Debunker Assistant: a support for detecting online misinformation

Arthur Thomas Edward Capozzi Lupi[1,2], Alessandra Teresa Cignarella[1,2], Simona Frenda[1,2], Mirko Lai[1,2], Marco Antonio Stranisci[1,2] and Alessandra Urbinati[1,3]

[1]aequa-tech, Turin, Italy
[2]Computer Science Department, University of Turin, Turin, Italy
[3]MOBS, Northeastern University, Boston, USA

Abstract
This paper describes the framework developed for the *Debunker-Assistant*, an application that allows users and newspapers to assess the trustworthiness of a news item starting from its headline, body of text and URL. The Debunker-Assistant adapts ideas from Natural Language Processing and Network Science to counter the spread of online misinformation. Its centerpiece is a set of four *News Misinformation Indicators* based on linguistically engineered features, models, network analysis metrics (Echo Effect, Alarm Bell, Sensationalism, and Reliability). In this short contribution, we describe the back-end structure on which the indicators are implemented.

Keywords
Misinformation, Debunker-Assistant, Linguistic Features, Web Domain Network

1. Introduction and Background

Fake news threatens democracies, public health, and news outlets' credibility. For this reason, tackling misinformation is an open challenge faced by governments, private companies, and the scientific communities [1].

There are many proposed approaches, some based on AI methods, others on fact-checking by human experts, still others on a combination of the two [2, 3]. However, the fact that fake news detection algorithms are often owned by private social media companies, and additionally the adoption of "black-boxed" algorithms contribute to the lack of transparency in the fake news identification and filtering process.

Debunker-Assistant (D-A) is a back-end AI tool which supports the analysis and detection of online misinformation. The D-A tool takes as an input the link of a news article and returns its misinformation profile based on Natural Language Processing (NLP) and Network Analysis (NA) features. Such an approach is inspired by the survey provided by [4].

D-A works with Italian and is not bounded on a specific topic: it is designed as a general purpose tool that can extract relevant features for assessing the quality of information[1].
Its main purposes are:

1. displaying the indicators to deal with misinformation;
2. de-biasing the mechanisms to make trustworthy the internet;
3. showing insights about a certain context to aid the search and discovery of information.

In this paper, we present the various features of NLP and NA embedded in the D-A tool and how these features have been designed and developed. As represented in Figure 1, D-A allows users to search for news and compare them against a given set of 4 macro-indicators of misinformation: Echo Effect, Alarm Bell, Sensationalism, and Reliability. These indicators are designed on the basis of specific linguistic and network features, such as the absence of sources, non-authority of references, the presence of specific figures of speech or flames, and other stylistic characteristics.

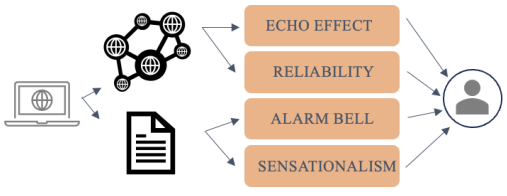

Figure 1: Overall pipeline of the D-A tool.

[1]The access to the D-A tool is offered through API, a public available version can be found at https://github.com/AequaTech/DebunkerAssistant

All the authors contributed equally.
CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

arthurtomasedward.capozzi@unito.it (A. T. E. C. Lupi); alessandrateresa.cignarella@unito.it (A. T. Cignarella); simona.frenda@unito.it (S. Frenda); mirko.lai@unito.it (M. Lai); marcoantonio.stranisci@unito.it (M. A. Stranisci); a.urbinati@northeastern.edu (A. Urbinati)

To operationalize our objectives, we focus on three main actions: **Inform** D-A can be freely used by citizens who want to check if a content they have read includes potential misinformation signals; **Support** D-A can also be adopted by journalists who work in the field of debunking, empowering their activity; **Teach** teachers are able to use D-A for designing information literacy activities in their classrooms.

2. News Misinformation Indicators

2.1. Echo Effect

Echo Effect encodes the coupled aspects of origin and dissemination. There are news sources that occupy a key role in the spread of misinformation [5]. This class aims to identify the most important domains in the news ecosystem, and measure the impact and range of these sources. We employ the network science framework, modeling each domain as a node of a network model, $G = (V, E, \varpi)$, where V is a set of nodes of N domains, E it the set of link $e_{ij} = (i, j)$ that encodes the pairwise interactions between two domains, and $\varpi : V \times V \to \mathbb{Y}$ is a function, defined for each pair of nodes $i, j \in V$, that maps each link e_{ij} into a weight w_{ij}, that stands for the total number of interactions from the domain i to j.

A well established method to assess the importance of a node in a network is to measure the value of "centrality" that a node has with respect to all the other nodes. There are many possible definitions of importance and so many centrality measures. In the context of news, we adopted the *hyperlink-induced topic search* [6] algorithm to infer the "origin" and "destination" of the echo effect, i.e., the set of nodes responsible for the amplification of signals traversing the network and the set of nodes in which the signal resonates the most. While the *betweenness centrality* [7] measures the extent to which a node can reach other locations in the network[2].

2.2. Alarm Bell

Alarm Bell takes into account the presence of possible pragmatic implications like flaming and ironic language in the news. Specifically, it analyzes whether the headline and news content contain these elements, and outputs an average score of the various probabilities.

