Fighting Misinformation, Radicalization and Bias in Social Media

Erica Coppolillo[1,2,*], Carmela Comito[1], Marco Minici[1,3], Ettore Ritacco[4], Gianluigi Folino[1], Francesco Sergio Pisani[1], Massimo Guarascio[1] and Giuseppe Manco[1]

[1]Institute for High Performance Computing and Networking, via P. Bucci 8-9/C, Rende, 87036, Italy
[2]University of Calabria, via P. Bucci, Rende, 87036, Italy
[3]Università degli studi di Pisa, Pisa, 56126, Italy
[4]Università degli studi di Udine, Via Palladio8, Udine, 33100, Italy

Abstract
Social media have become the ideal place for black hats and malicious individuals to target susceptible users through different attack vectors and then manipulate their opinions and interests. Fake news, radicalization, and pushing bias into the data represent some popular ways noxious users adopt to perpetrate their criminal intents. In this evolving scenario, Artificial Intelligence techniques represent a valuable tool to early detect and mitigate the risk due to the spreading of these emerging attacks. In this work, we describe the Machine Learning based solutions developed to address the problems mentioned above and our current research.

Keywords
Fake News Detection, Radicalization, Bias, Fairness

1. Introduction

Nowadays, the users' opinions and interests can be manipulated through a wide range of attack vectors: spreading fake news, radicalization due to specific suggestions in recommender systems, and pushing biased information in the training data for Machine Learning (ML) models represent just some recent examples of this phenomenon. Monitoring and early detection of these malicious behaviors are becoming crucial problems for political organizations and institutions as they can manipulate public opinion and change the results of events and campaigns.

In this complex and evolving scenario, Artificial Intelligence (AI) and Machine Learning techniques can play a key role in detecting and mitigating the risk due to these new emerging attacks. In particular, there is a growing interest in Deep Learning (DL) based solutions as their capability to effectively process different types of raw data (e.g., text, images, preferences, etc.) without the necessity of a feature engineering phase and the intervention of an expert.

In this work, we provide an overview of the above-mentioned problems and discuss current and future research lines. Specifically, in Section 2 we introduce the fake news detection problem and show two DL based solutions; Section 3 describes the radicalization phenomenon and presents a simulation framework to investigate its effects; Section 4 depicts a solution to mitigate the risk due to presence of bias in data; finally, Section 5 concludes the work and propose some interesting new research lines.

2. Neural Models for Fake News Detection

In recent times, social media channels, such as *Twitter*, *Facebook*, and *Instagram*, have been exploited to widespread false information and influence people's opinions. This phenomenon took different forms over time: clickbait, misinformation, and deceptive news are some examples, just to cite a few [1].

The exacerbation of this problem has attracted the attention of researchers and practitioners, especially because of the suspicion that several important recent events (e.g., the 2016 US election [2], the Brexit referendum [3], and the Vax campaign for the COVID-19 pandemic emergency [4]) were influenced by the diffusion of misleading information.

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy
*Corresponding author.
erica.coppolillo@unical.it (E. Coppolillo); carmela.comito@icar.cnr.it (C. Comito); marco.minici@phd.unipi.it (M. Minici); ettore.ritacco@uniud.it (E. Ritacco); gianluigi.folino@icar.cnr.it (G. Folino); francescosergio.pisani@icar.cnr.it (F. S. Pisani); massimo.guarascio@icar.cnr.it (M. Guarascio); giuseppe.manco@icar.cnr.it (G. Manco)
0000-0002-4670-8157 (E. Coppolillo); 0000-0001-9116-4323 (C. Comito); 0000-0002-9641-8916 (M. Minici); 0000-0003-3978-9291 (E. Ritacco); 0000-0002-8139-3445 (G. Folino); 0000-0003-2922-0835 (F. S. Pisani); 0000-0001-7711-9833 (M. Guarascio); 0000-0001-9672-3833 (G. Manco)

In this scenario, assessing the veracity and authenticity of news represents a crucial problem that can benefit from recent advances in AI and ML. In particular, the automatic detection of fake news is a relevant problem attracting great interest from the research community. This problem was traditionally addressed in the literature as a text classification problem [5] i.e., distinguishing between real and fake news documents.

However, learning reliable detection models able to identify misinformation requires coping with different complex issues. First, an effective solution should allow for handling low-level raw data frequently affected by noise, since the channels used to spread fake news typically allow for sharing only short text (e.g., Twitter). Moreover, the number of labeled training instances is limited; the labeling phase is a difficult and time-consuming task manually performed by domain experts. Finally, fake news can concern different topics; therefore, the features leveraged to perform the prediction should be domain independent to handle different topics. In addition, malicious contents represent only a limited portion of the data; then, the training set will exhibit an unbalanced distribution that makes the learning phase of the model more difficult.

