Adversarial Spam Generation Using Adaptive Gradient-Based Word Embedding Perturbations

Jonathan Gregory and Qi Liao
Department of Computer Science
Central Michigan University, USA
Email: {grego3j, liao1q}@cmich.edu

***Abstract*—In recent years, artificial intelligence (AI) and machine learning (ML) have become extremely promising in almost every aspect of our lives, including in cybersecurity. For instance, intrusion detection systems (IDS) and spam filters use machine learning algorithms to constantly monitor networks for abnormal behavior. However, the security of AI/ML-based solutions remains largely unknown and a cause for concerns. This study examines the possibility of cheating AI/ML-based cybersecurity solutions such as spam filters. In particular, we developed an Adaptive Gradient-based Word Embedding Perturbations (AGWEP) framework for automatically generating adversarial spam examples. AGWEP smartly chooses the optimal perturbations across all features of the word vectors to minimize the degree of modifications to real spam messages. The experimental study suggests the adversarial model is effective to generate meaningful adversarial examples to fool a CNN-based spam classifier.**

I. Introduction

While forms of electronic communication like email, text, and SMS messaging have greatly increased the ease of communication in the modern world, these advances have also brought negative side effects – namely, the proliferation of unwanted electronic messages, or spam. Unlike normal, "ham" messages, spam messages pose a severe risk to cybersecurity, seeking to deceive users into revealing personal information, installing malware, or otherwise exploiting the recipient of the spam. Therefore, spam generation and distribution must be suppressed – if not eliminated – for the safety, security, and productivity of users.

To counteract spam threats, the most common approach is to use a spam filter that is often based on machine learning (ML) models such as Naïve Bayesian classifiers, Support Vector Machines, and more recently, deep learning algorithms that have been adapted to the task of differentiating spam from ham messages [1]. However, these spam filters are not altogether impervious to attacks by adversaries. For instance, researchers demonstrated that by applying a few minimal changes to spam examples, such as replacing certain words indicative of spam with benign synonyms, a Naïve Bayesian classifier could be made to overlook a spam sample without significantly changing the spam's payload [2].

Since more and more cybersecurity solutions are based on machine learning and artificial intelligence, it is critical to evaluate the security of such approaches. Is it possible to fool the AI/ML-based solutions? It would be interesting to explore the possibility of automatically generating adversarial examples for AI/ML-based cybersecurity models such as spam filters. In this research, it is of our interest to find ways to autonomously generate adversarial variations of existing spam that will fool a spam classifier. After all, it is only through knowing the weaknesses of an AI/ML model that these weaknesses can be patched. For this reason, this research investigates the application of automatic, machine learning-based methods for generating adversarial spam examples. Specifically, this research focuses on using an adaptive variant of the Fast Gradient Signed Method (FGSM) adversarial technique, or Adaptive Gradient-based Word Embedding Perturbations (AGWEP), to slightly perturb the embedded word vectors of spam sequences. Essentially, the goal of this adversarial technique is to "poke" the embedded word vectors of a spam sequence in a meaningful direction so that the new placement of these vectors will both confuse a spam filter and remain as close as possible to the original spam sequence. Note that this problem of perturbing existing spam examples is strictly simpler than the more complicated Natural Language Generation (NLG) task of generating entirely new spam examples.

The initial results of experimental study indicate the automatically generated adversarial spam resulted in an evasion rate of 43% and most of the adversarial examples are meaningfully perturbed and keep all critical features of spam messages. While the findings demonstrate the effectiveness of the proposed adversarial spam generative model as a method to circumvent spam classifiers, more research is necessary to determine more effective and universal ways of finding the ideal perturbations for spam examples.

II. Related Work

Artificial intelligence and machine learning has much success in cybersecurity solutions. However, research suggests that the machine learning based approaches in spam detectors may suffer adversary attacks during the training or the prediction phase [3], a practice known as adversarial machine learning [4]. A survey reveals a new type of adversarial spam that can attack against online social networks such as Twitter spam detectors [3]. An attacker may target at high-value features with the knowledge of email distribution [5].

