EVALUATION OF WORD EMBEDDING TECHNIQUES FOR THE VIETNAMESE SMS SPAM DETECTION MODEL

Vu Minh Tuan[*], Do Thuy Duong[*], Tran Quang Anh[+]
[*]Hanoi University
[+] Posts and Telecommunications Institute of Technology

Abstract: The escalating issue of SMS spam in Vietnamese text messages has prompted the adoption of machine learning and deep learning models for effective detection. This paper investigates the impact of word embedding techniques on enhancing SMS spam detection models. Traditional statistical methods (BoW, TF-IDF) are compared with advanced techniques (Word2Vec, fastText, GloVe, PhoBERT) using a proprietary dataset. The evaluation focuses on accuracy, precision, recall, and F1 Score. PhoBERT integrated with CNN model showcased the highest accuracy of 0.968 and a remarkable F1 score of 0.941. The study sheds light on the role of word embeddings in constructing robust spam detection models, offering valuable guidance for model selection. The methodology, comparative analysis, and future directions are also presented in the paper.

***Keywords*:** Vietnamese SMS spam, word embedding, deep learning, CNN.

I. INTRODUCTION

In recent years, the proliferation of mobile devices and the widespread use of short message service (SMS) have led to an increase in SMS spam, posing a significant challenge for effective spam detection in Vietnamese text messages. To combat this issue, machine learning and deep learning models have been widely adopted for SMS spam detection, where the quality of word representation plays a crucial role in the model's performance.

Word embedding techniques have emerged as powerful tools for capturing the semantic meaning and contextual information of words in natural language processing tasks. In the context of SMS spam detection, word embeddings can be particularly beneficial in transforming raw text messages into meaningful numerical representations, enhancing the performance of models.

In this paper, Naïve Bayes (NB) and Convolutional Neural Network (CNN) were deployed as representatives of traditional machine learning and deep learning models to detect Vietnamese SMS spams. This study aims to explore and evaluate various word embedding techniques for developing an efficient SMS spam detection model for Vietnamese text. The performance of traditional statistical word embedding methods such as Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) against more advanced techniques such as Word2Vec, fastText, GloVe and PhoBERT will be assessed. The evaluation will be conducted using a private dataset of annotated SMS messages, comprising both spam and non-spam messages, with specific attention to accuracy, precision, recall, and F1 Score. The combined utilization of PhoBERT and the CNN model demonstrated the utmost accuracy value of 0.968, accompanied by a notable F1 score reaching 0.941. By shedding light on the significance of word embedding techniques in the context of Vietnamese SMS spam detection, the research provides insights into choosing appropriate word representations for building robust and accurate spam detection models.

The rest of the paper is structured as follows. The related works are reviewed in Section II. All about the methodology including data collection and pre-processing, word embedding techniques is presented in Section III. Section IV compares and discusses the results of the research. Section V includes conclusion and future works.

II. RELATED WORKS

Word representation techniques played an important role in the achievement of the spam detection models. These techniques were compared or evaluated as a factor of success in studies mentioned below.

Gauri Jain et al. [1] proposed using a convolutional neural network (CNN) with a semantic layer on top, creating a semantic convolutional neural network (SCNN). The semantic layer enriched word embeddings using Word2Vec for training random word vectors and WordNet and ConceptNet to find similar words when word2vec is unavailable. The SCNN architecture was evaluated on two corpora: SMS Spam dataset (UCI repository) and Twitter dataset. The approach achieved impressive results, outperforming the state-of-the-art with 98.65% accuracy on the SMS spam dataset and 94.40% accuracy on the Twitter dataset.

Sreekanth and Maunendra [2] presented several deep learning models based on convolutional neural networks (CNNs). In total, five CNNs, each employing different word embeddings (GloVe, Word2Vec), and one feature-based model were utilized in the ensemble. The feature-based model incorporates content-based, user-based, and n-gram features. The proposed approach is evaluated on two

Contact author: Vu Minh Tuan
Email: minhtuan_fit@hanu.edu.vn
Manuscript received: 8/2023, revised: 9/2023, accepted: 9/2023.

datasets, one being balanced and the other imbalanced. The experimental findings reveal that the proposed method surpasses existing techniques in terms of performance and accuracy.

