.1 shows the login page, where the admin is allowed to enter his username and password, then the login button is clicked. The authentication of the user is been done by connecting to the database in order to check whether the user is authorised or not and if the username or password is valid then it logins successfully

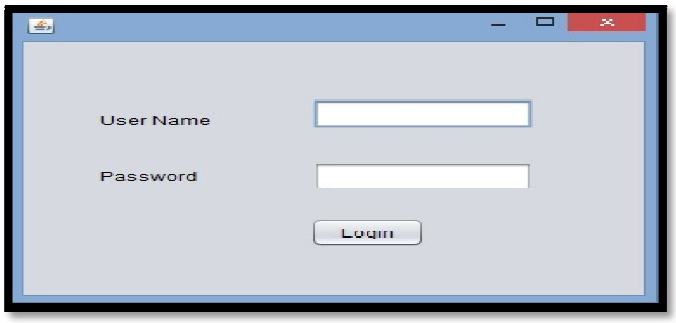


Fig.5.1 Log in page

If the username or password is not valid then a dialog box will be displayed stating invalid username and password as shown in figure 5.12

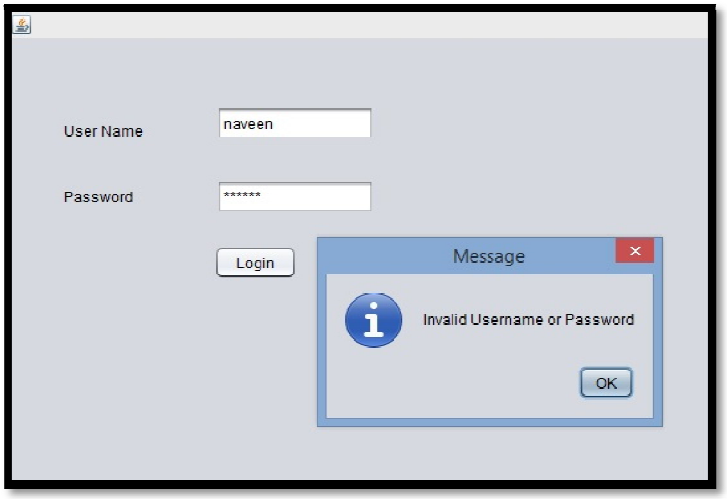


Fig.5.2 Log in page when wrong password entered

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue X, October 2017- Available at www.ijraset.com

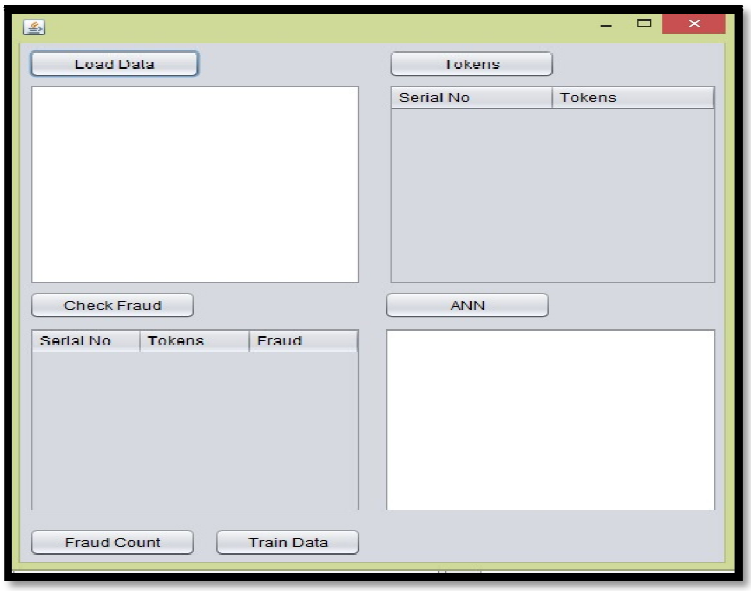


Fig.5.3 Overall process of extraction and analysis

The above figure 5.3 shows the overall process of extraction and analysis which includes

