Chapter 9
Adapting to Change:
Assessing the Longevity and Resilience of Adversarially Trained NLP Models in Dynamic Spam Detection Environments

Mahmoud Basharat
Capitol Technology University, USA & Houston Community College, USA

Marwan Omar
Capitol Technology University, USA & Illinois Institute of Technology, USA

ABSTRACT

The rapid evolution of cyber threats in digital communication necessitates robust and adaptive natural language processing (NLP) models, especially for spam detection. This chapter explores the effectiveness and sustainability of adversarial training in NLP models within dynamic spam detection contexts. The authors investigate how adversarially trained models illustrate the concept drift phenomenon. The findings reveal significant insights into the limitations and potential of adversarial training, providing a nuanced understanding of its long-term implications in real-world deployment scenarios. This research contributes to the broader understanding of NLP model resilience, emphasizing the necessity of continuous model evolution to maintain efficacy in changing cyber environments.

INTRODUCTION

The rapid advancements in Natural Language Processing (NLP) and the widespread deployment of its applications have ushered in a new era of challenges, particularly in the context of cybersecurity. Among these challenges, the resilience of NLP models against adversarial attacks in dynamic environments, specifically in spam detection, stands out as a critical area of concern. Adversarial training, initially a beacon of hope for enhancing model robustness, now faces scrutiny under the evolving landscapes of data and cyber threats (Goodfellow, Shlens, & Szegedy, 2014).

DOI: 10.4018/979-8-3693-1906-2.ch009

Spam detection, a longstanding and essential task in cybersecurity, has evolved from simple rule-based systems to sophisticated NLP models leveraging deep learning algorithms like BERT, RoBERTa, and CNERG (Barbieri, Camacho-Collados, Espinosa-Anke, & Neves, 2020; Mathew, Saha, Saha, & Mukherjee, 2020). While these models have shown remarkable accuracy in classifying spam, their robustness against adversarial attacks remains a vital concern (Zhu, Cheng, Gan, Sun, Goldstein, & Liu, 2019). Adversarial attacks, by subtly altering input data, can deceive these models, leading to serious security breaches (Goodfellow et al., 2014).

The concept of adversarial training emerged as a promising solution to this problem. It involves integrating adversarially generated examples into the training process, aiming to prepare the model for potential attacks (Zhou, Jiang, Chang, & Wang, 2019). This approach has shown success in several studies, where models trained with adversarial examples exhibited improved resistance against attacks (Dinan, Humeau, Chintagunta, & Weston, 2019; Jin, Jin, Zhou, & Szolovits, 2020). However, the long-term effectiveness of adversarial training in real-world scenarios, where models continuously encounter new, non-adversarial data, remains underexplored.

Recent literature indicates that while adversarial training initially enhances model robustness, its effectiveness may erode over time as the model interacts with new data types and distributions (Morris, Lifland, Yoo, Grigsby, Jin, & Qi, 2020). This phenomenon, known as 'concept drift', refers to the changes in the statistical properties of the target variable, which could lead to a decline in model performance (Lu, Liu, Dong, Gu, Gama, & Zhang, 2018). In the domain of spam detection, this is particularly relevant as spammers continually devise new strategies, causing the characteristics of spam to evolve.

This paper aims to investigate the temporal erosion of adversarial training's impact on NLP models in spam detection tasks. We hypothesize that while adversarial training initially improves model resilience, its benefits diminish as models encounter newer data types and distributions. This hypothesis is tested through extensive experiments using state-of-the-art NLP models like BERT, RoBERTa, and CNERG across various benchmark datasets like Enron spam, SMS spam, and Ling spam (Barbieri et al., 2020; Aluru, Mathew, Saha, & Mukherjee, 2020). By analyzing these models' performance over time and against evolving data distributions, we aim to uncover the limitations of adversarial training in realistic deployment scenarios.

Our study contributes to the field by providing a nuanced understanding of adversarial training's limitations in dynamic environments, a topic that remains relatively unexplored in current literature. Previous works have primarily focused on the immediate robustness against adversarial samples (Zhu et al., 2019; Jin et al., 2020), but have not adequately addressed potential degradation in model performance over time, particularly in realistic spam filter deployment scenarios.

Understanding the temporal dynamics of adversarial training is crucial for practitioners and researchers aiming to maintain resilient NLP models in an ever-changing cyber threat landscape. This research not only sheds light on the complexities of maintaining robust NLP models but also provides insights into optimizing strategies for deploying these models in real-world scenarios. By investigating the long-term implications of adversarial training, we contribute valuable insights to the field of cybersecurity and NLP, highlighting the need for continuous model adaptation in the face of evolving data and threats.