Lee and Kang [3] investigated the utilization of word embedding for feature vector construction and deep learning for binary classification. CBOW was employed as the word embedding technique, and a feedforward neural network was used to classify SMS messages as ham or spam. The experimental results indicate that the deep learning method (95.87%) outperformed the conventional SVM-light machine learning method (95.72%) in terms of accuracy for binary classification.

Surajit et al. [4] proposed multiple deep neural network models for spam message classification, utilizing Tiago's Dataset. Preprocessing steps involved text cleaning and tokenization. Two deep learning architectures, CNN and hybrid CNN-LSTM were used for classification, and BUNOW and GloVe word embedding techniques are incorporated to improve accuracy. The best accuracy of 98.44% is achieved by the CNN LSTM BUNOW model after 15 epochs on a 70%-30% train-test split. Other researchers applied machine learning and deep learning methods to tackle the challenge of detecting SMS spam including [5] [6] [7] [8].

While earlier research has introduced models employing word embedding methodologies for identifying spam messages in diverse languages, a noticeable gap exists in terms of comparative investigations concerning the impact of word representation techniques within the domain of Vietnamese SMS spam detection. This study seeks to assess the efficacy of six word representation methods in conjunction with various machine learning and deep learning classifiers. The outcomes of this inquiry will serve to enhance SMS spam detection technology, thus augmenting the security and user satisfaction of mobile communication among Vietnamese users.

III. METHODOLOGY

A. Dataset

The spam dataset consists of a total of 50,000 text messages. Among these, approximately 20,000 spam messages are collected through a honey pot system utilizing 32 phone numbers running on a sim-bank device. Additionally, around 25,000 spam messages are provided by the Vietnam Computer Emergency Response Team (VNCERT). Furthermore, approximately 5,000 spam messages were contributed by 30 volunteers over a period of 3 months.

Within the spam dataset, 65% of the messages contain diacritics, while the remaining 35% are without diacritics. To mitigate redundancy, a portion of the spam messages (about 35,000) is used for the experimentation.

The ham dataset comprises 45,000 legitimate messages. These non-spam messages are voluntarily contributed by 10 participants. Within the ham dataset, 95% of the messages contain diacritics, with the remaining 5% being without diacritics. The redundancy level is estimated to be around 10%, resulting in a selection of approximately 40,000 messages for experimentation. In terms of the message's length, the majority of usual messages typically consist of 0 to 10 words per message, with only a limited portion having around 30 to 40 words. Conversely, most spam messages tend to have a word count ranging from 30 to 40 words.

By meticulously organizing the spam and ham datasets, including details on their sources and redundancy ratios, we aim to provide a reliable and diverse dataset to thoroughly evaluate the effectiveness of Vietnamese SMS spam filtering methods in the upcoming experiments (Table 1). All experiments were conducted with the full-accented Vietnamese dataset. The dataset was split into 70% for training and 15% for validation and 15% for testing.

Table 1 Dataset Description

	Total	**Accented**	**Unaccented**
Spam	35.000	22.750	12.250
Ham	40.000	38.000	2.000

B. Text Preprocessing

Before training any model, data preprocessing is essential to improve overall accuracy. The preprocessing steps involve data cleaning to remove redundant content, stemming to bring documents into a consistent form, stop-word removal to eliminate irrelevant words, and word segmentation to determine the boundaries and types of words in a Vietnamese sentence.

Data cleaning focuses on removing noise data like paths, unnecessary phrases, or meaningless characters in SMS text messages. Stemming optimizes storage memory and accuracy by converting various word forms into a unified representation. Stop-word removal is crucial as it eliminates common and uninformative words from the dataset.

In Vietnamese text processing, word segmentation is vital to determine the grammatical structure and word types in a sentence. These preprocessing techniques ensure that the data is appropriately prepared for the model, leading to better performance in text classification tasks.

C. Word Embedding Techniques

Given that machine learning algorithms can solely process numerical data, the process of making words and text meaningful entails converting them into numerical representations. Proper preprocessing is essential to handle the challenges posed by noisy and unstructured SMS data, and word embeddings can aid in converting the SMS messages from the text format into the digital format by different mechanisms. We categorize word embedding methods into two groups: statistical approaches (BoW, TF-IDF) and contextual approaches (Word2Vec, GloVe, FastText, PhoBERT).

