Fraud Detection in Web Advertisement

Sanap Kanchan S
Students of BE Computer
Sanjivani Collage of Engineering,
Kopargaon, Savitribai Phule,
Pune University

Kuwar Shraddha A
Students of BE Computer
Sanjivani Collage of Engineering,
Kopargaon, Savitribai Phule,
Pune University

Kulkarni Chinamy U
Students of BE Computer
Sanjivani Collage of Engineering,
Kopargaon, Savitribai Phule,
Pune University

T.Bhaskar
Asst .Prof (Computer Engg)
Sanjivani Collage of Engineering,
Kopargaon, Savitribai Phule, Pune University

ABSTRACT
The improvement of the technology and web-based application over the crime and fraud give best result in online advertisement. In recent years fraud is major problem in online advertising. It can affect the trust, beliefs and encouragement of the customer on online marketing. In this thesis, the development of this system can be done using Naive Bayes classifier and Apriori algorithm .The system can find fraud or scam in web based marketing and advertisement .It can also give the solution to the fake advertisement. Main aim of development of this system is public awareness which is very important in today's market.

Keywords
Fraud, Naive Bayes, Apriori, Scam.

1. INTRODUCTION
The remarkable growth in online search has substantially lowered the cost of information acquisition. That provides tremendous information at low cost. Which has altered searching technics and raise finding and getting efficiency, so that we explore the website "Craigslist" using this one can post his/her own advertise related to any area. It is an online platform provided to classify the advertisements considering area of automobile [1].

The online advertisement provides powerful mechanism to advertisers to handle the web users. The web application can contain the vast data and related things. We have to collect and management these data. For that in online searching the most important thing is data management. These problem become more challenging when some form of uncertainty in data or their relationship in data exists. Also the searching topics can have different domain including databases, data mining, algorithms, networking and many more.

During early stages of commercialization, many companies made online brochures about their products and services available. Some early web pages provided email addresses and their contact number for direct communication and other allowed user to fill out online forms to gather information about customer needs. But now a days data related to each advertise can be directly stored with add, so that for users it is very easy to search the product. Every day lot of advertises are posted on sites. Here we find out the fraud in online buying and selling of product.

As we know the web is collection of data from these kinds of data extracting particular data or pattern is called as data mining. It is a most important tool for the business to transform data into business for giving informational advantages.

For detecting fraud or scam in online advertisement e used the technic like Naïve Bayes classifier and Apriori algorithm [4]. As the name suggest these technologies have logical base and their performance can tested on statistical character of databases. Naïve Bayes is efficient learning algorithm for data mining and also for information retrieval. Most of the text is digital form and there is need to access data in flexible ways for that text classifier is beneficial. Next is Apriori algorithm which is a classic algorithm for learning association rules [2]. It is mainly design to operate on databases containing transaction.

2. RELATED WORK
Major current issue in online advertisement is detecting and searching the fraud considering the victimization it is a problem. Most of people refer the online buying and selling because only for low cost and less time requirement. But many people can fail in that kind of fraud in online advertisement will have impact on social community. Now a day's younger generation do online purchasing which is a routine activity for them and usually they get fail [1]. This is happen only due to less knowledge of business marketing and related current topics.

Taking the ratio of income and poverty is also related to the victimization [1]. The people who are having high school degree and have graduated education are less affected by this kinds of attacks or fraud. Women's are also are also less likely to be approached by fraudsters because they have more risk affected than men. E-Commerce cannot fulfill the requirement of the women for that they are rarely refer the online purchasing .So that lower fraud can be posted, ultimately it will reduce percentage of getting or doing fraud.

There are many techniques and algorithm are their which is used for classifying the text. But these technics are not yet correct and still are on improvement. There for there is a technic which can reduce difficulty that is Naive Bayes Classifier [3] is included. It is a simplest and based on Bayed theorem.

Apriori is follow bottom-up approach. Apriori algorithm used for understanding of association rule in that collected dataset can be checked against the related database. Using association rule difference item set are generated and these are tested or check against data. After that frequency of each item set will counted. This count can be used for getting frequent item set [4].

3. METHODOLOGY

Project input represent in form of text. For detect the fraud in advertisement we used technic collect text data from given input, feature extraction, vector table generation and final we apply Naive Bayes classifier for detect final result. For that we follows different step:

3.1 Load Database

We take actual input of project is nothing but the actual advertisement that stored in form of separated word. We load the trainee dataset [5] [6] which is previous checked dataset. This is used to compare Scam and Non Scam advertisement probability.

3.2 Feature Extraction

In feature extraction first step is removing stop words [7] [8] [9] like a, an, the .This process reduce the processing time of text. Next step is stemming in this we make for example freely, likely then 'ly' get removed and word treated as 'like' and 'free'. Next step we applied the Apriori on formatted input.

3.2.1 Apriori Algorithm

In Apriori dataset get divided into different item set where count of each item can take. Next item set we need to apply join operation [4].

Join: - It is process for finding possible combination.

Pruning: - It is process eliminate the combination which is not present in previous item set.

Support: - Frequency count of word.

Above explained process easily understand using example.

Example: -

Set of Words: - {A,B,C,D},{A,B,C,D,E},{B,C,D},{B,C,E},

{A,B,D}.

Table 1. Item set I

Item	Support
A	3
B	5
C	4
D	4
E	4

According to the above frequency of words. Let, Consider minimum support value =2.Frist we apply joining so next item set.