A. Loading the data
B. Generate tokens
C. Check fraud profiles
D. ANN and SVM

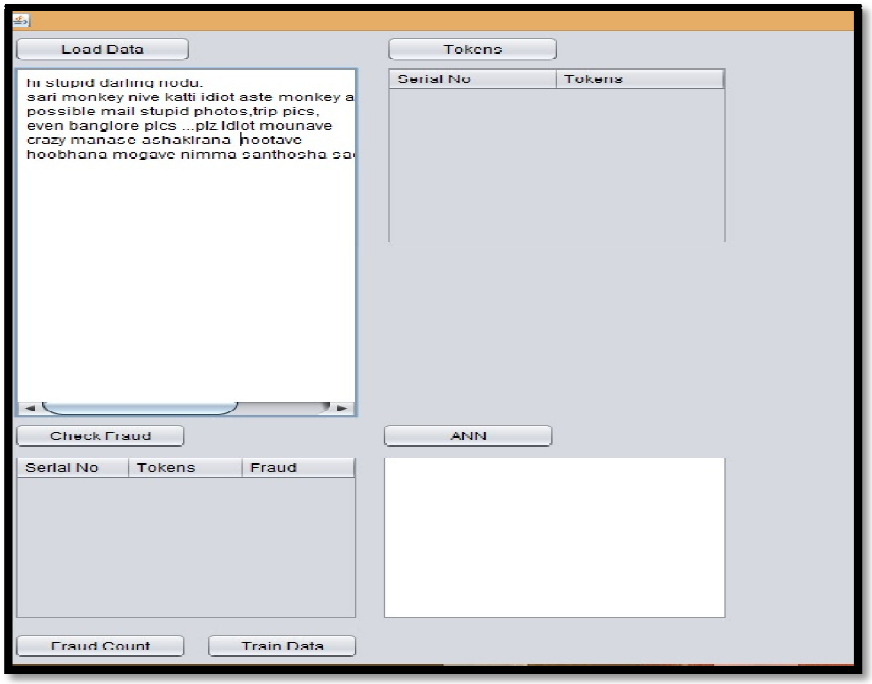


Fig.5.4 After load the user text data

above figure 5.4 shows the data is be loaded by clicking on the load data button from the untokenized profiles. The untokenized profiles consits of raw data stored in unstrctured manner which also contains irrelevant or redundent data .