Previous works focus on generating adversarial examples for images by adding perturbations using Fast Gradient Sign Method (FGSM) [6] and Generative Adversarial Network (GAN) [7]. As suggested by Goodfellow et al. [6], machine

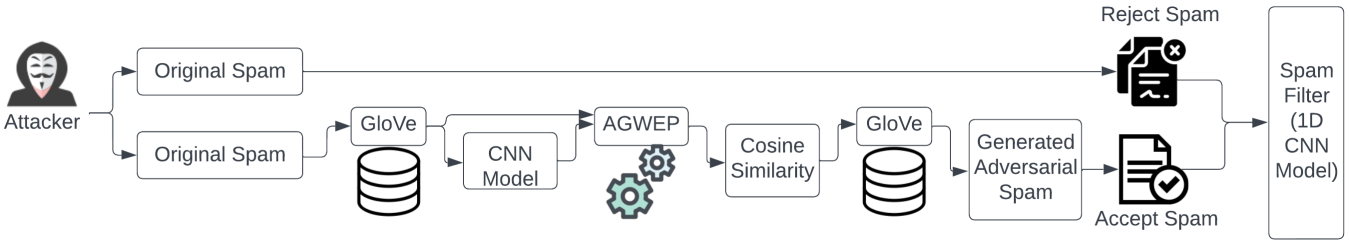


Fig. 1: System overview of creating adversarial spam sequences using Adaptive Gradient-based Word Embedding Perturbations (AGWEP).

learning models that behave linearly, such as LTSMs, ReLUs, and sigmoid networks, can be fooled by adversarial examples generated by perturbing inputs with a fast, straightforward method of perturbation. FGSM attack may be feasible against differentiable machine learning models like recurrent neural networks (RNNs) when using continuous inputs, such as word vectors in an embedding space [8]. Further, it has been demonstrated that another gradient-based method to perturb word vectors, i.e., the Forward Derivative method, calculated the direction to perturb the elements of a word vector using the vector's Jacobian tensor, and could effectively fool an RNN and allow the mapping of a perturbed embedded sequence back to English [8]. Similarly, a study [9] introduced interpretable adversarial perturbations for embedded text sequences, which perturbed embedded text using a version of Goodfellow et al.'s adversarial method for continuous spaces but constrained these perturbations only to the direction of other words in the embedding space. Moreover, researchers demonstrated that for better interpretation of these adversarial examples, the perturbed word embeddings in a sentence could be mapped back to discrete words and displayed as intelligible sentences.

In addition, optical character recognition (OCR) systems have been found vulnerable to modifications of image pixels or character-level perturbations that result in OCR models misclassifying the embedded text [10]. A black-box model based Word Substitution Ranking Attack (WSRA) [11] has been proposed against neural ranking models (NRMs) to promote a target document in rankings by adding adversarial perturbations to its text. Adversarial email may be generated with "magic words" [12]. Adversarial machine learning on spam filters using techniques such as synonym replacement, ham word injection and spam word spacing were proposed [2], but much of the work is a manual process rather than automated.

Lastly, the use of Global Vectors (GloVe) [13] in detecting malicious activities such as spams and other related tasks has been documented, e.g., for classifying online hate speech [14]; for recognizing phishing attacks on web pages [15]; and for helping classify Twitter spam [16]. As such, GloVe both simplifies classification tasks and enables meaningful word perturbation, making them of special use in this research.

III. Adversarial Spam Generative Model

This sections discusses the key algorithm and methodology of the proposed adversarial spam generative model. An overview of the system is illustrated in Figure 1. Traditionally, a network administrator sets up a spam filter based on up-to-date machine learning algorithms such as a Convolutional Neural Network (CNN) model, which effectively rejects nearly all spam messages from an attacker. Under the adversarial model, an attacker would process the spam sequence in the following way. First, GloVe is applied to map spam messages to word vectors, which are then perturbed using AGWEP algorithm. Then, the perturbed vectors are translated back to words using similarity measures to generate new adversarial spam samples, which will successfully be accepted and bypass the spam filter.

A. Word Embedding with Global Vectors

To be processed by a machine learning model, discrete inputs such as words in a spam message must be translated into a machine-readable format. While simple methods for translating text to numerical inputs exist, many of these approaches for vectorizing text sequences, such as using a one-hot encoding or bag-of-words technique, lack an awareness of the relationships between words. Further, these simple, discrete methods do not contain a continuous, navigable space to traverse when applying perturbations to vectorized text sequences. As such, a more complex, context-aware vectorization scheme is necessary for this research. The GloVe model for word embeddings fills this role by providing an embedding space for word vectors that captures word relationships using calculations based on the global word co-occurrences in a large corpus of text data. Since the embedding space developed in the GloVe model is continuous, later perturbations on the word vectors in this space are therefore meaningful, and perturbed word vectors in this embedding space can be mapped to nearby words using measures of vector similarity like Euclidean distance or cosine similarity.

