Understanding bias in Twitter-based intelligence analysis

Alexandros Karakikes*
Intelligent Systems Lab,
Dept. of Cultural Technology and Communication
University of the Aegean
Mytilene, Greece
cti22010@aegean.gr

Panagiotis Alexiadis
Intelligent Systems Lab,
Dept. of Cultural Technology and Communication
University of the Aegean
Mytilene, Greece
ctm21003@ct.aegean.gr

Theocharis Theocharopoulos
Intelligent Systems Lab,
Dept. of Cultural Technology and Communication
University of the Aegean
Mytilene, Greece
theocharopoulos@aegean.gr

Nikolaos Skoulidas
Department Faculty of Humanities
University of the Aegean
Rhodes, Greece
psed20004@rhodes.aegean.gr

Konstantinos Kotis*
Intelligent Systems Lab,
Dept. of Cultural Technology and Communication
University of the Aegean
Mytilene, Greece
kotis@aegean.gr

2023 14th International Conference on Information, Intelligence, Systems & Applications (IISA) | 979-8-3503-1806-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/IISA59645.2023.10345941

***Abstract*— Twitter has been lately engaged by the community of intelligence services worldwide that for monitoring areas of interest and identifying emerging social, political and security trends/threats. Over time, the Intelligence Community (IC) has identified bias as the major obstacle in information analysis, thus it has developed scientific and empirical methods for bias mitigation, in parallel to those developed by the information and communication technology (ICT) and artificial intelligence (AI) community. Both communities share the interest to accurately trace bias and ideally eradicate or moderate its effects. In this paper we introduce systemic parallels between Intelligence Analysis (IA) and Twitter Analytics (TA), to highlight similarities/dissimilarities, and investigate the feasibility of adapting and adjusting methodologies from the first field to the latter. Furthermore, we briefly introduce a novel framework for AI-augmented bias mitigation in the IC, proposing methods and tools, already adapted by the ICT community, for efficiently supporting bias mitigation methodologies adapted by the IC.**

Keywords—Twitter analytics, bias, intelligence analysis

I. INTRODUCTION

Since its inception, Twitter went through several developmental stages, to evolve to its current form, as a microblogging platform used by more than 300 million users on a global scale, posting circa 60.000 tweets per day. Its speed, accessibility to the average user and high volume of omnidirectionally flowing information, elated it from a social media network to a vast pool of information, to be collected, analyzed (predominantly via AI tools - scraping) and exploited for a wide array of purposes. Aforementioned purposes, fluctuate from marketing promotion [1] electoral [2], [3], financial [4] and other events prediction [5], to monitoring conflict zones, emerging and ongoing geopolitical tensions [6], [7].With regards to the latter field, the volume and the promptness of Twitter-derived data, combined with the collection and analysis capabilities that AI provides, has rendered it a highly efficient and exponentially attractive tool for intelligence services worldwide [8].

Bias distorts the validity of the information that are circulating through Twitter, subsequently hindering its utilization from the IC, as one of the available collection methods. Correspondingly, bias has been over the years, one of the main challenges that IC has been facing and mitigating through established practical methodologies, mainly deriving from the cognitive sciences, while employing its traditional disciplines. In order systemize to systemize its activities and mitigate bias, the IC instituted a comprehensive procedure of distinctly defined steps, called the intelligence cycle (called Cycle in the rest of the paper). The Cycle has been adopted, integrally or in slight differentiations, by the majority of the IC, and interestingly by private sector intelligence/risk assessment corporate entities. The core phases, that are commonly present in different versions of the Cycle, are: 'Planning & Direction", "Collection", "Processing and Exploitation", "Analysis & Production" and "Dissemination" [9], [10], [11]. Thus, for the purpose of the current study, these phases will be further examined for the potential presence of bias phenomena, and for their mitigation through proposed AI tools.

Presently, significant academic research has been performed towards identifying and mitigating bias in the Twitter ecosystem, predominantly through the usage of AI-powered tools, thus exploiting the processing possibilities they provide, towards examining vast datasets. These approaches address the bias issue either directly, or indirectly. Nonetheless, from an IC perspective, the majority of the aforementioned approaches seem to be fragmentary in nature, not addressing bias in all procedural steps that the IC is following, as part of its structured methodology. In addition, scarce research is devoted to exploring specifically Twitter as a potential intelligence collection and analysis platform.