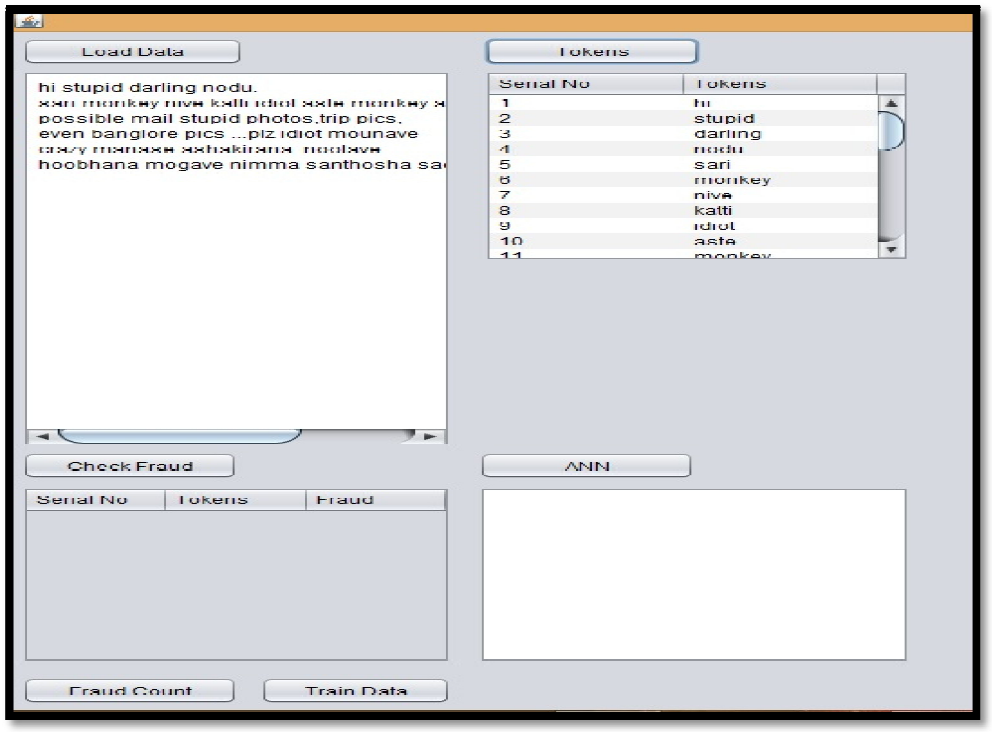


Fig .5.5 Text message in token format

Once the token button is clicked it generates tokens which are in the form of structured data. This is done by dividing individual words from entire sentence into tokens as shown in figure 5.5

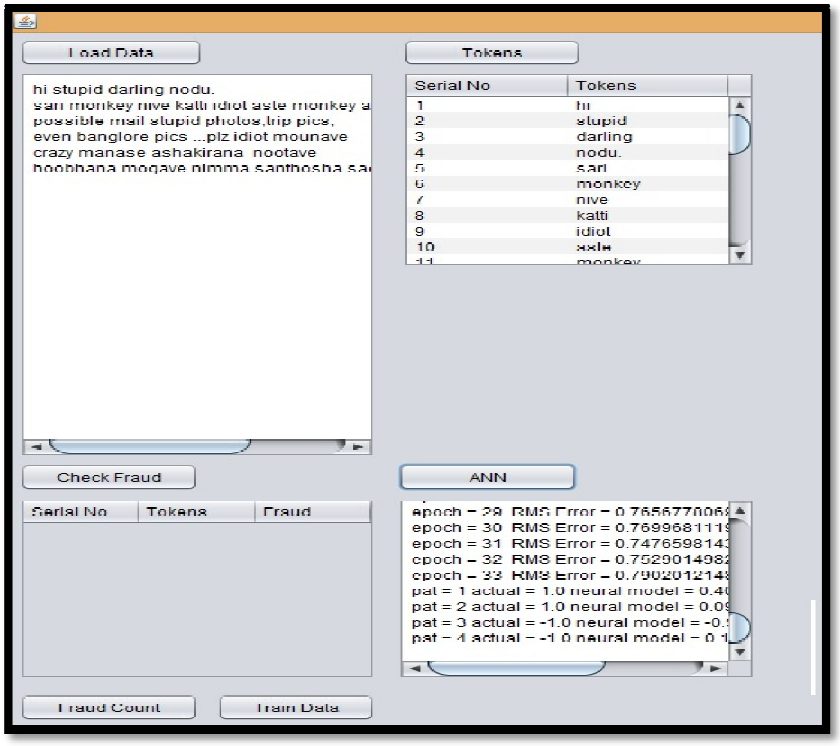


Fig.5.6 After applied ANN to token

Once the ANN button is been clicked a network model will be generated which consisting of input layer, hidden layer and output layer.