2.1. Leveraging semi-supervised techniques for learning reliable models from scarce data

As mentioned above, in real scenarios data scarcity and the lack of labeled data can greatly affect the performances of Fake News Detectors. Recent approaches propose the usage of semi-supervised techniques to leverage the huge amounts of unlabelled data to boost the detection capability of the learned models. In [6], we proposed to employ a pre-trained instance of BERT model [7] as a backbone for classifying (as either fake or not) short news documents coming from a specific domain. However, instead of simply trying to fine-tune the BERT instance for this classification task, it is integrated into a self-training scheme. In more detail, a *Pseudo-Labelling* approach is used to map a number of unlabelled data instances, sampled from a given instance bucket. The pseudo-labels are assigned by a classification model, which is iteratively trained against a growing collection of both originally-labeled examples and pseudo-labeled ones. This process is repeated (in a self-training cycle) until a suitable stop criterion is satisfied (e.g., *the maximum number of iterations*, *loss convergence*, etc.).

This approach is depicted in Figure 1, where the initial set of labeled data is enriched with unlabeled data while using the classifier trained on the labeled data to assign "artificial" labels to unlabeled ones. Then, a new version of the classifier is built by using both the originally-labeled data and the newly automatically-labeled ones. Extensive experiments conducted on two different real datasets confirmed the effectiveness of the proposed approach in discovering accurate enough classification models even when the fraction of labeled data is relatively small.

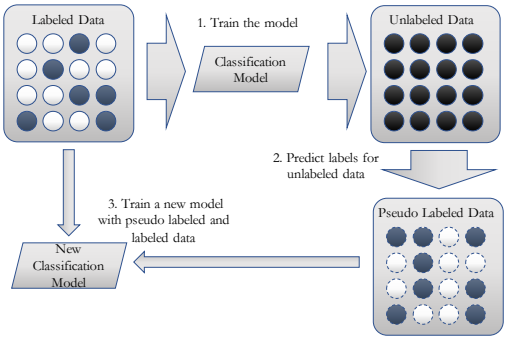

Figure 1: High-level view of the pseudo-labelling approach used to train the fake news classifier

2.2. Deep Learning models for Cross-Domain Fake News detection

As aforesaid, another challenge in automatic fake news detection is that they can vary across different domains. The problem is particularly relevant as they may emerge in contexts for which no prior evidence is available [8, 9, 10, 11, 12, 13, 14]. Despite several proposals, state-of-the-art exhibit several limitations. Most of the existing approaches are indeed designed for specific domains and, as a result, poorly perform in cross-domain scenarios.

To overcome this drawback, the proposed work introduces an end-to-end DL based framework for fake news detection tailored for cross-domain applications. The adoption of the DL paradigm [15] represents a natural solution to address the above issues, as it permits the learning of accurate classification models also from raw data (in our solution, the words composing the news) without requiring heavy intervention by data-science experts. Basically, these DL models are structured according to a hierarchical architecture (consisting of several layers of base computational units i.e., the artificial neurons are stacked one upon the other), allowing for learning features at different abstraction levels to represent raw data.

Relying on the success of deep neural networks, we combine recent advancements in the field, like autoencoders and adversarial generative approaches, for learning latent domain feature representations to use on new domains. Moreover, we derive features that are domain-invariant and, thus, benefit the detection of fake news on newly arrived, emergent events for which only a few verified news are available.

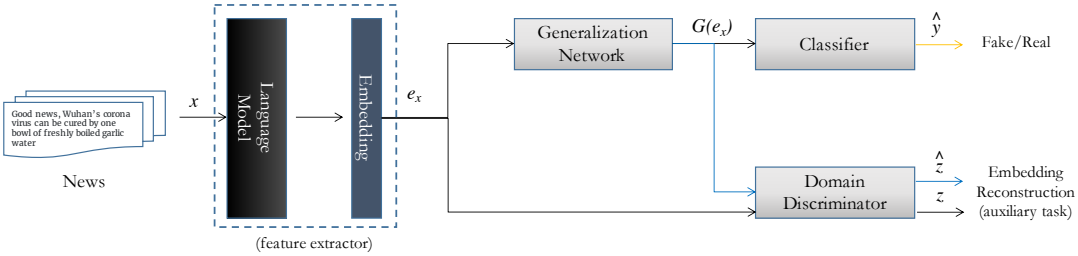

Figure 2: Overview of the solution for Cross-Domain Fake News Detections

The proposed solution is composed of three neural components (see Figure 2) that collaborate to solve two tasks simultaneously: the main one is to recognize fake information, and the second (auxiliary) task aims to produce domain invariant features. Preliminary experimentation conducted on two real datasets shows promising results and encourages further studies.