The aim of the paper is a) to introduce a research study towards understanding bias in Twitter-based intelligence analysis, and b) to propose an AI-augmented framework for mitigating bias in IA, based on the capabilities offered by the ICT community. Eventually, the goal is to evaluate the suitability of the proposed framework based on the needs of the IC. The research is performed under the principal postulation that all intelligence gathering activities are to be executed through Twitter, while the proposed AI toolset for supporting the framework is to be comprised of tools suitable to each discernable step of the IA cycle. To accomplish that, a number of academic research papers and studies have been systematically reviewed, to identify trends in bias mitigation methodologies. This research also includes governmental reports and other unclassified, releasable to the public, content pertaining to the IC.

The structure of this paper is as follows: in Section II related work is summarized. In Section III, the proposed framework for bias mitigation in IA is presented, and section IV concludes the paper.

II. Related Work

In [12], one of the first attempts to introduce the concept of ontologies in various phases of the Cycle is presented, advocating that this would enhance communication among them, especially between the phases of collection and analysis. In [13], the establishment of an ontology-based Integrating Semantic Framework (ISF) is proposed, utilizing the Basic Formal Ontology (BFO) as a basis. By using BFO, authors contend that intelligence analysts can establish a common understanding and language for discussing intelligence-related concepts, which can improve communication and collaboration among analysts from different agencies and organizations. In [14], authors propose the semantic enhancement of data, using a formal ontology to improve the accuracy and consistency of intelligence analysis across various domains. the paper demonstrates the potential of ontology-based approaches to enhance the quality and effectiveness of intelligence analysis.

In [15], authors propose a generic framework to automatically and in real-time execute credibility analysis of posted messages in social media platforms, including Twitter, which is used as reference. The proposed model, based on natural language processing (NLP) and sentiment analysis techniques, will measure the level of acceptance of the topic or event referenced in the text. In [16], authors propose a mainly NLP-based approach, to examine the credibility of Twitter user accounts. At the credibility analysis component of the approach, source (Twitter account) credibility evaluation is proposed, by examining 12 features of each account, through the use of K-means clustering method. Authors state that their approach achieves an average accuracy of 68% in assessing a Twitter account's credibility. In [17], authors evaluate a dataset of 1.206 tweets comparatively with two different methods, a machine learning (ML) algorithm in R language and a two-human agents team (critics) who applied a set of manually assigned rules, with the goal to assign to each tweet one of four credibility levels (HC – highly credible, HNC – highly non credible, N – neutral, C – controversial). By merging different methods, ML-based and human agents-based, the achieved results were enhanced, raising the precision in the HC and HNC credibility classes by 8%–10% and decreasing low controversial (C) class error by 0,3%. In [18], authors propose an approach called 'Reliability Index for Twitter', which ascribes every Twitter profile with a numeric value, towards determining a profile's authenticity. This study has been exclusively directed in investigating Twitter profile's authenticity, and can be valuable in this particular field, nonetheless supplementary proposals in examining the veracity of the content are to be investigated in another relevant research. In [19], authors propose a methodology to automate military intelligence confidence assessments for Twitter messages. They determine the independence of sources by observing for explicit and implicit retweets, shared URLs, and distance between users in the Twitter graph.

Garimella et al. [20], suggest a content-based recommendation approach as alternative to Twitter's built-in recommendation, to mitigate confirmation bias and to assist towards increasing average Twitter user's exposure to conflicting views and beliefs, independently of Twitter user's viewpoint on a specific issue. Authors in [21] propose a method for analyzing bias in tweets using NLP techniques, which consists of three steps: data collection, feature extraction, and bias analysis In the bias analysis step, authors use the extracted features to analyze the bias in the tweets and additionally use a supervised ML algorithm to classify the tweets as biased or non-biased. The authors also use a sentiment analysis algorithm to identify the sentiment of the tweets. The limitations according to the authors include inability to detect contextual bias, non-textual bias (e.g. video, image), and elusive bias (sarcasm). In [22], authors present a framework for mitigating biases in ML systems, based on conditional Generative Adversarial Networks (cGANs) which allows the generation of new high-quality synthetic data related to the targeted population groups (minorities). The suggested framework enables ML systems engineers to estimate the actual distribution of the original data pertaining to the targeted population groups (population groups that are victims of biases) through establishing a termed "two-player game". Experimental results indicated that the proposed framework efficiently mitigated the biases against targeted population groups, while at the same time enhancing the prediction accuracy of the ML classifiers.

III. The Proposed Framework

Instances of bias can be manifested at each individual phase of the Cycle, either overtly or subconsciously. Thus, it is considered essential to dissect the Cycle, identify the respective manifestations of bias that might be encountered at each individual phase, and investigate the available AI-powered mitigation tools.