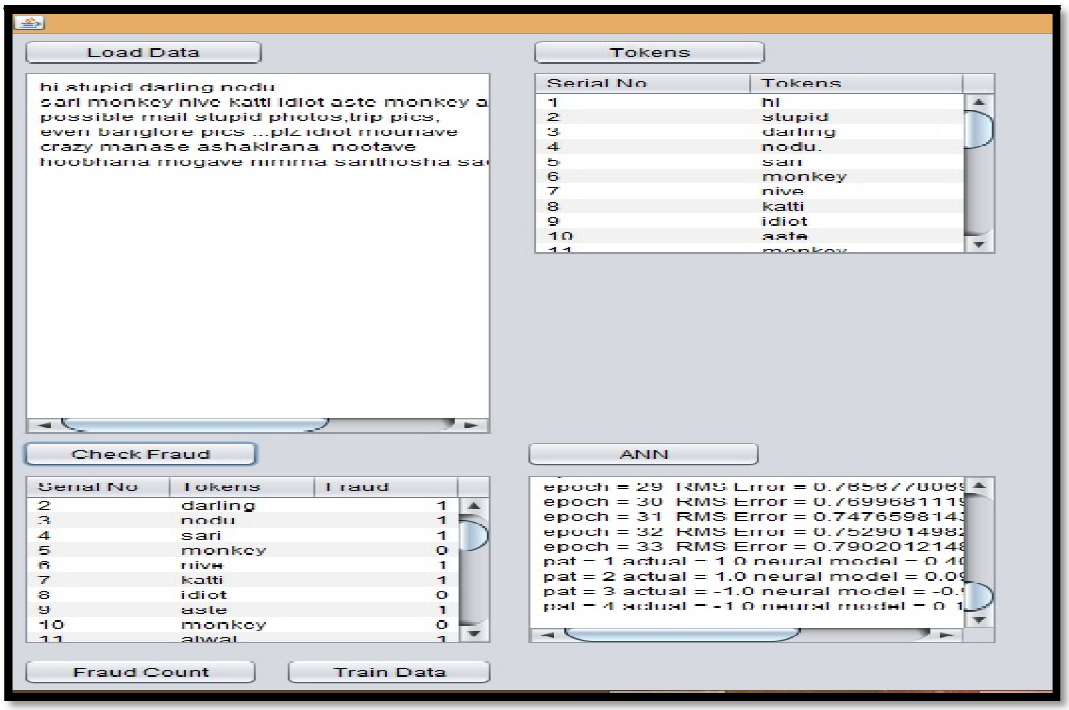


Fig.5.7 Check fraudulent words

Once the check fraud button is been clicked, the outcome generated will be in terms of either 1 or 0 which indicate good words or fraud words respectively as shown in the figure 5.7

