CHAPTER 2

REVIEW OF LITERATURE

2.1 Introduction

Instant messengers (IMs) have been there for decades, bringing together people around the world through the internet, right from ICQ to AOL Instant messenger (AIM) and MSN messenger to Skype. IM users have been able to send and receive messages to each other instantly. Secure communication was always the prime concern in IMs and to have a reliable and trustworthy service, efforts have been made on a regular basis to improve its efficiency and performance. This chapter firstly deals with the overview of the relevant methodologies of phishing detection in instant messengers and then discusses about the various approaches and improvements made from time to time in the direction of providing effective solutions to the existing problems.

2.1.1. Instant Messaging

*Definitions*: Instant messaging can defined according to Webopedia as a type of real-time text based communication service over the internet between two or more people [5]. According to Wikipedia IM is a real time text transmission through online chatting over the internet. It’s a text based communication using a set of communication methodologies, between two or more users, over the internet [6].
Also, a system that exchanges typed messages instantly electronically through the web using a shares application from a personal computer. This message when available instantly to the recipient is the instant message [54].

**Brief History and applications:** Instant messaging has been constantly evolving over the decades, IRC (Internet Relay Chat) [55] was developed in 1988, by Jarkko Oikarinen. The year 1996 saw the launch of ICQ (I seek you) [56], America Online (AOL) [57] later acquiring it in 1998 and selling it off to Digital Sky Technologies in 2010. It had the basic features of multi user chats, file transfers and searchable directories. AOL launched AIM (AOL IM) in 1997, with additional basic services like user profiles, buddy lists, away messages and interactive icons. In 1998, Yahoo launched Yahoo Messenger with additions to the basic services like customized environments (IMvironments) and address book integration [58]. In the same year Pidgin launched Gaim where users using different operating systems could be communicated. Still today it is in use with features like file transfer, away messages, buddy icons apart from many plug-ins. It is compatible with all the leading IMs [59]. In 1999 Microsoft released MSN Messenger which gave the status as to when friends, family etc. were online. Later in 2005, it was renamed Windows Live Messenger, and additional features like photo sharing and storing, games and social networking were added [60]. The year 2000 saw the emergence of Jabber, which was a multi-protocol IM base on Extensible Messaging and Presence Protocol (XMPP), but tragically in August 2012 due to user abuse and DoS attacks closed down new registrations as the
migration of the account database is under process. Any other XMPP service can be used to create accounts [61][62]. Then came the iChat from Apple in 2002 and later released an updated version iMessage in 2010 which lets the users to exchange messages between iPads, iPhone, iPod touch and Mac apart from the latest features of any IM [63]. In the year 2003 Skype was released, where voice and video were added to instant messaging apart from file sharing and other standard features. It also has paid services to call any landline and mobile and text messaging through-out the world [64]. Meebo an IM service was acquired by Google in June 2012, and its products and features are no longer available ever since the take-over [65]. Google Talk (also referred to as Google chat or Gchat) were introduced in 2005 provided both text and voice communications. In 2013 Google replaced its client software offerings with Google + Hangouts [66]. Google Hangouts provide better group conversations and video calls and includes emoji (a new Japanese emoticon or a pictograph) during messaging [67]. Myspace, a social networking site launched its own IM MyspaceIM with the standard features along with skinnable interface, picture cropping, and customized emoticons and zap (combination on of picture, words and sound bites) [68] [69]. Facebook launched Facebook chat in 2008 and Facebook messenger in 2012 with the standard features for desktops, Facebook chat’s unique feature is that it is minimal [70] [7]. There are many more to be mentioned and yet to come, but with respect to the relevance of the work, the main concern here is to address the security issues.
Apart from information exchange and chatting at personal level and file sharing and other discussions at workplace level, IMs have also been used by the US at the tactical level for command and control purposes, exchange of tactical information, technical failure reports and status reports [8].

**Architecture:** IMs use the client server architecture where the major components are the Web Server, the Web Browsers, IM multiplexor, IM Server and the Lightweight Directory Access Protocol (LDAP) Server.

![Basic Instant Messaging Architecture](image)

**Figure 2.1 Basic Instant Messaging Architecture**

The figure 2.1 shows the basic architecture of Instant Messaging system with the flow of the authentication requests. An explanation of the steps in this authentication process is as follows:
1. The client accesses the IM applet URL from the web browser and invokes the client to choose a method.