1) Statistical word embedding approaches

The Bag-of-Words (BoW) is a prevalent Natural Language Processing (NLP) technique used for text modeling. It represents text by capturing the word occurrences within a document, disregarding grammatical details and word order. The term "bag" refers to the fact that any information pertaining to word order or structure is discarded, and the model solely focuses on whether known words occur in the document, not their position within the text. Employing the BoW technique facilitates the conversion of a text into a numerical vector representation, enabling further analysis and processing in the NLP domain.

TF-IDF is a significant text representation technique in Natural Language Processing (NLP). It calculates

word importance within a document relative to its occurrence in the entire corpus [9]. TF measures word frequency in a document, while IDF reflects word rarity across the corpus. The TF-IDF score for each word is the product of these two factors. Common words appearing in multiple documents get lower TF-IDF scores, while rare words with high occurrences in specific documents receive higher scores. Employing TF-IDF converts texts into numerical vectors, facilitating NLP tasks like document classification and information retrieval.

2) Contextual word embedding approaches

Word2Vec aims to represent words in a continuous vector space, capturing semantic and contextual relationships between words. Word2Vec employs a neural network architecture, where words are mapped to dense vectors, often of a fixed dimensionality [10]. The key idea behind Word2Vec is that words with similar meanings or that appear in similar contexts should be represented as vectors that are close to each other in this vector space. There are two main approaches in Word2Vec: Continuous Bag-of-Words (CBOW) and Skip-gram. CBOW predicts a target word given its context, while Skip-gram predicts context words given a target word.

Developed by researchers at Stanford University, GloVe aims to represent words as vectors in a continuous space, capturing both global word co-occurrence statistics and their semantic meaning [11, 12]. Unlike traditional word embedding methods that only consider local word relationships, GloVe leverages the entire corpus to build a global word-word co-occurrence matrix. It then employs matrix factorization techniques to generate meaningful word vectors. These vectors preserve semantic similarities between words and can be used for various NLP tasks, such as word analogy, sentiment analysis, and text classification.

fastText is an advanced word embedding technique developed by Facebook AI Research (FAIR). It extends the concept of traditional word embeddings by representing words as character n-grams, allowing it to handle out-of-vocabulary words and morphological variations effectively. This unique approach makes fastText particularly suitable for languages with rich morphology like Vietnamese. The model learns to generate dense vector representations for words, subwords, and character n-grams, capturing their semantic and syntactic properties. FastText's efficient implementation enables fast training on large-scale datasets, making it popular for tasks like text classification, sentiment analysis, and machine translation.

PhoBERT is a cutting-edge pre-trained language model specifically designed for the Vietnamese language. Developed by the Research Institute of AI (AI2) and VinAI, PhoBERT is based on the powerful BERT architecture, fine-tuned on a massive amount of Vietnamese text data [13]. It leverages the Transformer model to learn contextual word embeddings, enabling it to capture complex linguistic nuances and semantic relationships in Vietnamese sentences. PhoBERT is capable of handling various NLP tasks such as text classification, named entity recognition, and sentiment analysis.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. Experiment Design

Within this investigation, we conducted experiments involving widely utilized Bag-of-Words (BoW) based weighting techniques, including count vectorizer and TF-IDF. These methodologies were coupled with a prevalent machine learning classifier, the Naïve Bayes (NB) classifier, commonly employed in text classification scenarios.

The spam detection process involves collecting a dataset of spam and legitimate messages, preprocessing the data, extracting numerical features, and building a model using machine learning techniques. This process is repeated for both the training and testing phases, with separate datasets, to develop and evaluate a reliable model for identifying spam messages accurately. The whole process was shown in Figure 1.