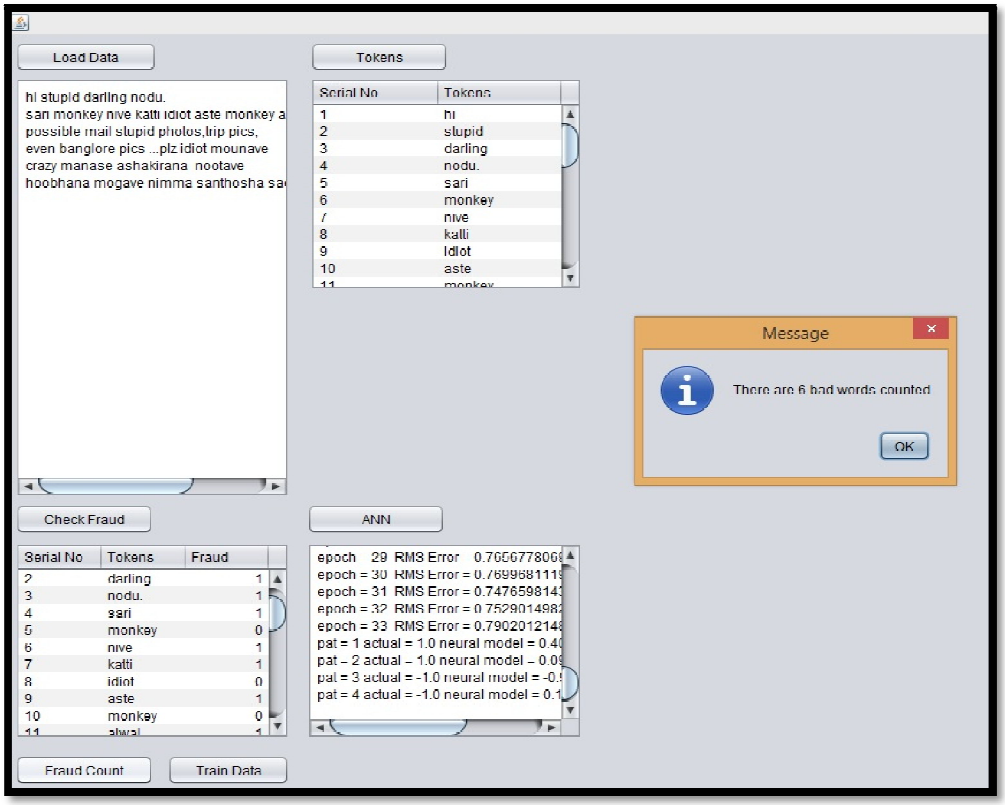


Fig.5.8umber of fraud words in user text data

The above figure 5.8 shows the count of fraud words which in turn compares with the customized tokens and if there is a match of fraud words the count increases. The number of fraud words count is displayed on the new dialog box.

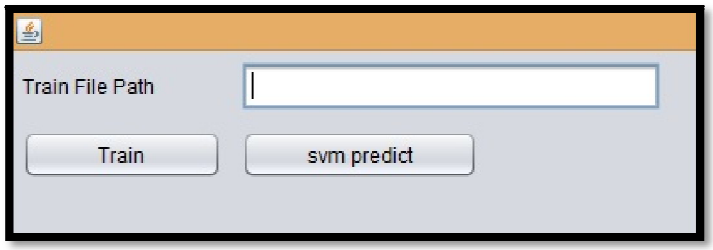


Fig.5.9 Trained file path for SVM

The above figure 5.9 shows a text box for the admin to enter train file path for SVM prediction.

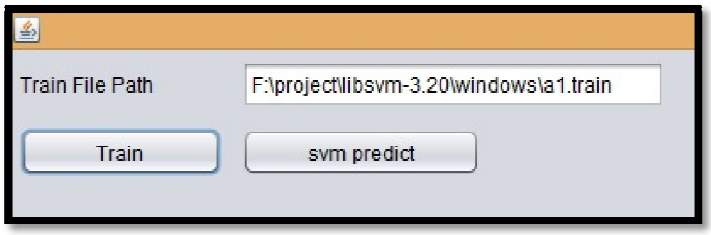


Fig.5.10 After loading trained file path for SVM

once admin enter the train file path and then by clicking the train button, three files will be generated automatically i.e. text file, model file and output file path for a specific file as shown in figure 5.11

Fig.11 Test, model and output files for SVM

Once all the files path has been loaded, test button is clicked to obtain the accuracy of the classifications. If there are 20 data sets and after classifying if we get 20 classified data we get 100% accuracy as shown in figure 5.12

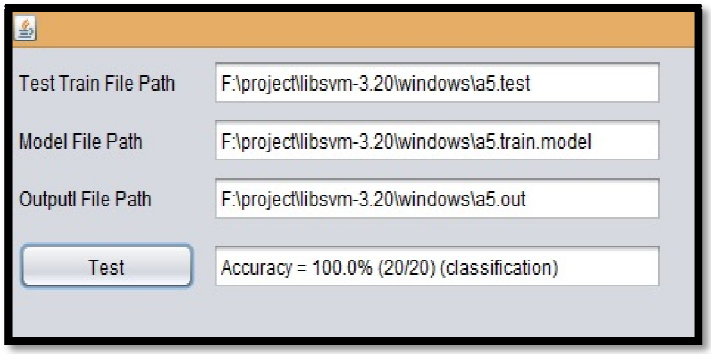


Fig.5.12 Final result of SVM classification accuracy

The below figure 5.13 shows an example of getting 17 classification among 20 data sets and after the test button is been clicked the accuracy is been obtained which is 85%.