2. The Java plugin is invoked by the browser.

3. The required Instant Messenger resource files are downloaded from the Java plugin and the IM gets started.

4. The client enters the login details in the login window, which is sent through the multiplexor to the IM server.

5. In order to authenticate the client the IM server communicates with the LDAP server and requests the user’s contact lists.

After the authentication is completed, the IM main window is displayed along with the user’s contact list. The Instant Messaging chat sessions hence get started with other chat mates [9].

**Threats and vulnerabilities:** IMs are not different from other internet applications and services when it comes to threats and vulnerabilities. IMs are prone to threats like worms and phishing when deceitful extraction of confidential information is on the rise making it unreliable and insecure. With existing weak intrusion detection mechanisms and ineffective anti-viruses, the computers become vulnerable and can be remotely accessed by hackers [10].

Vulnerabilities are common programming mistakes made by developers, usually causing DoS attacks and allowing unauthorized remote access to hackers. Many web applications are likely carriers of malwares and worms. IMs and their clients spread approximately 30 worms by using four methods:
• Exposing functionality of IMs from the APIs to create worms.
• Programming maliciously to click the required buttons to list out the buddy list or sending files through IM client window.
• Sending an instant message with a URL link, to a worm file on an infected node instead of a file
• Creating patches for the client DLLs (a malicious file) to be sent whenever an instant message is sent [10].

Backdoor Trojan Horses [43] also send themselves or sensitive information, replicate and perform malicious actions in IMs.

In IMs a phisher can elicit deceitfully passwords and confidential data, which is extracted without the IM user realizing it [10].

2.1.2. Phishing Overview

Phishing attacks originated from the theft of AOL accounts and evolved over the years by making its presence felt in more profitable domains like internet banking and ecommerce. This phished information would be used for financial gain, making purchases through stolen identities and if not for any other thing, merely for the purpose of peer recognition and receiving fame [3]. The increase in phishing attack instances as surveyed by the Global Phishing survey 1H2013 of the Anti-Phishing Work Group (APWG) has been discussed in Section 1.1 [4].
Definitions: According to the APWG, phishing is a crime which uses socially engineered messages and technical trickery to extract the user’s personal data and financial details [11].

Phishing can also be defined according to Mahmoud Khonji et.al., as a type of system attack which sends socially engineered messages to users through instant messengers and emails, to make them perform certain actions for the phishers benefit [3].

In the above definitions, socially engineered messages refer to either deceitful instant messages or spoofed emails supposedly from legal sources, to trick users into revealing their confidential details like passwords and other important data. Technical trickery or subterfuge refers to implanting malware onto the victim’s PC by directly stealing or intercepting the confidential information [11].

As discussed in Section 1.1, the term phishing is derived from fishing, where the phisher elicits personal data from the end user, here the initial two letters ph became the replacement for the character f in common context [3].

Live examples of phishing in emails, URLs, websites and Instant messengers are demonstrated below.
Figure 2.2 A fake website showing phishing attack in Facebook URL

Figure 2.3 An Email phishing attack example
Figure 2.4 An Email phishing attack in Yahoo! Mail for account details

Figure 2.5 An instance of IM phishing attack in MSN Live Messenger.
Most IMs play a vital role for spreading phishing or virus files. When chatters receive some file attachments as shown in Figure 2.5, from their trusted buddies they download it, subsequently they may be monitored by phishers.

Consider for example, a typical password phishing scenario, in which a phisher has gained the trust of a gullible chat mate during multiple sessions of chatting.

He might ask the chatter: “Hi! I am having trouble in creating my Gmail account, as it says my password is very weak. Could you suggest me some hints.”

This gullible chatter who wants to help out his dear friend suggests certain tips, which he himself used for his various account creation purposes.

So he replies: “Why don’t u try some special characters, or use some caps. You could use somebody’s name who’s very special to you for easy recalling”.

Here it is observed that the gullible chatter falls prey to the foxy inquiry by the phisher and ends up revealing keys details due to lack of technical awareness.

In all the above instances shown it is observed that the phishers try to use all possible means to deceive the user and elicit confidential data through phishing emails, fake websites, phishing URLs and instant messages.
**Issues:** User ignorance towards the use of communication services like IMs, emails, websites, social networks etc. is becoming the biggest advantage for phishers. As there does not exist any permanence in solving the phishing problems, efforts could be made in the direction to minimize the impact of phishing. The possible solution to this issue is firstly to create technical awareness in users by educating them and secondly to create a phishing detection application which would generate phishing alerts for the user so that phishing messages are not ignored. The approach would classify the message as phishing or normal [3].

Active research work is in progress for various phishing detection approaches and considerable efforts have been made in services like emails and websites. With respect to the increasing phishing menace in instant messengers more research contributions are still needed, not only for detection and prevention of phishing but also for regaining the lost trust of e-users in using internet banking and ecommerce, this would be discussed in detail in Section 2.2.

### 2.1.3. Ontology Based Information Extraction (OBIE)

An ontology contains a set of concepts within a domain are signified by knowledge. Ontology is a structural framework for organizing information and represents knowledge about the world as a whole or anything contained in it. It is used widely in areas like the Semantic Web, software engineering, information architecture, AI etc. [15]. Ontology is defined as a specific description of concepts in a domain. It is a type of collection of information
used by the semantic web in order to discover common meanings for whatever knowledge bases encountered. It captures the domain knowledge so as to help extract relevant information for the given keyword [14].

Ontology Based Information Extraction (OBIE) is a mechanism which extracts and presents information from numerous data sources guided by ontologies. OBIE integrates processing of text with information retrieval and semantic web techniques in order to extract useful knowledge from various text sources efficiently [48]. OBIE is a subset of Information Extraction (IE). In this process the information has to be extracted from the unstructured text and converted into machine usable form. For this the pre-processing steps involved are stemming, removing the stop-words and storing the machine usable keywords. Semantic lexicon is used to identify the domain of the keywords under process. It is a collection of words pertaining to a particular domain. In the semantic lexicon the domain is uniquely identified as a part of the domain vocabulary. For instance the semantic lexicon for online banking includes words like password, account type, savings bank account, debit card number, credit card number, code etc. The OBIE architecture and extraction workflow will be discussed in detail in Chapter 3. After the domain is identified, the instance information is extracted and a Resource Description Framework (RDF) [44] node is formed which eventually updates the ontology using any of the ontology editors like Protégé [45]. The Triplet Extraction algorithm is used to extract the Subject, Predicate and Object from a sentence or message. In this process the Natural Language Processing (NLP) is used to
extract the Subject and maps it to a semantic class and the predicate and object are used as attribute name and value respectively. Ultimately the domain ontology is inferred from the theme concept. For example, for the sentence “Novotel is located in Hyderabad”, the identified subject is Novotel, predicate is located, and object is Hyderabad. The theme concept identified here is Novotel and domain class extracted will be “hotel” [12][15].

The additional task to be accomplished, apart from identifying the domain ontology, is the extraction of the context or the intention of using the word, which may help a great deal in identifying the relevant phishing words with the least number of false positives and false negatives. The identification of the context too will be discussed in detail in the Chapter 3.

2.1.4. Data Mining

Data mining can be defined as the extraction of interesting patterns from large repositories of information. These interesting patterns can be unknown previously, non-trivial, implied and useful. It is alternatively known by many names, but most prominently known as Knowledge Discovery of Databases (KDD). It is actually the knowledge mining from data [16]. Data mining (DM) can be applied to any type of data as long as it is relevant and has meaning. It adopts techniques from various allied subjects like statistics, information extraction, artificial intelligence, machine learning, pattern recognition, high performance computing and many more application areas.
The functionalities of DM are used to elaborate the kinds of patterns to be found in DM tasks. They include class / concept description, mining of frequent patterns or association rule mining, associations and correlations, classification and prediction, clustering, and outlier analysis. The main application areas are Business Intelligence (BI), search engines, bioinformatics, software engineering, health informatics, banking, finance, marketing surveys, digital libraries, digital governments etc. Consider for example the Business Intelligence, which provides the historical, present and predictive view of the various business operations viz., ad hoc reporting, Online Analytical Processing (OLAP), managing business performance, market analysis, prediction and forecasting etc. [16].

Discovering frequent patterns that customers tend to purchase item sets, first a cell phone, followed by a pen drive and then a memory card together frequently as depicted from the transactional data, gives the associations between the given item sets.

Classification finds a model which describes important data classes. It is a type of data analysis model for instance, to categorize whether issue of credit card to a customer is risky or safe, thus giving a better understanding of the data at a larger scale. It has a number of applications ranging from fraud detection to medical diagnosis [16].

Associative classification is a better option where frequent patterns can be used for classification. The association rules are generated from frequent
patterns and then used for classification. The different methods used are Classification based on Associations (CBA), Classification based on Multiple Association Rules (CMAR), and Classification based on Predictive Association Rules (CPAR). In the current research work of phishing detection in IMs, it becomes necessary to find frequent phishing patterns for the discovery of interesting association rules in order to construct a classifier using CBA. These rules are ordered based on decreasing precedence depending on their support and confidence [16].

2.1.5. Association Rule Mining

Association Rule Mining (ARM) discovers the association relationships from item sets in transactional data and other information repositories with its easy to understand nature of rules. Frequent patterns are mined or searched for discovering recurring relationships in a given set of data [16]. In ARM, the associative relationships between itemsets in a dataset are discovered which are in the form of simple and easy to understand rules [17]. Initially ARM was introduced to discover associations and correlations among items for the market basket analysis, but was applied in many other areas like credit card fraud, network intrusion detection, market basket analysis, cross-marketing, catalog design, store layout design, customer shopping behavior analysis, genetic data analysis, microarray data analysis etc. [16]. Market basket analysis is used to analyze the buying habits of a customer by discovering the associations between the various items a customer keeps in his / her shopping baskets, in order to develop better
marketing strategies for customer shopping behavior analysis, catalog and store layout design [16]. It is also used to attract more and more people to the stores and shopping malls, better advertising and promotions, increase the size and value of the shopper’s baskets, to use the market as a laboratory, and to make the retailers take better decisions [46].

For instance 60% of the customers who buy bread, are likely to buy butter as well on the same trip to the stores; 5% of the transactions include bread and butter. These types of associations when analyzed may increase the sales by planning their shelf space and store layout. The phrase “customers who buy bread” is known as the rule antecedent, and “buy butter as well” is the rule consequent. The 60% of the association rule here represents the strength of the rule and referred to as the rule’s confidence. The 4% is the rule’s support which signifies the statistics [17].

The confidence and support are the ARM performance measures for rule interestingness. These rules would be interesting if they satisfy user defined thresholds like minimum support threshold (min_sup) and minimum confidence threshold (min_conf). ARM is thus a two-step method, firstly finding all frequent itemsets (with support greater than equal to min_sup) and secondly generating strong association rules from these itemsets (satisfying both min_sup and min_conf) [16]. Apriori [18] and FP-growth [74] are the two important association rule mining algorithms. In Apriori the property used is: all not empty subsets of the itemsets which are frequent
should also be frequent for pruning the search space. It uses the breadth first search (BFS) whereas the FP-growth follows the Depth first search (DFS) [72].

2.1.6. Classification

Classification is a form of data analysis which extracts models which describe the data class. It is a process which involves a learning step (or the training phase) and a classification step. A classification model or classifier is built in the learning step, whereas class labels for the given data are predicted in the classification step. Classification implies learning a function that classifies or maps a data instance into one of class labels which are predefined [71]. The dataset from which learning of a classification function or model is done is known as the Training set. For testing the classifying capability of the learned function or model, a Testing set is used [72]. In the first step training data are analyzed by the classification algorithm. In the second step test data are used to find the accuracy of the classification rules. If the accuracy is satisfactory the rules will be applicable to classify new data [16].

For example if data analysis is required to find whether a person with a certain profile is going to purchase a laptop. The data analysis here signifies the classification where the model i.e. the classifier is built to predict the class label like “yes” or “no”. Here the class label of the training data is provided, this refers to supervised learning. Whereas is the class label of the training data is unknown i.e. the count of the classes is not known, it is unsupervised learning or clustering [16].
The applications of classification include fraud detection, target marketing, medical diagnosis, performance prediction etc. The various classification techniques include decision tree induction, Bayesian classification and rule based classifiers. Among the various decision tree algorithms, Iterative Dichotomiser (ID3), C4.5 or J48 (successor of ID3 and introduced by Quinlan) and Classification And Regression Trees (CART) used the top-down approach. More advanced classification techniques include Bayesian belief networks, Back propagation, Support Vector Machines (SVM), Associative classifications like CBA, CMAR and CMAR (classification using frequent patterns). These classifiers are examples of eager learners, whereas instance based classification methods like nearest-neighbor classifiers (k-nearest neighbor kNN) and case based reasoning classifiers are lazy learners [16].

When a classifier is constructed from rules, often it is represented as a list of rules where the order of the rules is based on the significance of the rules. Classification rules exist in the form P → c, where P is a pattern in the training data and c is the predefined class label [72].

Classifier Performance

The performance of a Classifier is measured by accuracy, i.e. the percent of correct predictions to the total no. of predictions made. Sensitivity, specificity, precision and recall also, are the various measures used to understand the varying aspects of the generated classifier model. Of these, precision and recall are used in this work and are defined as follows [72]:
Precision = true positives / (true positives + false positives) ....Eq. 2.1

Recall = true positives / (true positives + false negatives) ....Eq. 2.2

In classifier construction, a model is constructed from a training set where each instance has a class label. This is then tested to know how the new instances are predicted. Testing can either be done over the training set or an independent test set. The former way of testing is usually not recommended since the classifier has been generated from the same data. The latter testing which uses a separate test set is recommended as it shows the classifiers performance on new instances [72].

Based on the amount of available data, training and testing can be done by splitting it in the ratio of two thirds is to one third respectively, in case of large of data. For limited data, $n$ fold cross-validation is recommended to maximize the use of the data to produce a good classifier. Here the data is divided into $n$ folds where the individual folds are used for testing and the other folds are used for training. The accuracy found will be average of the $n$ iterations [72].

Even though active research has been done in the comparative analysis of the above classifiers with respect to various parameters and characteristics, associative classification like CBA has been used in the current research work, based on the study to be discussed in the following sections 2.1.7 and 2.2.
2.1.7. Classification Based Association (CBA)

ARM generates a large number of rules in some instances, but many of them either are repeated or do not portray the true relationship among data itemsets. At times even strong association rules can also be misleading. Alternatively ARM when used to build classifiers is called as Associative Classification (AC). According to Thabant et al, AC combines association rule mining and classification to discover the knowledge, and then choosing a group of rules to construct a classifier [20]. This method explores all the associations between attribute values and their classes in the training dataset in order to build large classifiers. Association rule based classification was proposed in [73] through two algorithms: firstly an algorithm called CBA-RG (similar to Apriori) to generate rules and secondly algorithm for building the classifier called CBA-CB. The first algorithm is a Rule Generator and the second a Classifier Builder. The rules generated by Rule Generator are called Classification Association Rules (CARs), and have a predefined class label or target.

The CBA-RG algorithm finds all rule items of the form < conditionset, y > where conditionset is an item set, and y ∈ Y where Y is the set of class labels. The supportcnt of the rule item is the no. of instances in the data set D that contain the conditionset and are labeled with y. Each rule item corresponds to a rule of the form: conditionset → y. A subset of rules is selected from the generated CARs, based on the criteria which the subset of rules can classify the training set precisely [72].
The main difference between ARM and AC is that AC considers only the class attribute in the rules consequent, whereas ARM allows multiple attribute values in the rules consequent [19] [73].

AC incorporates the following steps: firstly, finding the frequent itemsets (finding the commonly occurring attribute value-pairs in the itemsets). Secondly, generating association rules per class by analyzing the frequent itemsets, supporting min_sup and min_conf. Lastly, create a rule based classifier by organizing the rules [16].

According to empirical studies, AC is more efficient method than traditional classification approaches such as decision tree and The k-nearest neighbor algorithm (k-NN), which constructs better and predictive classification systems. Moreover, AC generated outputs are of the simple form of if-then rules, making it easily understood [19], [20].

The various methods used for AC are CBA, CMAR, and CPAR, but the simplest among these is the CBA (Classification based on associations). Here the iterative approach of ARM i.e. Apriori is used by the CBA, where the rules are ordered according to the decreasing precedence depending upon the confidence and support, in order to create a classifier. According to Han and Kamber, experimentally the CBA was more accurate than the C4.5 on large number of datasets [16].