B. Adaptive Adversarial Gradient-based Perturbations

To begin the perturbation process for spams, this research first experiments with the baseline approach using Fast Gradient Sign Method (FGSM) that was originally designed for perturbing image pixels. From preliminary tests, the confidence with which the spam filter classified these adversarial examples

as spam was drastically lower than their unperturbed counterparts. Interestingly, these perturbations sometimes pushed certain spam examples into a regime that the spam classifier could detect more easily as spam, which was detrimental. However, subtracting, rather than adding, the perturbations from the embedded sequence was an effective way to reverse this phenomenon and once again generates adversarial examples that could fool the spam detector.

A key limitation of this baseline approach is that the perturbations are applied evenly across all features of the word vectors. Specifically, the perturbation to each index in a word vector will be the same regardless of the magnitudes of the raw gradients at those indices. These large-grained perturbations are thus insensitive to the nuances of the gradients on which these adversarial examples depend and may deliver suboptimal results. A natural solution to this problem would be to add the true gradient at each index (not just the sign of this value) to its corresponding index in each word vector. Unfortunately, an examination of these gradients reveals that they are typically minuscule in magnitude, so a perturbation using the pure gradients would essentially leave no perceivable change to an embedded text sequence.

The solution developed in this research is to allow the largest gradient value in magnitude for each word vector to be adaptively set to an absolute value of 1 (the largest value possible using the base FGSM approach) by multiplying this value by its reciprocal. All lesser gradient values are then multiplied by this reciprocal, thus preserving the measure of scale in the gradients. Since the largest values of the gradients are still close to 1, this approach maintains FGSM's core ability to create perturbations that can measurably disrupt an embedded text sequence without causing too much change in the sample. However, in testing this approach, a larger constant of multiplication, denoted in the following algorithm as ϵ, was necessary to perturb the embedded spam to a reasonable extent. This research experimentally defines ϵ equal to 1.5 to allow the spam to fool the spam detector but preserve its original structure and purpose. However, other reasonably sized constants should work as well. A formal definition of this Adaptive Gradient-based Word Embedding Perturbations (AGWEP) approach for creating adversarial examples is given in Algorithm 1. To the best of the authors' knowledge, this specific approach to perturbing spam examples is a novel application of the overarching FGSM algorithm.

Algorithm 1 Adaptive Gradient-based Word Embedding Perturbations (AGWEP)

procedure AGWEP(model, spam, label)
  $grads \leftarrow getFGSMGradients(model, spam, label)$
  for $i \leftarrow 0$ to $spam.length - 1$ **do**
    if $spam[i]$ is not a zero vector **then**
      $grads[i] \leftarrow \epsilon * 1/\max|grads[i]| * grads[i]$
      $spam[i] \leftarrow spam[i] + grads[i]$
    end if
  end for
  if model.predict(spam) is less than label **then**
    revert spam to original values
    for $i \leftarrow 0$ to $spam.length - 1$ **do**
      if $spam[i]$ is not a zero vector **then**
        $grads[i] \leftarrow \epsilon * 1/\max|grads[i]| * grads[i]$
        $spam[i] \leftarrow spam[i] - grads[i]$
      end if
    end for
  end if
end procedure

C. Word Translation with Similarity Measure

Importantly, after adaptive adversarial gradient-based perturbation techniques, generating raw perturbed examples is not the end. In order to be properly evaluated, these perturbed spam examples must be translated back to English. While these embedded word vectors no longer map to specific words once perturbed, by iterating through all known words in the GloVe embedding space and comparing the similarity between every known word vector and each perturbed vector, the perturbed vector may be cast to its closest neighbor. As aforementioned, cosine similarity or Euclidean distance can be used for this similarity calculation. Specifically, the cosine similarity is defined as follows:

$$(P * W)/(||P|| * ||W||)$$

where P is the perturbed word vector and W is the vector for a word in the GloVe embedding space. The perturbed word vector is then translated back to the word in the GloVe vocabulary whose vector yields the greatest cosine similarity to P.