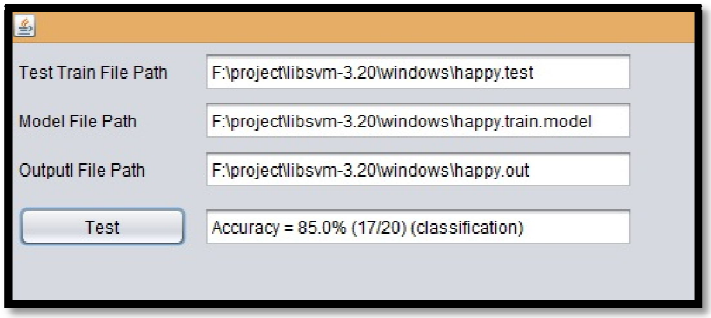


Fig.5.13 Partial accuracy of SVM prediction

Fig.14 Example of fake profiles 1

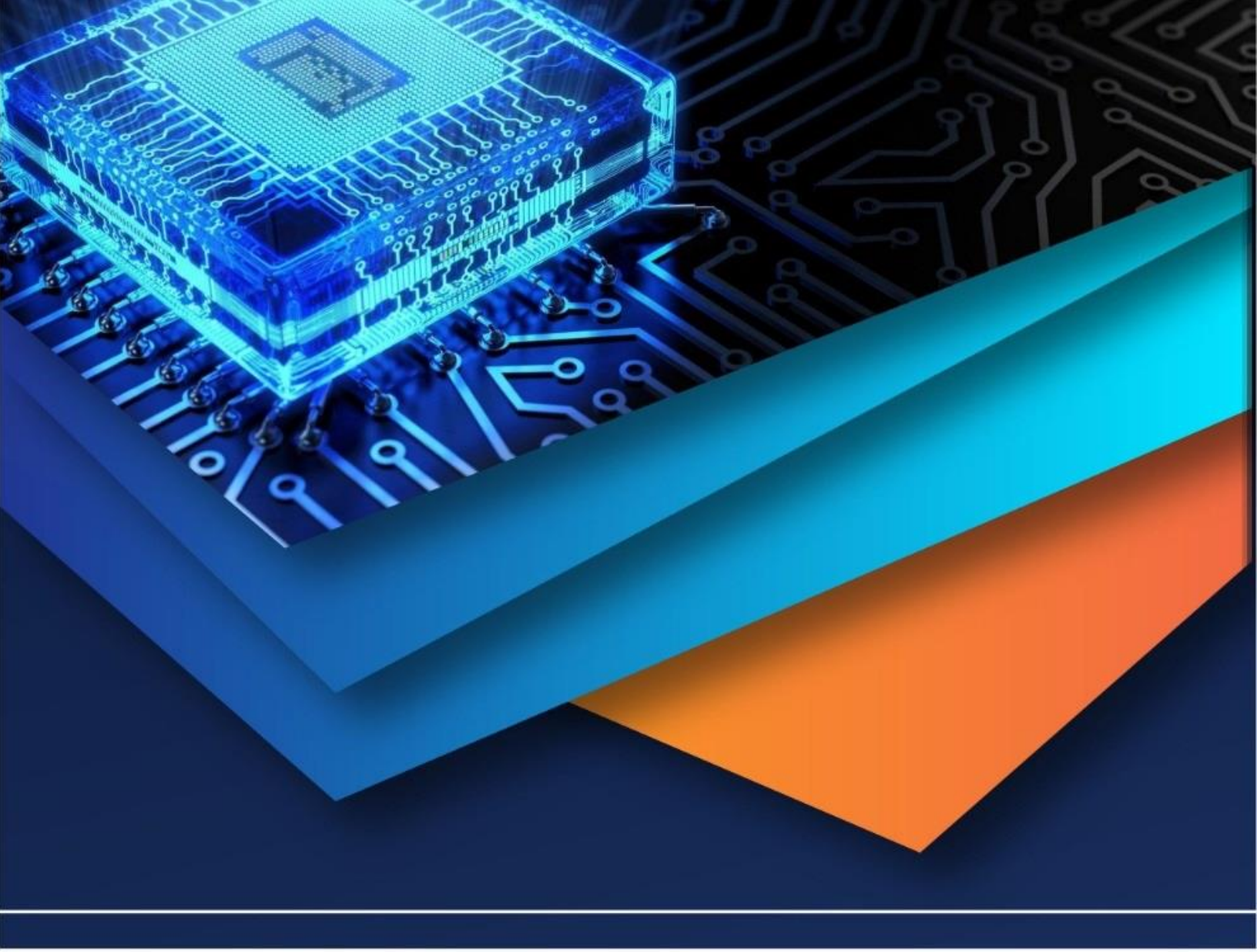
VI. CONCLUSION

Social networking websites contains lot of textual documents in the form of emails, comments, chats and blogs, etc. Many technologies are developed for the extraction of information from the huge collection of textual data using different text mining techniques and the extraction of user data becomes a challenging factor when the textual information is not structured according to the grammatical construction. Tokenization technique is used for extraction of logical patterns from the unstructured and grammatically incorrect textual data and also hybrid classifier is applied on the generated token to convert textual data to numerical data (ANN). Further the obtained numerical data is given as input to SVM which detects fraudulent activities in order to achieve better performance.

REFERENCES

[1] Sorensen, L. 2009. User managed trust in social networking comparing facebook, MySpace and LinkedIn. In Proceedings of 1st International Conference on Wireless Communication, Vehicular Technology, Information Theory andAerospace & Electronic System Technology, (Wireless VITAE 09), Denmark, 427-431.
[2] Han.J, Kamer.M. .Data Mining Concepts and Techniques.BeiJing:Higher education press, 2001. 285-295.
[3] Chang, M. & Poon, C. K. 2009. Using phrases as features in e-mail classification. Journal of System and Software, 82(6).
[4] Forman, G. & Kirshenbaum, E. 2008. Extremely fast text feature extraction for classification and indexing. In Proceedings of 17th ACM Conference on Information and Knowledge Management, California, USA.
[5] Khalessizadeh, S. M., Zaefarian, R., Nasseri, S. H. & Ardil, E. 2006. Genetic mining: Using genetic algorithm for topic based on concept distribution. Journal of Word Academy of Science, Engineering and Technology, 13(2), 144-147.
[6] metaphor in political blogs. In Proceedings of 28th International Conference on Human Factor in ComputingSystems (CHI 2010) ACM, Atlanta, GA, USA, 34-45