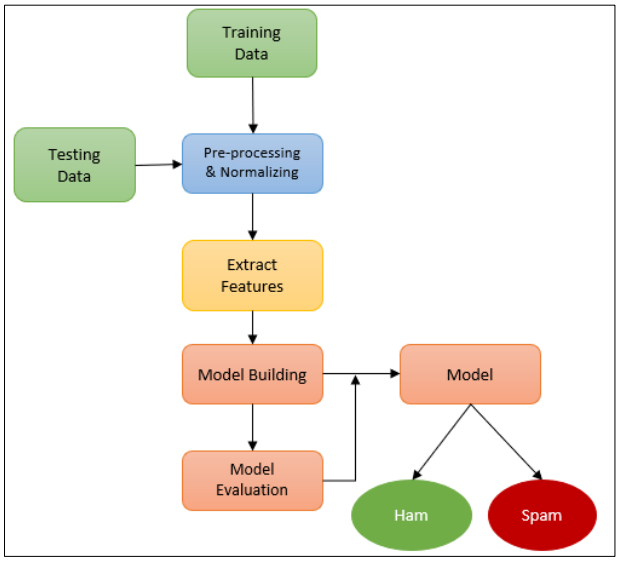

Figure 1 Traditional machine learning model

For the contextual word representation techniques including Word2Vec, GloVe, fastText and PhoBERT, the Convolutional Neural Network (CNN) was deployed to evaluate the performance of the Vietnamese SMS spam detection model.

The CNN structure used in the experiment consists of an input layer receiving preprocessed word embedding feature vectors (Figure 2). The process begins with the use of word embeddings, such as Word2Vec, GloVe, fastText, or PhoBert, to represent words in the input text, capturing semantic meaning and contextual information. Before feeding the word embeddings into the CNN, the input data undergoes preprocessing and normalization, including tokenization and handling special characters Multiple convolutional layers extract local features from the input, followed by activation functions like ReLU to introduce non-linearity. Max pooling layers down sample the feature maps, and a flatten layer prepares the data for fully connected layers. The fully connected layers perform high-level feature extraction and decision-making for binary classification. The output layer, with a sigmoid activation function, produces the probability of spam or non-spam. CNN is trained using backpropagation with binary cross-entropy loss to optimize model parameters. The architecture

is tuned during experimentation to achieve optimal performance in spam detection.

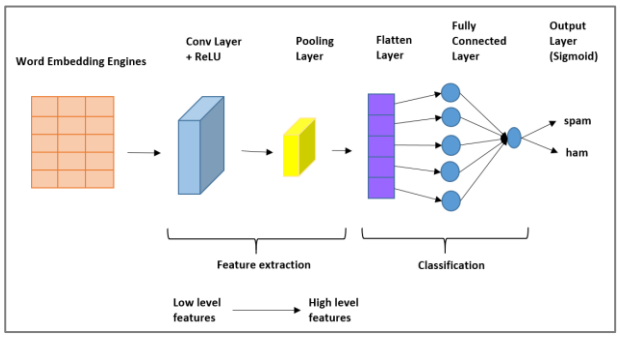

Figure 2 Spam Detection Model with CNN and Word Embedding Techniques.

B. Evaluation Metrics

The experiment employs four evaluation metrics: accuracy, precision, recall, and F1-score. Accuracy measures overall correctness, precision evaluates true positives within predicted positives, recall assesses true positives within actual positives, and the F1-score balances precision and recall for a comprehensive performance assessment in spam message classification.

$$Accuracy\ (A) = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision\ (P) = \frac{TP}{TP + FP} \quad (2)$$

$$Recall(R) = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - Score = \frac{2 * P * R}{P + R} \quad (4)$$

C. Results and Discussion

The experimental results of utilizing Bag-of-Words (BoW) and TF-IDF with the Naïve Bayes classifier showcase distinct performance characteristics. BoW achieves an accuracy of 0.913, precision of 0.892, recall of 0.879, and an F1 score of 0.885. In contrast, the TF-IDF approach demonstrates higher performance, with an accuracy of 0.932, precision of 0.906, recall of 0.892, and an F1 score of 0.899. The TF-IDF technique outperforms BoW in all aspects, showcasing its enhanced ability to classify spam messages effectively (*Table 2*).