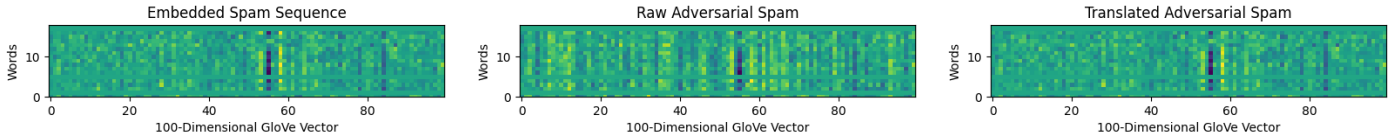
As an example. Figure 2 illustrates visualizations of a spam message before perturbations (Figure 2a), after perturbations (Figure 2b), and after translations (Figure 2c). When comparing Figures 2a and 2c, one can clearly see how the adaptive adversarial generative model yields a perturbed example that is visually similar to the original spam sequence.

IV. Evaluation

This section discusses the implementations, dataset and machine learning model used to train spam filter. Most importantly, experiments are conducted to evaluate the effectiveness of the proposed generative adversarial system in terms of various metrics. A comparison of original and generative adversarial examples are also examined.

A. Data, Model and Implementations

For experimental study, a SMS dataset [17] from the University of California - Irvine, is used for testing. The dataset contains 5,574 messages, labeled either as ham (legitimate) or spam, which were converted to 0 for ham and 1 for spam. To obtain a balanced dataset, the ham subset was shuffled and trimmed to the same size as the spam subset, which contained 747 samples. These subsets were then recombined into a single balanced SMS dataset, shuffled, and split into training and testing samples using an 80-20 split.



(a) Original spam message (b) Raw adversarial spam message after perturbations (c) Final adversarial spam message after translations

Fig. 2: Visualizations of a spam message: (a) original spam "Dear U've been invited to XCHAT. This is our final attempt to contact u! Txt CHAT to 86688", which is classified as spam with a score of 0.999889; (b) the raw adversarial spam after adaptive perturbation algorithm; (c) the translated adversarial embedded spam sample for the raw sequence. The sequence now reads "dear uve been invited to xchat this is our final attempt to contact u unzip chat did 86688", which is classified as ham with a score of 0.007325036.

The SMS sequences in these sets were all converted to lowercase, stripped of their punctuation, split into tokens using whitespace as a delimiter, vectorized, and truncated or padded to sequences of length 30 (a length that accommodated most SMS samples in the dataset) with the TextVectorization layer from the Keras API. The vectorized words in each sequence were then translated into 100-dimensional word vectors using the Stanford NLP pretrained GloVe word embeddings trained on Wikipedia and Gigaword data.

Following the conversion of the SMS sequences to matrices of word embedding vectors, a machine learning (ML) classifier was prepared for the task of differentiating spam from ham sequences. Although several different ML options exist for spam classification, a 1D Convolutional Neural Network (CNN) text classification model was chosen for this task due to its generality and popularity. This model was compiled using a binary cross-entropy loss function and an Adam optimizer and was trained on the training ham and spam samples for 15 epochs with validation data to monitor overfitting. The resulting model performed well on the test data, misclassifying only 19 out of 299 SMS samples with a precision measure of 0.96 and a recall measurement of 0.92 with a decision threshold of 0.5. Most importantly, this model only classified 13 of the 167 test spam examples as false negatives, indicating it can detect spam with reasonable accuracy and is therefore a good model to test adversarial spam examples against.

B. Experimental Results and Discussions

We ran the adaptive adversarial generation method for perturbing spam examples on 167 test spam sequences. The decision boundary for considering an example as spam or ham was set at 0.5, as is a standard threshold for binary classification. After perturbations, the raw adversarial spam sequences were automatically converted back to English, and tested on the spam classifier. The results suggest that the spam classifier is lured into classifying 72 of these adversarial examples as false negatives, i.e., as ham. In other words, these adversarial examples decreased the recall of the model from 0.92 to 0.57, a noticeable and intriguing drop of 0.35 in recall. Overall, whereas the original spam resulted in a spam evasion success rate of about 0.08, the perturbed adversarial spam resulted in an evasion rate of 0.43, an increase of 0.35 in spam evasion effectiveness.

Additionally, this research qualitatively evaluated whether the resulting spam that did fool the spam detector were still meaningful and effective following their perturbations and translations – that is, whether these messages were intelligible, reasonably close to their unperturbed versions, and still carried the necessary information to manipulate the recipient of the spam. From this manual evaluation, 27 of the 59 successfully perturbed spam sequences (omitting the 13 spam sequences that were originally misclassified by the model) were considered adequate in this regard. Table 1 shows a sample of the best adversarial spam sequences resulting from this research.

As the results indicate, this method for perturbing spam was effective on this SMS dataset. Moreover, as the results in Table I show, the perturbed spam translated back to English is reasonably close to the original spam examples by keeping all critical information of the original spam messages. For instance, consider the following spam example, which was marked as spam by the classifier with a score of 0.9995413: "Hi this is Amy, we will be sending you a free phone number in a couple of days, which will give you an access to all the adult parties...." The system converted this sequence into the following, which was classified as ham with a score of 0.0017231683: "hi this is amy we will be sending i it not phone number in a couple of days which we give you an access to all the adult parties." (Note that the differences in punctuation and capitalization between the original and perturbed spam are due to the preprocessing necessary to convert the spam to GloVe vectors.) In this case, changing the words "you" to "i," "a" to "it," "free" to "not," and "will" to "we" was enough to turn this sequence from spam to ham. Moreover, although the grammar of this sentence is disrupted by these changes, the core original message of the spam (that someone has an advantageous resource that they would like to distribute) is retained despite these alterations. Although portions of this spam sequence degrade with these perturbations, the fact that the adversarial spam keeps its message but fools the spam classifier is nonetheless impressive.

The perturbations were not always successful, however. For instance, the original spam "You have 1 new message. Call 0207-083-6089.", which was confidently marked as spam with a score of 0.9999785, was translated into an adversarial example "i have 1 it i i 02070836089.", which was marked as ham with a score of 1.8200574e-10. Note that this ad-

TABLE I: Generated Adversarial Spam Examples

Original Spam (Detected)	**Adversarial Spam (Undetected)**
3. You have received your mobile content. Enjoy	2 i have received my mobile content enjoy
Hi this is Amy, we will be sending you a free phone number in a couple of days, which will give you an access to all the adult parties...	hi this is amy we will be sending i it not phone number in a couple of days which we give you an access to all the adult parties
Someone U know has asked our dating service 2 contact you! Cant Guess who? CALL 09058097189 NOW all will be revealed. POBox 6, LS15HB 150p	someone u know has asked our dating it 2 contact i cant guess who they 09058097189 now all will be revealed pobox 6 ls15hb 150p
dating:i have had two of these. Only started after i sent a text to talk sport radio last week. Any connection do you think or coincidence?	datingi have had two of these only started after i sent it text to why sport radio last last because connection do you thing or coincidence
Save money on wedding lingerie at www.bridal.petticoatdreams.co.uk Choose from a superb selection with national delivery. Brought to you by WeddingFriend	save money on wedding lingerie at wwwbridalpetticoatdreamscouk assume from a superb selection with national delivery brought to i by weddingfriend
Twinks, bears, scallies, skins and jocks are calling now. Don't miss the weekend's fun. Call 08712466669 at 10p/min. 2 stop texts call 08712460324(nat rate)	twinks bears scallies skins and jocks are saying now dont miss the weekends fun done 08712466669 at 10pmin 2 it texts did 08712460324nat rate
Check Out Choose Your Babe Videos @ sms.shsex.netUN fgkslpoPW fgkslpo	check out choose my babe videos smsshsexnetun fgkslpopw fgkslpo
Natalie (20/F) is inviting you to be her friend. Reply YES-165 or NO-165 See her: www.SMS.ac/u/natalie2k9 STOP? Send STOP FRND to 62468	natalie 20f is inviting i to be her friend reply yes165 or no165 see her wwwsmsacunatalie2k9 stop could stop frnd way 62468
Dear U've been invited to XCHAT. This is our final attempt to contact u! Txt CHAT to 86688	dear uve been invited to xchat this is our final attempt to contact u unzip chat did 86688
Oh my god! I've found your number again! I'm so glad, text me back xafter this msgs cst std ntwk chg å£1.50	oh my god ive found my number again im so glad text me back xafter this msgs cst std ntwk chg å£150

versarial example is unintelligible compared to its original form, so despite fooling the spam detector, it cannot be called a meaningful adversarial example. Further, some adversarial examples failed even to fool the detector. For example, the adversarial example for "Win a å£1000 cash prize or a prize worth å£5000" only changed "cash" to "money," resulting in both the original and adversarial spam being detected by the spam classifier. Examples like these indicate that this model still requires refinement.