[7] Li, J., Wang, H. & Khan, S. U. 2012. A semantics-based approach to large-scale mobile social networking, Mobile Networks and Applications, 17(2), 192-205
[8] https://www.csie.ntu.edu.tw/~cjlin/libsvm/
[9] Forman, G. & Kirshenbaum, E. 2008. Extremely fast text feature extraction for classification and indexing. In Proceedings of 17th ACM Conference on Information and Knowledge Management, California, USA, 26-30.
[10] Zhao, Y. & Dong, J. 2009. Ontology classification for semantic-web-based software engineering. IEEE Transactions on Service Computing, 2(4), 303-317

[11] Brucher, H., Knolmayer, G. & Mittermayer, M. 2002. Document classification methods for organizing explicit knowledge. In Proceedings of 3rd European Conference on Organizational Knowledge, Learning and Capabilities,Athens, 1-25
[12] Durga, A. K. & Govardhan, A. 2011. Ontology based text categorization-telugu document. International Journal of Scientific and Engineering Research, 2(9), 1-4
[13] , Y., Kakkonen, T. & Sutinen, E. 2011. MinEDec: A decision-support model that combines text-mining technologies with two competitive intelligence analysis method. International Journal of Computer Information System and Industrial Management Applications, 3, 165-173.
[14] arabic script. International Arab Journal of Information Technology ,5(1) , 92- 101
[15] Ling, H. S., Bali, R. & Salam, R. 2006. Emotion detection using keywords spotting and semantic network. InProceedings of International Conference on Computing and Informatics IEEE (ICOCI), Kuala Lumpur, 1-5.
[16] Xu, X., Zhang, F. & Niu, Z. 2008. An ontology-based query system for digital libraries. In Proceedings of IEEE, Pacific-Asia Workshop on Computational Intelligence and Industrial Application, Wuhan, 222-226.
[17] Miao, D., Duan, Q., Zhang, H. & Jiao, N. 2009. Rough set based hybrid algorithm for text classification. Journal ofExpert Systems with Applications, 36(5), 9168-9174.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY