STARLIGHT

Leveraging Continuous Learning for Fighting Misinformation

Evgenia Adamopoulou[1], Theodoros Alexakis[1], Nikolaos Peppes[1], Emmanouil Daskalakis[1], Konstantinos Demestichas[2]

1. Institute of Communication and Computer Systems, School of Electrical and Computer Engineering,
2. Department of Agricultural Economics and Rural Development, Agricultural University of Athens

1. Introduction

The eruption of digitization and the establishment of Social Media as a major content production and reproduction means has led to new paradigms of journalism and news spreading. The rapid changes that took place in the last twenty years led to an environment of pluralism without borders, where, also, may threats are lurking. One of these threats is the rapid spreading of misinformation/disinformation. It is proven that fake news is spreading even to six faster than credible information [1]. This phenomenon consists a major concern firstly for media organizations and professionals as well as for Law Enforcement Agencies (LEAs) due to the fact that the rapid spread of disinformation can severely threaten many aspects of society. According to the European Commission the spread of both disinformation and misinformation can feature a range of harmful consequences, such as the threatening of our democracies, the polarization of debates, and the setting of the health, security and environment of EU citizens at risk [2].

As the practices of misinformation and disinformation evolve it is of utmost importance to design, develop and engage new technologies and solutions in order to tackle such phenomena. In this light, numerous approaches engaging Machine Learning (ML) in order to address this problem from different viewpoints have emerged. Even though, from a technical perspective, several diverse solutions for fake news detection and identification of misinformation, such as transfer learning, multi-task learning, reinforcement learning, online learning etc., do exist, no universal solution to fit in all the aspects of this issue has been developed so far. Almost each and every single solution aims to address the problem in or a very specific topic or domain and based on a limited dataset. The purpose of this study is to present an approach which combines and evaluates the results of different Machine Learning prediction models into a common environment named "Meta-Detection Toolset". This solution relies on the calculation of a meta-score by using weights-based voting among different prediction models, usually referred to as verification services. The weights of the verification services are constantly updated based on the annotation procedure by the end-users of the Toolset. This leverages the current solution into a lifelong learning approach which is future-proof and adaptable as the Machine Learning models improve or aggravate through the course of time or perform better or worse for different topics or styles of writing.

2. Proposed solution

As mentioned, the proposed solution of the Meta-Detection Toolset engages different verification services. These diverse verification services serve as predictors of credibility for a given content source e.g. URL or a text. Based on the integration and implementation of a weighted majority algorithm [3], equal weights are initially assigned to each verification service. During the continuous training process, the weight assigned to a verification service is automatically adjusted according to the accuracy of its predictions. Verification Services with more correct predictions during the training phase, are provided with higher weights, thus playing a more important role when the MDT is calculating the credibility of a certain URL or a text. End-users, e.g. fact-checkers, LEA officers, etc., also play an active role in the training process. More specifically, end-users can insert their credibility evaluation of specific URLs (i.e., indicating whether a specific URL represents legitimate or fake news). The aforementioned users' evaluations are provided in the form of a ground truth label (Legitimate/Fake), are stored in a database, and are utilized during the continuous training phase for updating the weights assigned to each verification service. Thus, a growing number of these annotations will lead to improved verification results of the Meta-Detection Toolset.

The accumulated experience is envisioned to lead to the generation of a model which extensively utilizes contemporary AI technologies for combatting the spread of fake news on the web. This model is comprised of multiple specialized verification services and has the ability to combine verification services based on different methods, aiming to evaluate the truth based on a complex scoring mechanism. This AIbased process is called Meta-Detection and achieves continuous improvement

established by annotation processes performed by specialized end-users.
In the context of the Meta-Detection Toolset, an integrated management environment of the verification services utilized is developed, where the Meta-Detection scores are also determined according to the annotations provided by fact-checkers. More specifically, for a specific URL for example, the annotation of a ground truth label is provided (Legitimate/Fake) by certified fact-checkers. A growing number of these annotations can lead to more accurate verification results in real time. Data ingestion can be achieved either at the end-users' side over the HTTPS protocol or by using data connectors (Kafka topics and/or REST APIs). Then, the input data can be consumed by various verification services integrated in or connected with the toolset. Following the completion of the verification services' computation processes, the prediction results are sent to the MDT and the results are combined in order to compute a meta-score that reflects the credibility of the digital content.

Acknowledgments
The work described in this paper is performed in the H2020 project STARLIGHT ("Sustainable Autonomy and Resilience for LEAs using AI against High priority Threats"). This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 101021797.

References
1. Vosoughi, S.; Roy, D.; Aral, S. The Spread of True and False News Online. Science 2018, 359, 1146–1151, doi:10.1126/science.aap9559.
2. European Commission Tackling Online Disinformation Available online: https://digitalstrategy. ec.europa.eu/en/policies/online-disinformation (accessed on 18 April 2023).
3. Littlestone, N.; Warmuth, M.K. The Weighted Majority Algorithm. Information and Computation 1994, 108, 212–261, doi:https://doi.org/10.1006/inco.1994.1009.

Tools4LEAs

Kriptosare: Behaviour analysis in cryptocurrency transactions

Francesco Zola[1], Jon Elduayen[2], Igor Pallin[1], Raúl Orduna-Urrutia[1]

1. Vicomtech Foundation
2. European Anti-Cybercrime Technology Development Association (EACTDA)

Undoubtedly, the cryptocurrency industry is experiencing rapid innovation and constant evolution, derived from its power and utility. Despite being backed by blockchain technology that promises security, immutability, and full transparency, some cryptocurrencies, Bitcoin imprimis, have been used as enablers for many licit and illicit activities such as trading, buying of goods, money laundering, scam, terrorism financing, ransomware payments, etc. In this scenario, the analysis of the transactions, as well as the entities that have generated them, became a crucial step for Law Enforcement Officer (LEO) investigations. However, the (pseudo) anonymity of the network, the lack of regulatory authority, the employment of anonymizer mechanisms, the evolution of entities' behaviour, and the emergence of new dynamics, are just some of the main elements that make this task challenging. At the same time, the huge amount of information to be analyzed can result in a waste of time and resources, slowing the investigations.
For this reason, in this work, we present Kriptosare, a tool able to classify entity behaviours belonging to three main cryptocurrencies (Bitcoin, Bitcoin Cash, and Litecoin). Kriptosare is composed of a module called Kriptosare.class which makes use of state-of-the-art Machine Learning (ML) techniques to reduce anonymity in the considered cryptocurrencies. This ML model extracts behaviours (or classes) from interactions and dynamics of different known entities involved in the transactions and then predicts the behaviours of new unseen entities. Pre-defined ML models are provided for a first classification, although users can train new ones, and so they can re-classify the whole blockchains. For this task, the blockchain information