3. Radicalization in Recommender Systems

Today, digital platforms such as social media and e-commerce websites integrate advanced Recommender Systems (RecSys) to improve the user experience and make the items search more effective. With the growing amount of available content, RecSys have become a valuable tool to support users in navigating the large volumes of information proposed by these platforms.

Alas, the social consequences deriving from their usage are still object of study. Some recent works analyze their impact in terms of harmful phenomena such as echo chambers [16, 17], rabbit holes [18, 19], filter bubbles [20], and radicalization [21, 22, 23], showing that the long-term interaction with the RecSys can drift users towards polarized opinions and noxious contents. As a relevant example, it has been observed that continuous interactions with the recommendation algorithm can lead to severe extremization of users' political leaning [24].

In the following, we describe our current research line consisting of a simulation framework to evaluate the effects of RecSys on the users' preferences.

3.1. Algorithmic Drift

In our research, we consider a typical recommendation scenario in which some top-ranked items [25] are proposed and accepted by the user [26]. Items (that can take the form of media, news, videos, etc.) are tagged as *harmful* or *neutral*, while users are categorized as *non-*, *semi-* or *radicalized*, based on the percentage of harmful interactions in their preference history. Here, we investigate whether and how RecSys can affect and modify the users' preferences, as they could be used to manipulate their opinions and behaviors. In more detail, we devise a simulation framework for modeling the evolution of the interactions between users and items. In our setting, we assume that such interactions are fully driven by a (black-box) collaborative filtering algorithm. The simulation model is intended to start from an initial group of heterogeneous user preferences, from which the recommender system induces initial transition probabilities between item categories. The interaction between the simulation process and the recommendation algorithm allows for simulating the evolution of such preferences.

By analyzing the resulting transition probabilities, we are interested in assessing the impact of the algorithm on the initial population of user preferences.

In a nutshell, given the initial users interactions, our goal is to generate a probabilistic graph G_u, whose nodes include the subset of items for which the user u had at least one interaction during the simulation.

In this way, estimating the users' leaning drift built on this graph corresponds to estimating the recommender's influence in the long term. Figure 3 depicts an overview of the simulation framework we intend to implement.

Our final aim is to evaluate how much the recommendation algorithm changes user preferences after a certain number of interactions. We refer to this tendency as *algorithmic drift*.

In practice, given a recommendation algorithm pretrained with the initial user interactions (classified as neutral or harmful), the RecSys will tend to suggest more extreme content to initially unbiased users, and, vice-versa, neutral and harmless content to already radicalized ones.

To quantify the algorithmic drift induced by the recommender system, we intend to define novel graph-based metrics to be applied over the probabilistic graph G_u of a user u. Basically, we are interested in measuring the probability to move from harmful to neutral items (resp. from neutral to harmful) and to remain in cliques of neutral (resp. harmful) ones. As a consequence, we

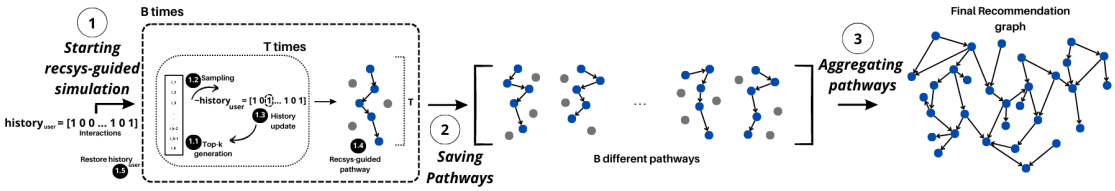

Figure 3: Overview of the simulation framework modeled to analyze the Algorithmic Drift phenomenon induced by Recsys

want to estimate the change in users' leaning before and after interacting with the recommendations.

The underlying idea is that the larger the value, the more the recommender is shifting the user's choices towards harmful items with the respect to its initial preferences.

3.2. Mitigating Radicalization

Radicalizing factors of RecSys can be mitigated on the basis of different strategies. Here, we define two different approaches: a *pre-processing* mitigation strategy and a *post-processing* one.

Pre-processing based solution. This strategy aims to modify the initial interactions dataset prior to training the recommender. As it can be viewed as a bipartite graph consisting of connections between users and items, we hypothesize that hindering the flow between non-radicalized users and items on the bipartite graph will also hinder the noxious effects underlying algorithmic drift of non-radicalized users in the long term.

This approach is implemented by adopting a rewiring strategy that, for each non-radicalized user u, randomly deletes a percentage $p\%$ of its edges toward items shared with semi-radicalized users. It then rewires them randomly choosing from the subset of neutral items which are exclusively connected to other non-radicalized users, i.e., an edge does not exist between these neutral items and semi-radicalized or radicalized users.

Once the rewiring is done for all non-radicalized users, we obtain a new altered dataset, which we use for training the recommender.

Post-processing based strategy. Based on the solution proposed in [27] for popularity debiasing, we propose to penalize the recommendation scores assigned to items. The rationale lies in boosting the final recommendation scores associated with neutral items and reducing the ones associated with harmful items. In practice, for each user, the penalization assigned to harmful items in the final recommendation score is inversely related to the radicalization level of the user (i.e., the percentage of harmful items in their initial history). To more accurately regulate the re-ranking process, we introduced a hyper-parameter to tune the importance of the re-ranking. By applying this mitigation strategy, we expect to reduce the noxious exposure obtained in the long term, especially towards initially unbiased users. Moreover, unlike the pre-processing strategy, this one exhibits the relevant benefit that does not require additional retraining phases of the model.

4. Bias and Fairness

ML-based systems typically aim for higher accuracy, but a recent research trend focuses on balancing the quality performances with ethical or discriminatory effects of ML-based recommendations. Indeed, unfairness in Decision Support Systems and RecSys allows for performing sophisticated attacks able to manipulate users' opinions or segregate subpopulations based on irrelevant characteristics [28]. Typical examples are limited access to credit or healthcare due, e.g., to race or gender. For example, a recent research[1] highlighted that an algorithm adopted to predict patients likely to need extra medical care heavily favored white patients over black patients.

Unfairness may come into play mainly for two reasons: inductive biases of algorithms (i.e., inheriting intrinsic biases of data) or through adversarial attacks (i.e., when a malicious actor purposefully poisons an ML system by exploiting some model weakness).

In [29], we addressed this problem in the case of RecSys. Specifically, we modeled a DL architecture that significantly improves the exposure of low-popular items. The proposed technique is based on two main aspects: resampling negative items and ensembling multiple instances of the algorithm.

5. Conclusions and future works

In this work, we presented the current research activities ICAR-CNR is developing in the framework of the SERICS project to mitigate the risk due to attacks aiming at manipulating the users' behaviors. Specifically, we have

[1] https://www.scientificamerican.com/article/racial-bias-found-in-a-major-health-care-risk-algorithm/.

shown as AI-based tools could represent a promising tool to detect, evaluate and reduce the influence of these new emerging attacks.

As future works, we are interested in extending our preliminary studies concerning radicalization phenomena, relying on more sophisticated and realistic simulation frameworks. In addition, as regards the fake news detection task, we aim to exploit multi-modal data like images, video, news propagation patterns, social context representing the user engagements of news on social media (e.g. the number of followers, hashtags, friendship networks, retweets) to improve the detection capability of the neural models.

Acknowledgments

This work was partially supported by project SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU.

References

[1] X. Zhou, R. Zafarani, A survey of fake news: Fundamental theories, detection methods, and opportunities, ACM Comput. Surv. 53 (2020).

[2] H. Allcott, M. Gentzkow, Social media and fake news in the 2016 election, Journal of Economic Perspectives 31 (2017) 211–36.

[3] H. MARSHALL, A. DRIESCHOVA, Post-truth politics in the uk's brexit referendum, New Perspectives 26 (2018) 89–106.

[4] C. Ngai, R. Singh, L. Yao, Impact of covid-19 vaccine misinformation on virality on social media: Content analysis of message themes and writing strategies (preprint), Journal of Medical Internet Research 24 (2022).

[5] C. Liu, X. Wu, M. Yu, G. Li, J. Jiang, W. Huang, X. Lu, A two-stage model based on bert for short fake news detection, in: C. Douligeris, D. Karagiannis, D. Apostolou (Eds.), Knowledge Science, Engineering and Management, Springer International Publishing, Cham, 2019, pp. 172–183.

[6] P. Zicari., M. Guarascio., L. Pontieri., G. Folino., Learning deep fake-news detectors from scarcely-labelled news corpora, in: Proceedings of the 25th International Conference on Enterprise Information Systems - Volume 1: ICEIS,, INSTICC, SciTePress, 2023, pp. 344–353. doi:10.5220/0011827500003467.

[7] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: NAACL-HLT, Association for Computational Linguistics, 2019, pp. 4171–4186.