A decision list is formulated by the set of rules which makes a classifier. Some of the rules are:

- If a rule set contains the same antecedent then the highest confidence
The rule with the lowest precedence specifies the default class for new tuples, which are not satisfied by any other rule in the classifier, which is the default rule.

- Also when a new tuple is classified, the first rule itself, which satisfies the tuple, is used to classify it [16].

The CBA assigns class labels after discovering all frequent item sets, and generates the Class Association Rules (CARs) after calculating the confidence for each CAR. The rule item that is greater than the \text{min\_conf} value becomes a rule. These rules are then ranked according to confidence and support values.

Classifiers based on association rules have been made for different domains i.e. for classification of mammography images [75], for classification of web documents [76], for recommender classification systems [77], for classification of spatial data [78], for classification of documents [79], and text categorization [80].

2.1.8 Weka

Weka [87] is an open source software developed by University of Waikato, New Zealand, used to perform data mining tasks using a set of machine learning algorithms, which can be applied to datasets or can be called from Java source code. The tools are used to perform data preprocessing, association, classification, clustering, regression and visualization on the datasets.
2.2 Phishing Detection Overview

A general overview of phishing was presented in the Section 2.1.2, where its origin, concepts and history were traced. Various issues related to the lack of awareness among the users and the need to educate them, were also discussed. This section provides a study of the ongoing research work in the phishing detection scenario from the recent past as well as the present proposals.

Creation of technical awareness among the users by education them has been discussed in the earlier sections, but according to Mahmoud Khonji et al [3], the essence of global education is very large, and does guarantee the prevention and detection of phishing as such. It may be a part of the solution and not the solution. When the user education material i.e. Anti-phish phil [81] was used, which taught the employees the method of identifying malicious links and Fake URL addresses, thereby creating a considerable drop in the false negatives from 43% - 29%, which was not up to the satisfactory levels [3].

The phishing detection in the past has not been dynamic i.e. pre-defined blacklisted phishing words were used for matching and could not detect the zero-hour phishing attacks, which is critical from the detection accuracy point of view. Also, when there is increase in the false positives the system would be more harmful rather than being useful, as the users will start ignoring the system warnings due to wrongly reported phishing alerts [3].

All the previously detected phishing words, URLs and IP addresses are listed
and updated from time to time and stored as *blacklists*. These blacklists cannot provide protection from the zero-hour phishing attacks i.e. the phishing words or URLs which are not available in the blacklists previously. It took about 12 hours to detect and blacklist the new phishing words or URLs, which is a considerable delay owing to the fact that about two-thirds of the phishing attacks end within two hours [21].

Google Safe Browsing API, DNS-based Blacklist (DNSBL), PhishNet are the examples of services using the blacklists. Google Safe Browsing API allows client application to check whether the phish words are there in the blacklist which is continuously updated by Google, and used in Google Chrome and Mozilla Firefox [22]. The DNS-based Blacklist (DNSBL) uses the DNS protocol and if the listed entries count is large, and the DNSBL server is not optimized for handling large amounts of Resource Records, it may face performance problems [23]. Phishnet uses predictive blacklisting [82], i.e. whenever changes in the phishing URL, result in no matching, the exact matching limitation in the blacklist, is addressed by the Phishnet [23]. The blacklisted URLs are processed to produce multiple variations of that URL through different varying heuristics.

In the above blacklisting approaches protection from zero-hour phishing attacks was not available. To overcome this problem, phishing heuristics i.e. experience based methods of problem solving were used like SpoofGuard, Phishguard [83], Phishwish, and Cantina [3]. Of these the first two i.e.
SpoofGuard and Phishguard are plugins for web browsers [24] [25].

Phishwish [26] uses minimum number of heuristic phishing rules in order to identify dynamically whether the message is a phishing message, with least number of false positives.

Cantina [27] is basically a toolbar for Internet Explorer which performs content based phishing webpage detection. It minimizes false positives by using heuristics, web based search engines and Term Frequency-Inverse Document Frequency (TF-IDF). The heuristics used are: domain age more than a year is permissible, and it is a phishing website if suspicious page Uniform Resource Identifier (URI) contains (-, @, more than 5 dots).