Table 2 Item set II

Item	Support
{A,B}	3
{A,C}	2
{A,D}	3
{A,E}	1
{B,C}	4
{B,D}	4
{B,E}	2
{C,D}	3
{C,E}	2
{D,E}	1

Here We Consider minimum support =2.So support >2 that combination consider for next item set. Above item set {A, E}, {D, E} is eliminated for further processing. On item set we apply joining and pruning.

Table 3. Item set III

Item	Support
{A,C,D}	2
{B,C,D}	3
{B,C,E}	2

Above Example explained working Apriori algorithm. Last Item set is used as input to Naïve Bayes classifier.

3.3 Result Generation

Here in our paper we used Naive Bayes classifier [3].

3.3.1 Naive Bayes Classifier

Which helps to take decision advertisement is fraud or not fraud. In this we calculate the probability means we check the possibility of outcome. Depend on probability value we can take the decision.

Example:-

Consider small Example advertise.

Scam advertisement: -

"Car sell, please contact to Mahindra, cash on delivery."

Non Scam Advertisement: -

"Car for sell, for purchase contact to Mahindra, Car purchase cash on delivery."

Take frequency of word occurrences.

Table 4. Word with frequency

Words	Scam Count	Non Scam Count
Car	1	2
Sell	1	2
Please	1	0
Contact	1	1
Mahindra	1	1
Purchase	0	2
Cash	1	1
Delivery	1	1

Depend on input we get frequency count of each word. We can calculate probability.

1. **Non Scam**
   P (Non Scam| Car, sell, contact, Mahindra, purchase, cash, delivery) =

   P(Non Scam) X P(Car |Non Scam) X P(sell| Non Scam) X P(contact| Non Scam) X P(Mahindra| Non Scam) X P(Purchase| Non Scam) X P(cash| Non Scam) X P(Delivery| Non Scam).

2. **Scam**
   P (Scam| Car, sell, contact, Mahindra, purchase, cash, delivery) =

   P(Scam) X P(Car| Scam) X P(Sell| Scam) X P(contact | Scam) X P(Mahindra |Scam) X P(purchase| Scam) X P(cash |Scam) X P(delivery |Scam)

   Above calculation can give final output of advertisement that is Scam or Non Scam.

4. RESULTS

When user want to check the advertise, firstly user should login to the system.

Fig. 1: Login screen for User.

User must give project inputs have given in text format only. After giving input we performed the pruning and streaming operation on a data. So we get the important and small data for processing. After that Apriori algorithm can be used for find combination which can present in advertisement and we get or generated vector table in that we get data in form of table like**,**

Table 5. Words with Frequency

Words	Counts
W1,W2,W3	4
W1,W5,W6	5

Then Naive Bayes classifier is used and we get final output given as advertisement as scam or non-scam. After click on checks then view of our website like show in snapshot.

Fig. 2:Non-Scam advertise check snapshot.

Fig. 3: Scam Advertise check snapshot.

5. FUTURE SCOPE

In this paper we find the scams in limited areas i.e. car and truck but in addition to this we can take the mobile, laptops etc. Also here we only take the input in the form of text, in that we use images also, for that we have to do image processing for the image. To provide the more security to user authentication here we use the OTP and many other ways.

6. CONCLUSION

In this paper we find the fraud in online advertising. It is one of the regnant and most done problems today. It's impact on the trust of the victim in online advertising. To find the fraud in online advertising we used two technic like Apriori Algorithm and Naïve Bayes classifier. These technics are not physical one so we get result dependent with dataset. Here we take input in the form of text and get the output as scam or non-scam. The main aim for implementation of this is public awareness. Good fraud management will provide great advantage over ad networks.

7. REFERENCES

[1] Vaibhav Garg and Shirin Nilizadeh, "Craigslist Scams and Community Composition: Investigating online Fraud Victimization". IEEE Transactions on security and privacy workshops year 2013.

[2] Metwally , D.Agrawal ,and A.El Abbadi ," Using Association Rules for Fraud Detection in Web Advertising Network". Technical Report 2005-13, University of california .Santa Barbara, 2005.

[3] Jingnian Chen, Houkuan Huang, Shengfeng Tian, Youli Qu," Feature Selection for text classification with Naïve Bayes", Department of Information and computing Science, Shandong University of finance,Jinan,Shandong,250014,china.

[4] Ishatiq Ahmed, Donghai Guan and Tae choong chang,"SMS classification Based on Naïve Bayes classifier and Apriori Algorithm Frequent Itemset",International Journal of machine Learning and computing, Vol. 4,No. 2,April 2014.

[5] http://www.carwale.com/

[6] http://www.cartrade.com/

[7] http://www.ranks.nl/stopwords

[8] http://adsabs.harvard.edu/abs_doc/stopwords.html

[9] http://tools.seobook.com/general/keyword-density/stop_words.txt

8. AUTHOR'S PROFILE

Sanap Kanchan is pursuing B.E Computer Engg in Sanjivani College of engineering, Kopargaon, Maharashtra, India. Her areas of research interests include Information Security, Data mining.

Kuwar Shraddha A is pursuing B.E Computer Engg in Sanjivani College of engineering, Kopargaon, Maharashtra, India. Her areas of research interests include Information Security; Data Mining.

Kulkarni Chinmay U is pursuing B.E Computer Engg in Sanjivani College of engineering, Kopargaon, Maharashtra, India. His areas of research interests include Information Security; Data Mining.

T.Bhaskar received a M.Tech (CSE) from JNTU Hyderabad. .He is currently working as Asst. Professor in Computer Engineering Department, Sanjivani College of Engineering, Kopargaon and Maharashtra India. His research interest includes data mining, network security.