Table 2 Performance of BoW and IF-IDF with Naive Bayes Classifier

	Acc.	Pre.	Rec.	F1-Score
Bow	0.913	0.892	0.879	0.885
TF-IDF	0.932	0.906	0.892	0.899

Moving to word embedding techniques integrated with a Convolutional Neural Network (CNN), the word2Vec approach demonstrates high accuracy at 0.958 and an impressive F1 score of 0.921, showcasing its potential. GloVe achieves an accuracy of 0.896 and an F1 score of 0.897, while fastText performs well with an accuracy of 0.931 and an F1 score of 0.917. However, the standout performer is PhoBert, attaining the highest accuracy of 0.968 and a remarkable F1 score of 0.941. This underscores the potency of PhoBert's contextual representation, which effectively captures intricate linguistic nuances and relationships within Vietnamese SMS messages (Figure 3). The significantly elevated performance of PhoBert underscores its capability to comprehend the subtleties of the Vietnamese language, thus positioning it as a superior choice for enhancing accuracy in SMS spam detection tasks.

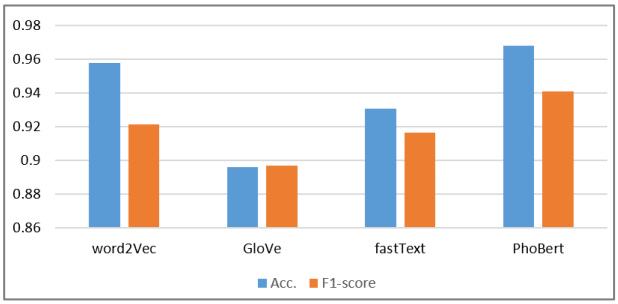

Figure 3 Performance of Contextual Word Embedding Techniques with CNN

The comparison of the two groups reveals the varied performance of techniques in Table 3. TF-IDF surpasses BoW with higher accuracy (0.932 vs. 0.913) and precision (0.906 vs. 0.892). Word embedding with CNN sees word2Vec scoring well (accuracy: 0.958, F1: 0.921), followed by GloVe (accuracy: 0.896, F1: 0.897), and fastText (accuracy: 0.931, F1: 0.917). PhoBert stands out with the highest accuracy (0.968) and an impressive F1 score (0.941), showcasing its capacity to capture nuances in Vietnamese SMS, making it a top choice for enhancing SMS spam detection accuracy.

Table 3 Performance comparison of all word representation techniques

	Acc.	Pre.	Rec.	F1-score
Bow	0.913	0.892	0.879	0.885
TF-IDF	0.932	0.906	0.892	0.899
word2Vec	0.958	0.926	0.917	0.921
GloVe	0.896	0.893	0.901	0.897
fastText	0.931	0.935	0.899	0.917
PhoBert	0.968	0.943	0.939	0.941

V. CONCLUSION

This paper introduces a comprehensive comparative analysis of widely used word embedding techniques across an array of classifiers to detect Vietnamese SMS spam. Our study encompasses two statistical embedding approaches, namely TF-IDF and BOW, along with four contextual techniques: word2vec, fastText, GloVe, and PhoBert. In a broad context, our findings reveal that the amalgamation of contextual word embedding techniques with Convolutional Neural Networks (CNN) yields notable performance advantages over the utilization of statistical word embedding methods with the Naive Bayes classifier. Within the category of contextual techniques, PhoBert emerges as a standout performer due to its unique ability to adeptly capture complex linguistic nuances and intricate semantic relationships embedded within Vietnamese sentences, resulting in superior performance compared to its counterparts.

Additionally, we are eager to extend our research by exploring alternative deep learning algorithms, and further refining and optimizing our methodologies to continuously elevate the efficacy of our approach. This forward-looking perspective underscores our commitment to advancing the realm of SMS spam detection through ongoing exploration and innovation.

REFERENCES

[1] Gauri Jain, Manisha Sharma, Basant Agarwal, "Spam Detection on Social Media Using Semantic Convolutional Neural Network," International Journal of Knowledge Discovery in Bioinformatics, vol. 8, no. 1, pp. 12 - 26, 2018.

[2] Sreekanth Madisetty, Maunendra Sankar Desarkar, "A Neural Network-Based Ensemble Approach for Spam Detection in Twitter," IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS, vol. 5, no. 4, pp. 973 - 984, 2018.

