# Vulnerabilities Assessment of Deep Learning-based Fake News Checker under Poisoning Attacks

Lelio Campanile[1], Pasquale Cantiello[2], Mauro Iacono[1], Fiammetta Marulli[1], and Michele Mastroianni[1]

[1] Università degli Studi della Campania ”L. Vanvitelli”,
Dept. of Maths and Physics, Caserta, Italy
[2] Istituto Nazionale di Geofisica e Vulcanologia, Napoli, Italy
{lelio.campanile,mauro.iacono,fiammetta.marulli,
michele.mastroianni}@unicampania.it pasquale.cantiello@ingv.it

**Abstract.** In this work, we envise an effective case study concerning a data and a model poisoning attack, consisting in evaluating how much a poisoned word embeddings model could affect the reliability of a deep neural network-based Fake News Checker; furthermore, we plan to train three different word embeddings models among the most performing in the Natural Language Processing field, in order to investigate which of these models can be considered more resilient and robust when such kind of attacks are applied.

**Keywords:** Natural Language Processing· Fake News· Adversarial Attacks· Data Poisoning Attacks· Deep Neural Networks Resilience

## 1 Introduction

Adversarial machine learning (AML) can act by prompting a Deep Neural Network with intentionally manipulated inputs, with the aim of fooling it and reducing its accuracy in accomplishing its task.
Adversarial examples, that represent purposely designed inputs able to exploit vulnerabilities of machine and deep learning models, got great popularity in performing adversarial attacks against image classification systems[3].
As for the Natural Language Processing (NLP) field, adversarial attacks are more difficult to be acted: while an image perturbation can lead to small changes of pixel values, that are hardly perceived by human eyes, small perturbations applied on texts are easier to sense. A text perturbation could consist, for example, in replacing characters or words within a sentence, thus generating invalid words, meaningless sequences or syntactically incorrect sentences. Therefore, perturbations on natural language texts are easily perceived.

[3] Goodfellow, I. J., Shlens, J., Szegedy, C. (2014). Explaining and harnessing adversarial examples.

## 2 The Envised Case Study

Recent studies have evidenced that also NLP systems can be fooled by AML attacks[4]. More precisely, they are vulnerable to poisoning attacks acted by the means of knowledge transferred, in the shape of pre-processed models, as an input during the training step of a deep learning model While small perturbations expressly introduced in a training data set could be easily recognized, a model poisoning, performed by the means of altered data models and typically provided in the transfer learning step to a deep neural network, are harder to be detected. A notable instance of a scenario is represented by the *poisoned word embeddings models* attack[5]. So, in this work we envise an effective case study to analyze this kind of attack to NLP systems; more precisely, we considered a NLP application of great interest in the current times, provided by the Fake News Detectors (FNDs). These systems typically apply several NLP techniques to analyze the features of a set of news and information published and disseminated, mainly through the web. FNDs represent an ideal target to adversarial attacks, since malicious users are interested to bypass news checker, to go on spreading unnoticed misleading information. We considered a model poisoning attack consisting in poisoning a word embeddings model, used to train a deep neural network-based Fake News Checker; we trained two different word embeddings models, based respectively on fastText[6] and BERT[7] algorithms; we performed experiments both on the trusted and the poisoned versions of these WEs models, in order to investigate the vulnerability level of these NLP systems to these poisoning attacks and the level of resilience of the WEs models to these kinds of traps.

## 3 Conclusion and Future Works

This work explores the effects of a poisoned word embeddings model attack to fool a a deep neural network implementing a Fake News Checker, that generate alerts when potential Fake News are detected. The rise of the False Positives rate, while reducing the reliability of a Fake News Checker, is among the goals of such attacks. We considered, as case study, a Fake News Checker, that performs a stylometric analysis of texts written in natural language, to analyze the severity of a WEs poisoning attack and how much it can affect the accuracy in recognizing real and fake news. We evaluated two different WEs models, to build a vulnerability assessment of NLP systems to poisoning attacks.

The research described in this work is funded and performed within the activities of the the Research Program "Vanvitelli V:ALERE 2020 - WAIILD TROLS", financed by the University of Campania "L. Vanvitelli",Italy.

[4] Zhang, W. E., Sheng, Q. Z., Alhazmi, A., Li, C. (2020). Adversarial attacks on deep-learning models in natural language processing: A survey. ACM Transactions on Intelligent Systems and Technology (TIST), 11(3), 1-41.

[5] Zhao, Z., Dua, D., Singh, S. (2017). Generating natural adversarial examples.

[6] https://fasttext.cc/docs/en/crawl-vectors.html

[7] https://blog.google/products/search/search-language-understanding-bert/