V. Conclusion

Artificial intelligence is promising in many areas including cybersecurity. However, the trustworthiness and security of many AI/ML-based solutions remain challenging. In this study, we designed a new adaptive adversarial model that systematically perturbates word embedding vectors to generate adversarial spam examples. The results demonstrated the viability of such approach and showed that spam filters are susceptible to such adversarial attacks. The findings of this research suggest that further tuning of this method could result in a very effective approach for generating adversarial text examples which will be our future work.

References

[1] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, "Machine learning for email spam filtering: review, approaches and open research problems," *Heliyon*, vol. 5, no. 6, p. e01802, Jun 2019.

[2] B. Kuchipudi, R. T. Nannapaneni, and Q. Liao, "Adversarial machine learning for spam filters," in *Proceedings of the 15th International Conference on Availability, Reliability and Security (ARES) - 15th ACM International Workshop on Frontiers in Availability, Reliability and Security (FARES)*, no. 38, Dublin, Ireland, August 25-28 2020, pp. 1–6.

[3] N. H. Imam and V. G. Vassilakis, "A survey of attacks against Twitter spam detectors in an adversarial environment," *Robotics*, vol. 8, no. 3, 2019.

[4] N. Martins, J. M. Cruz, T. Cruz, and P. H. Abreu, "Adversarial machine learning applied to intrusion and malware scenarios: A systematic review," *IEEE Access*, vol. 8, pp. 35 403–35 419, February 18 2020.

[5] J. PENG and P. P. K. CHAN, "Revised Naive Bayes classifier for combating the focus attack in spam filtering," *Proceedings of the IEEE International Conference on Machine Learning and Cybernetics*, pp. 610–614, july 2013.

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint 1412.6572*, Mar 20 2015.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, November 2020.

[8] N. Papernot, P. McDaniel, A. Swami, and R. Harang, "Crafting adversarial input sequences for recurrent neural networks," in *IEEE Military Communications Conference (Milcom)*, Baltimore, MD, Nov 1 2016, pp. 49–54.

[9] M. Sato, J. Suzuki, H. Shindo, and Y. Matsumoto, "Interpretable adversarial perturbation in input embedding space for text," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, July 2018, p. 4323–4330.

[10] N. H. Imam, V. G. Vassilakis, and D. Kolovos, "OCR post-correction for detecting adversarial text images," *Journal of Information Security and Applications*, vol. 66, no. C, p. 103170, May 2022.

[11] C. Wu, R. Zhang, J. Guo, M. D. Rijke, Y. Fan, and X. Cheng, "Prada: Practical black-box adversarial attacks against neural ranking models," *ACM Transactions on Information Systems*, vol. 41, no. 4, pp. 1–27, April 2023.

[12] Q. Cheng, A. Xu, X. Li, and L. Ding, "Adversarial email generation against spam detection models through feature perturbation," in *IEEE International Conference on Assured Autonomy (ICAA)*, Fajardo, PR, March 22-24 2022, pp. 83–92.

[13] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, Oct 25-29 2014, pp. 1532–1543.

[14] N. Badri, F. Kboubi, and A. H. Chaibi, "Combining fasttext and glove word embedding for offensive and hate speech text detection," *Procedia Computer Science*, vol. 207, pp. 769–778, 2022.

[15] E. Benavides-Astudillo, W. Fuertes, S. Sanchez-Gordon, D. Nuñez-Agurto, and G. Rodríguez-Galán, "A phishing-attack-detection model using natural language processing and deep learning," *Applied Sciences*, vol. 13, no. 9, April 2023.

[16] S. Madisetty and M. S. 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, December 2018.

[17] U. M. Learning, "Sms spam collection dataset," https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset, [Online; accessed 20-April-2023].