[3] Hyun-Young Lee, Seung-Shik Kang, "Word Embedding Method of SMS Messages for Spam Message Filtering," in International Conference on Big Data and Smart Computing (BigComp), Kyoto, Japan , 2019.

[4] "SMS Spam Classification–Simple Deep Learning Models With Higher Accuracy Using BUNOW And GloVe Word Embedding," Journal of Applied Science and Engineering, vol. 26, no. 10, pp. 1501-1511, 2022.

[5] Neelam Choudhary, Ankit Kumar Jain , "Towards Filtering of SMS Spam Messages Using Machine Learning Based Technique," Advanced Informatics for Computing Research, vol. 712, pp. 18-30, 2017.

[6] P. Poomka, W. Pongsena, N. Kerdprasop, K. Kerdprasop, "SMS Spam Detection Based on Long Short-Term Memory and Gated Recurrent Unit," International Journal of Future Computer and Communication, vol. 8, no. 1, pp. 11-15, 2019.

[7] Luo GuangJun, Shah Nazir, Habib Ullah Khan, Amin Ul Haq, "Spam Detection Approach for Secure Mobile Message Communication Using Machine Learning Algorithms," Security and Communication Networks, vol. 11, pp. 1-6, 2020.

[8] T. Huang, "A CNN Model for SMS Spam Detection," in 4th International Conference on Mechanical Control and Computer Engineering (ICM- CCE), Hohhot, China, 2019.

[9] A. Aizawa, "An information-theoretic perspective of tf–idf measures," Information Processing & Management, vol. 39, no. 1, pp. 45-65, 2003.

[10] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, "Efficient Estimation of Word Representations in Vector Space," Computation and Language, 2013.

[11] Jeffrey Pennington, Richard Socher, Christopher Manning, "Glove: Global vectors for word representation," in Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 2014.

[12] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, Armand Joulin, "Advances in pre-training distributed word representations," in the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 2018.

[13] Dat Quoc Nguyen, Anh Tuan Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," in Findings of the Association for Computational Linguistics: EMNLP 2020, 2020.

ĐÁNH GIÁ KĨ THUẬT NHÚNG TỪ SỬ DỤNG TRONG MÔ HÌNH PHÁT HIỆN TIN NHẮN RÁC TIẾNG VIỆT

Tóm tắt: Vấn nạn tin nhắn rác SMS tiếng Việt đã thúc đẩy việc áp dụng các mô hình học máy và học sâu để phát hiện hiệu quả. Bài báo này nghiên cứu tác động của các kỹ thuật nhúng từ trong việc cải thiện mô hình phát hiện tin nhắn rác SMS. Các phương pháp thống kê truyền thống (BoW, TF-IDF) được so sánh với các kỹ thuật tiên tiến (Word2Vec, fastText, GloVe, PhoBERT) sử dụng một tập dữ liệu riêng. PhoBERT kết hợp với mô hình CNN đã thể hiện độ chính xác cao nhất là 0.968 và F1 score ấn tượng là 0.941. Nghiên cứu này thể hiện rõ vai trò của các kĩ thuật nhúng từ trong việc xây dựng các mô hình phát hiện rác mạnh mẽ, là cơ sở quan trọng cho việc lựa chọn mô hình. Phương pháp triển khai, các phân tích đánh giá và hướng phát triển trong tương lai được trình bày chi tiết trong bài báo.

Từ khóa: Tin nhắn rác tiếng Việt, nhúng từ, học sâu, CNN.

Vu Minh Tuan graduated from Hanoi University in 2010. He completed his Information System Design MSc at the University of Central Lancashire, UK. His research interests include but not limited to spam detection, machine learning, system analysis and design…

Email: minhtuan_fit@hanu.edu.vn

Do Thuy Duong is working at the Faculty of Information Technology, Hanoi University. She completed her Master program at University of Engineering and Technology, Vietnam National University. Her research topics involve natural language processing, spam filtering, machine learning....

Email: duongdt@hanu.edu.vn

Tran Quang Anh is currently the Vice President of Posts and Telecommunications Institute of Technology. He completed his Master and Doctoral programs at Tsinghua University, China. His research areas include network security, evolutionary algorithms, anti-spam.