A. Bias in Planning & Direction

Bias can be introduced by prejudiced framing of research questions, due to decision makers' own biases, which in latter phase may influence the selection of sources or the analysis performed by analysts (confirmation bias).

NLP tools are proposed, for an initial review of an area-of-interest's social media, to recognize prevalent trends, issues, and events pertinent to the intelligence requirement. Subsequently, decision makers are briefed on the findings, to direct their requirements in a more objective way, to all areas of interest (e.g., if an antagonistic country is a traditional military power, decision makers might direct their requirements principally on this sector, thus neglecting the social/political/economic sectors, where potential underlying weaknesses might be missed).

Additionally, it is essential for the organization to engineer an ontology, which will form the linguistic and semantic basis of all internally exchanged information and knowledge, not only to avoid misinterpretations/ambiguities, but to be able to present information concisely, in a sentimentally neutral language, which will not instigate further bias.

B. Bias in the Collection

In this phase, availability bias may manifest. For instance, algorithmically favored twitter profiles of key figures of an antagonistic country, are more easily available and accessible to domestic IC, leading to overrepresentation of the views they're propagating or the information they're broadcasting in latter stages of the Cycle. Therefore, at this phase it is considered rather advantageous to provide IC the capability of monitoring even algorithmically unfavored, but yet influential

profiles, thus collecting, ideally, as-close-as-possible to the entirety of available information, and opposing views, while at the same time overriding limitations set by Twitter API [24].

Finally, bots and virtual assistants may be employed to interact with Twitter users, to gather information/survey opinions or respond to user questions for disinformation purposes. These tools can be custom-built to provide specific responses based on keywords or other criteria.

C. Bias in Processing

Bias can be introduced at this stage through the selection of which data to include or exclude, or the interpretation of data.

NLP tools can be used to automatically categorize the assembled data to recognize pertinent information and pre-process data before the main processing phase. More specifically, text classification/mining and sentiment analysis tools are proposed. Text classification automatically categorizes text data into specific thematic classes, a valuable task in identifying new, relevant information, sources (influential twitter profiles or ascending twitter profiles), or tracking the activities of specific individuals or groups, especially when the intelligence objective is related to counterterrorism or counter- disinformation field. Text mining techniques such as keyword extraction and named entity recognition can help recognize key concepts and entities that pertain to a certain intelligence objective. It should be emphasized, that findings from this task, should be forwarded back to the collection phase, to be further exploited by the collectors (e.g., exploiting new sources, direct collect efforts to new trends etc.).

Additionally, source/information evaluation, open-source frameworks, such as [15] are proposed for this phase. Ranking the reliability of each source and the information content, by comparing it with already validated sources/information, only the input/output phases of the process are examined, thus bypassing bias in the intermediate process. In addition, all scenarios, even seemingly improbable ones, are retained within the procedural loop, to be evaluated by the analysts in a subsequent stage. This feature might prove to be exceptionally useful, particularly taking into account that IC's history is full of failures stemming from exclusion of "low probability information/scenarios", most tragic of which the 9/11 attacks [27].

D. Bias in Analysis

Bias can also be present in the analysis of data, such as through the selection of analytical methods or the interpretation of results. Analysts may subconsciously apply their personal assumptions or prejudices, to the analysis, converging on hypotheses that confirm their predeterminations. To mitigate bias in this phase it is deemed necessary to select which analytic technique is to be implemented in order to opt for the appropriate AI tool. Our research, will focus into the Analysis of Competing Hypotheses (ACH), as it holds a highly prominent position in the IC for its efficiency and is the most preferable structured analytic technique (SAT). ACH was developed for use at the United States Central Intelligence Agency (CIA). By examining the eight distinct steps of ACH as described in [28], the use of supervised ML is proposed for:

- Hypothesis generation: ML algorithms may be utilized to generate multiple, heterogeneous hypotheses established upon available data and preceding knowledge. These algorithms can be trained on historical information (e.g., rival countries' historical bilateral relations) to distinguish patterns and generate new hypotheses accordingly.
- Hypothesis testing: Supervised ML algorithms may additionally be utilized to test competing hypotheses against available evidence. ML algorithms may also be trained on historical information to identify which hypotheses are most consistent with the available evidence and generate corresponding probability estimates.

Another approach proposed is the use of algorithm adversarial training, for the implementation of SATs that fall within the contrarian techniques or imaginative thinking categories. Contrarian techniques involve challenging assumptions and exploring differentiated interpretations, whereas in a similar manner adversarial training may ensure ML models are not excessively dependent on a single, biased perception. Adversarial training entails producing adversarial examples that are designed to prompt the model to erroneous predictions - and using these examples to train the model to be more effective. This may assist in identifying and rectifying biases in the model.

In addition, adversarial training, may be examined as a potential supportive tool for imaginative thinking techniques, which aim at creating new insights, diversified perspectives and/or suggest alternate outcomes. By generating adversarial examples (alternative scenarios), models can be trained to manage an array of hypothetical situations, and to produce more precise understanding of their potential outcomes.

E. Bias in Dissemination

Bias may appear in the dissemination of intelligence products, such as through the selection of which information to emphasize or the wording of the report. This can affect how decision-makers perceive the intelligence and how they act on it. In addition, organizational bias among various sub-groups of the intelligence service (e.g., collectors vs analysts, planners vs analysts etc.) or IC members (law enforcement, military intelligence, civilian intelligence etc.) can influence the decision to whom the intelligence products will be disseminated. Support in this phase may be offered by AI-powered tools that can be used to automatically disseminate intelligence reports to relevant stakeholders, such as government agencies, military organizations, and law enforcement agencies. Natural language generation (NLG) tools can be used to generate reports, summaries, and warnings, especially when they pertain to topics of repetitive nature (e.g., Daily Situation Report on a given geographical area, with standard recipients), to keep stakeholders informed about forthcoming threats and trends in a timely manner. To this end, the use and customization of open-source NLG tools like [29], is proposed.

Across all phases, but especially in the Planning & Direction and the Dissemination, it is crucial to maintain an established fair ontology. As stated above, this will be the standardized vocabulary used in all information circulating within the agency, and among collaborating agencies (Intelligence Services of a nation or national Intelligence Services of an Alliance's members), to avoid

misinterpretation, information overlapping and time consumption. As emphasized in [30] though, it is of the outmost criticality to engineer a fair ontology, since human or data bias may be encoded in engineered ontologies. Consequently, an unfair ontology may perpetuate bias in IA, instead of mitigating it, thus creating a paradox.

As a conclusive remark it should be stated that, among all phases of the Cycle, a continuous evaluation and feedback has to be performed and monitored, while at the same time the highest possible degree of human-machine collaboration is to be pursued.

IV. CONCLUSION

In this paper we propose a framework for AI-augmented bias mitigation in the IC. Also, we propose a toolset, already adapted by the ICT community, for supporting this framework of bias mitigation methodologies adapted by the IC. Our future plans are oriented towards evaluating the proposed tool-supported framework, detect potential weaknesses, examine the need for accordingly customizing its AI tools and exploring other AI tools that may enhance the effectiveness of the Cycle towards mitigating bias. Emphasis will be given on assessing Twitter scraping tools' role in the Collection phase and source reliability/information credibility frameworks' role in the Processing phase.

REFERENCES

[1] "Twitter Marketing." https://marketing.twitter.com/en (accessed Feb. 02, 2023).

[2] E. Tunggawan and Y. E. Soelistio, "And the winner is …: Bayesian Twitter-based prediction on 2016 U.S. presidential election," in 2016 International Conference on Computer, Control, Informatics and its Applications (IC3INA), Oct. 2016, pp. 33–37. doi: 10.1109/IC3INA.2016.7863019.

[3] D. A. Kristiyanti, Normah, and A. H. Umam, "Prediction of Indonesia Presidential Election Results for the 2019-2024 Period Using Twitter Sentiment Analysis," in 2019 5th International Conference on New Media Studies (CONMEDIA), Oct. 2019, pp. 36–42. doi: 10.1109/CONMEDIA46929.2019.8981823.

[4] X. Guo and J. Li, "A Novel Twitter Sentiment Analysis Model with Baseline Correlation for Financial Market Prediction with Improved Efficiency," in 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), Oct. 2019, pp. 472–477. doi: 10.1109/SNAMS.2019.8931720.

[5] G. A. Ruz, P. A. Henríquez, and A. Mascareño, "Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers," Future Generation Computer Systems, vol. 106, pp. 92–104, May 2020, doi: 10.1016/j.future.2020.01.005.

[6] J. S. Pohl, M. V. Seiler, D. Assenmacher, and C. Grimme, "A Twitter Streaming Dataset collected before and after the Onset of the War between Russia and Ukraine in 2022," SSRN Electronic Journal, 2022, doi: 10.2139/ssrn.4066543.

[7] S. Sazzed, "The Dynamics of Ukraine-Russian Conflict through the Lens of Demographically Diverse Twitter Data," in 2022 IEEE International Conference on Big Data (Big Data), Dec. 2022, pp. 6018–6024. doi: 10.1109/BigData55660.2022.10020274.