Data mining techniques like association rule mining (ARM), classification and clustering have also been used to detect phishing through algorithms like Apriori, C4.5, Support Vector Machines (SVM), k-Nearest Neighbor (k-NN) and Density-Based Spatial Clustering of Applications with Noise (DBSCAN).

ARMs Apriori algorithm is used for finding phishing patterns dynamically in Instant messengers from not only text messages but also from voice chatting, by integrating the IM system with speech recognition system [28]. This is purely a content based phishing detection system with increased FPs.

In the k-NN approach, training instances are stored in the memory, which exist in the form of multi-dimensional vectors. Classification is then performed on the test dataset by obtaining the distance between both the
testing and training instances. The classes of three nearest neighbors are obtained with \( k = 3 \) [3]. In the survey, a comparison was between k-NN, SVM, Naïve Bayesian and linear discriminant analysis based on the network performance of the website characteristics in order to detect phishing [29]. Phishing detection in emails using C4.5 decision tree induction was done in [30] based on the highest information entropy, but it could not determine the phishing threat type.

Advanced classification method like Fuzzy Set Approach has been used efficiently by Maher et al. [32] to detect phishing websites in the E-banking sector. The intelligent fuzzy-based classification system also uses associative classification algorithms to detect phishing websites in the e-banking, and extracts the phishing features and classifies them into phishing rules with a layered structure [32].

As discussed in section 2.1.7, the Associative Classification (AC) uses CBA, CMAR, and CPAR to generate classifier rules, but the according to [16], the Classification based on associations (CBA) was more accurate than the C4.5 on large number of datasets. The iterative approach of Apriori is used by the CBA to create a classifier [16]. This AC approach has been used in [32] along with fuzzy based scheme for detection of phishing websites but nowhere for phishing detection in instant messengers.

In the above discussions phishing detection has been surveyed to a large extent in emails, URLs and websites but it has been observed that very less
efforts have been made in the phishing detection in instant messengers. The purpose of studying the types of phishing approaches in emails, URLs and websites is to understand the extent of the research progress made in the three communication methods and how much progress remains to be achieved in instant messengers. Both the static and the dynamic approaches have been discussed, where blacklisting of phishing words and URLs represented the former, and zero-hour phishing detection represented the latter. But the main observation here is that even though the zero-hour phishing detection was dynamic, but still there was no considerable decline in the false positives and the false negatives, as all these phishing detection approaches including phishing heuristics and data mining algorithms were content based methodologies.

Natural Language Processing (NLP) [47] techniques play a crucial role in overcoming the shortcomings mentioned above, by giving the advantage of utilizing the semantics of the messages exchanged between chatters in instant messaging. In the discussions of section 2.1.3, the Semantic Web uses the ontology and could incorporate the NLP to extract triplets i.e. subject, predicate and the object of the message under process, as a part of Ontology Based Information Extraction (OBIE). During the research study it has been observed that NLP has not been utilized to its full potential and at times it has only been used as an alternative to information extraction.

The NLP is used for Intrusion detection and phishing detection in emails
through pattern matching using Ontosem [84]. The word matching is by the keyword’s literal meaning only. It does not utilize the domain concept [31].

Another frame work which detects suspicious messages in Instant messengers and social networking sites using Ontology and Association Rule Mining (ARM) is used to detect the domain of the message keywords [33]. Pre-defined phishing rules have been considered which are to be mapped with the ontology generator through NLP, to obtain the domain of the suspected word. The user is not alerted instead the e-crime department is notified if any suspicious activity is detected [33]. The overall detection is content based and thus in spite of using the NLP, the context part is not explored, the use of ontology is restricted to information extraction and domain identification only. The overall performance does not reduce the false positives and false negatives.

To conclude, the various concepts, history, architecture, threats and vulnerabilities of IMs were discussed. The concepts of phishing and its practical scenarios and its issues were also covered. The ontology concepts with their examples were studied with respect to the OBIE, in order to identify the domain and context of the IMs. Concepts of data mining, ARM, classification, CBA were also discussed. Lastly, the various research efforts relating to phishing detection with respect to blacklisting, phishing heuristics, data mining approaches and NLP were also discussed.