[8] K. Shu, L. Cui, S. Wang, D. Lee, H. Liu, Defend: Explainable fake news detection, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '19, 2019, p. 395–405.

[9] C. Raj, P. Meel, Arcnn framework for multimodal infodemic detection, Neural Networks 146 (2022) 36–68.

[10] T. Sachan, N. Pinnaparaju, M. Gupta, V. Varma, Scate: Shared cross attention transformer encoders for multimodal fake news detection, in: Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM '21, 2021, p. 399–406.

[11] R. Kumari, A. Ekbal, Amfb: Attention based multimodal factorized bilinear pooling for multimodal fake news detection, Expert Systems with Applications 184 (2021) 115412.

[12] Z. Jin, J. Cao, H. Guo, Y. Zhang, J. Luo, Multimodal fusion with recurrent neural networks for rumor detection on microblogs, in: Proceedings of the 25th ACM International Conference on Multimedia, Association for Computing Machinery, New York, NY, USA, 2017, p. MM '17.

[13] Q. Jing, D. Yao, X. Fan, B. Wang, H. Tan, X. Bu, J. Bi, Transfake: Multi-task transformer for multimodal enhanced fake news detection, in: IJCNN, 2021, pp. 1–8.

[14] J. Wang, H. Mao, H. Li, Fmfn: Fine-grained multimodal fusion networks for fake news detection, Applied Sciences 12 (2022).

[15] Y. Le Cun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.

[16] E. Pariser, The filter bubble: How the new personalized web is changing what we read and how we think, Penguin, 2011.

[17] H. Ferraz de Arruda, F. Maciel Cardoso, G. Ferraz de Arruda, A. R. Hernández, L. da Fontoura Costa, Y. Moreno, Modelling how social network algorithms can influence opinion polarization, Information Sciences 588 (2022) 265–278.

[18] M. H. Ribeiro, R. Ottoni, R. West, V. A. Almeida, W. Meira Jr, Auditing radicalization pathways on youtube, in: Proceedings of the 2020 conference on fairness, accountability, and transparency, 2020, pp. 131–141.

[19] M. Haroon, A. Chhabra, X. Liu, P. Mohapatra, Z. Shafiq, M. Wojcieszak, Youtube, the great radicalizer? auditing and mitigating ideological biases in youtube recommendations, arXiv preprint arXiv:2203.10666 (2022).

[20] H. Zhang, Z. Zhu, J. Caverlee, Evolution of filter bubbles and polarization in news recommendation, 2023. arXiv:2301.10926.

[21] H. Hosseinmardi, A. Ghasemian, A. Clauset, M. Mo-

bius, D. M. Rothschild, D. J. Watts, Examining the consumption of radical content on youtube, Proceedings of the National Academy of Sciences 118 (2021) e2101967118.

[22] M. H. Ribeiro, V. Veselovsky, R. West, The amplification paradox in recommender systems, 2023. arXiv:2302.11225.

[23] F. Fabbri, Y. Wang, F. Bonchi, C. Castillo, M. Mathioudakis, Rewiring what-to-watch-next recommendations to reduce radicalization pathways, Association for Computing Machinery, New York, NY, USA, 2022.

[24] M. Haroon, A. Chhabra, X. Liu, P. Mohapatra, Z. Shafiq, M. Wojcieszak, Youtube, the great radicalizer? auditing and mitigating ideological biases in youtube recommendations (2022).

[25] Z. Guan, E. Cutrell, An eye tracking study of the effect of target rank on web search, Association for Computing Machinery, New York, NY, USA, 2007.

[26] A. Agarwal, X. Wang, C. Li, M. Bendersky, M. Najork, Addressing trust bias for unbiased learning-to-rank, Association for Computing Machinery, New York, NY, USA, 2019.

[27] H. Abdollahpouri, Popularity bias in ranking and recommendation, AIES '19: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (2019) 529–530.

[28] L. Oneto, N. Navarin, B. Biggio, F. Errica, A. Micheli, F. Scarselli, M. Bianchini, L. Demetrio, P. Bongini, A. Tacchella, A. Sperduti, Towards learning trustworthily, automatically, and with guarantees on graphs: An overview, Neurocomputing 493 (2022) 217–243.

[29] L. Caroprese, G. Manco, M. Minici, F. S. Pisani, E. Ritacco, Unbiasing collaborative filtering for popularity-aware recommendation (discussion paper), in: Proceedings of the 29th Italian Symposium on Advanced Database Systems, SEBD 2021, Pizzo Calabro (VV), Italy, September 5-9, 2021, volume 2994 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 450–457.