[8] J. Rovner, "Intelligence in the Twitter Age," International Journal of Intelligence and CounterIntelligence, vol. 26, no. 2, pp. 260–271, Jun. 2013, doi: 10.1080/08850607.2013.757996.

[9] United States of America, Office of the Director of National Intelligence, "National Intelligence Strategy of the United States of America," 2019.

[10] "Intelligence Cycle Graphic," United States of America, Federal Bureau of Investigations. https://www.fbi.gov/image-repository/intelligence-cycle-graphic.jpg/view (accessed Feb. 01, 2023).

[11] "Joint Intelligence, Joint Publication 2-0," USA, Chairman of the Joint Chiefs of Staff. 2013.

[12] R. Desimone and D. Charles, "Towards an Ontology for Intelligence Analysis and Collection Management," 2002.

[13] B. Mandrick and B. Smith, "Philosophical foundations of intelligence collection and analysis: a defense of ontological realism," Intelligence and National Security, vol. 37, no. 6, pp. 809–819, Sep. 2022, doi: 10.1080/02684527.2022.2076330.

[14] "Ontology for the intelligence analyst," 2012. [Online]. Available: https://www.researchgate.net/publication/290966313

[15] Y. Cardinale, I. Dongo, G. Robayo, D. Cabeza, A. Aguilera, and S. Medina, "T-CREo: A Twitter Credibility Analysis Framework," IEEE Access, vol. 9, pp. 32498–32516, 2021, doi: 10.1109/ACCESS.2021.3060623.

[16] M. Wijesekara and G. U. Ganegoda, "Source credibility analysis on Twitter users," in Proceedings - International Research Conference on Smart Computing and Systems Engineering, SCSE 2020, Sep. 2020, pp. 96–102. doi: 10.1109/SCSE49731.2020.9313064.

[17] L. Krzysztof, S.-W. Jacek, J.-L. Michal, and G. Amit, "Automated Credibility Assessment on Twitter," Computer Science, vol. 16, no. 2, p. 157, 2015, doi: 10.7494/csci.2015.16.2.157.

[18] K. Sharma, "Reliability Index for Twitter – Twitter Handles' Credibility Assessment," HELIX, vol. 8, no. 5, pp. 4094–4099, Aug. 2018, doi: 10.29042/2018-4094-4099.

[19] M. M. Kokar and B. Ulicny, "Automating Military Intelligence Confidence Assessments for Twitter Messages," 2014. [Online]. Available: http://blogs.aljazeera.net/twitter-dashboard

[20] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis, "Mitigating Confirmation Bias on Twitter by Recommending Opposing Views," in WSDM 2017 - Proceedings of the 10th ACM International Conference on Web Search and Data Mining, Feb. 2017, pp. 81–90. doi: 10.1145/3018661.3018703.

[21] E. Tankard, C. Flowers, J. Li, and D. B. Rawat, "Toward Bias Analysis Using Tweets and Natural Language Processing," in 2021 IEEE 18th Annual Consumer Communications & Networking Conference (CCNC), Jan. 2021, pp. 1–3. doi: 10.1109/CCNC49032.2021.9369461.

[22] A. Abusitta, E. Aïmeur, and O. A. Wahab, "Generative Adversarial Networks for Mitigating Biases in Machine Learning Systems," May 2019.

[23] H. Mansourifar and S. J. Simske, "GAN-Based Object Removal in High-Resolution Satellite Images," Jan. 2023.

[24] "Twitter API Rate limits," https://developer.twitter.com/en/docs/twitter-api/rate-limits.

[25] "Twint." https://github.com/twintproject/twint (accessed Mar. 16, 2023).

[26] "Twitter-scraper." https://github.com/bisguzar/twitter-scraper (accessed Mar. 16, 2023).

[27] United States Senate Intelligence Committee, "H. Rept. 107-792 - Joint Inquiry into Intelligence Community Activities before and after the Terrorist Attacks of September 11, 2001," Washington DC, Dec. 2002.

[28] Heuer RJ, Psychology of Intelligence Analysis. Washington D.C: Central Intelligence Agency, Center for the Study of Intelligence, 1999.

[29] "Natural Language Summary Generation from Structured Data." https://github.com/akanimax/natural-language-summary-generation-from-structured-data (accessed Mar. 16, 2023).

[30] E. Paparidis and K. Kotis, "Towards Engineering Fair Ontologies: Unbiasing a Surveillance Ontology," in 2021 IEEE International Conference on Progress in Informatics and Computing (PIC), Dec. 2021, pp. 226–231. doi: 10.1109/PIC53636.2021.9687030.