To obtain these values, we designed specific classifiers to account for the presence in the text of irony, hate speech and stereotypes. For training our models, we used two well-known benchmark datasets: IronITA [8] and the extended version of the HaSpeeDe2 [9, 10]. These datasets are annotated for all the dimensions necessary to our purposes: IronITA contains Italian tweets about political and immigration issues, annotated for irony; the extended version of HaSpeeDe2 contains Italian tweets and news headlines about the integration of minority communities, annotated for hate speech, stereotypes, and irony.

To create the models, we fine-tuned the base version of BERT for Italian[3], optimizing basic hyperparameters like learning rate and batch size for each phenomenon. To make the models smalls and easy to call at each request of the API, we reduced the number of trainable parameters, adopting the LoRA technique implemented by HuggingFace [11]. If the aim of the training is to minimize the loss function[4] on the validation set, the evaluation of the models is focused on the analysis of f1-macro on the test sets of news headlines of HaSpeeDe2 (0.722 for hate speech and 0.749 for stereotype) and of tweets of IronITA (0.769 for irony)[5].

The tool returns values in a range from 0 to 1, representing the probabilities that these phenomena (irony, hate speech and stereotypes) are present in the news. The presence of them could be an alarm bell on the seriousness and trustworthiness of the news [12], as well as of their malicious intentions [4].

2.3. Sensationalism

Sensationalism includes three groups of features that are indicative of the diffusion of clickbait, especially while observing online headlines: informal style, syntactic complexity, emotion profile. Inspired by [13], who studied the linguistic and typological features commonly associated with clickbait in online news headlines, we designed the following features.

1. `informal_style`:
   use of upper case - The ratio between number of uppercase words and total number of words in a news headline;
   repeated letters - The ratio between words with a repeated letter for $n > 3$ (e.g., *SVEGLIAAAAA!* or *diiiiici*) and the total number of words in a news headline.
   distinctive punctuation patterns - In particular the count of 'weird' punctuation such as ! ... * + = $, and the count of 'normal' punctuation such as . , ; : ?
   presence of emojis - The normalized count of emojis in a news headline.

[2] See Appendix for details.

[3] https://huggingface.co/dbmdz/bert-base-italian-cased

[4] The function used in our experiments is the Binary Cross Entropy with different weight for each class, considering the constant imbalance of the positive class in all the datasets.

[5] The models are available in https://github.com/AequaTech/DebunkerAssistant

2. `syntactic_complexity`:
   avg length - A metric that computes the textual length of the news headline and compares it with the average textual length of a news headline stored in a proprietary database. The value is normalized between 0 and 1.
   shortest/longest - A metric that shows the comparison of the length of the headline and collocates it in one of the four quartiles of the database with all the other news headlines previously analyzed.

3. `affective_profile`:
   overall emotion - Inspired by the study Vosoughi et al. [14] about the presence of specific emotions in false stories, this feature returns the averaged score about the presence of emotions in the news. This value is based on the identification of the eight primary emotions—anger, fear, sadness, disgust, surprise, anticipation, trust, and joy— following the psycho-evolutionary taxonomy proposed by [15] through the multilingual NRC lexicon of [16]. Emotional signals have also been explored to detect credibility by Giachanou et al. [17].
   overall sentiment - As the previous feature, it is a normalized value of the presence of sentiment, exploiting the intensity score from Sentix lexicon [18]. The sentiment score to detect misinformation has already been used in previous works such as Baly et al. [19], and Ghanem et al. [20].

2.4. Reliability

Inspired by the different source cues that influence people when making credibility evaluation decisions [21], we also leverage the context of certain sites, because they could impact the type of content spread. In this line, this class aims to quantify its overall "reliability". Firstly, supported by debunking websites, we compiled two separate lists summarizing the established positions of specific URL domains regarding the spread of misinformation: the *white list* containing sites considered mostly safe, and the *black list* containing sites known for disseminating intentionally misleading information. Secondly, to frame the context, we assess two aspects linked to a URL domain.

1. `neighborhood`: We employ the network analysis framework to evaluate the neighborhood of a web page. We retain the same model described in Section 2.1 to detect special agglomerations of nodes, or communities (locally dense connected sub-graphs), such that the nodes that belong to the same community have a higher probability to be linked than the nodes in different communities. We employed the stochastic block model [22], a generative model that regards in a formal context the actual presence of a specific not random map into which the network can be partitioned. Once having the different communities, we characterized them by assigning to each a label, *white* or *black*, counting the majority of their nodes (web pages) belonging to either the white or the black list.

2. `solidity`: To quantify solidity, we take advantage of the "whois" metadata attached to each URL domain, which is the country of registration, creation date, and expiration date. We calculated the amount of time (days) that a specific domain maintains the same characteristics.

3. Conclusions

D-A is a tool developed in support of detecting online misinformation. Its social impact is twofold. First, it contributes to counter misinformation and improve the quality of information. In fact, the tool is aimed at supporting who deal every day with this issue. Second, it provides a better understanding of how these artificial intelligence technologies work in the context of the news media.

Given the complex and ever-changing nature of content creation and information dissemination, there are several directions for improvement. For example, users could be involved in providing anonymous feedback on the news itself and on the characterization of the evaluated articles, improving the overall performing skill of the tool, more so for those features that are less explored in the literature. In addition, this type of interaction makes the user think about important aspects of the online information, thus increasing awareness. Over time, as users search for new URLs, the core data that feed the models will expand to cover larger and more diverse sets of domains, incorporating a richer perspective on news consumption. To help the above research directions, we plan to develop a user-friendly interface and evaluate the general user experience. Finally, a future challenge would be to scale the model for other languages starting from English.

Ethics Considerations D-A provides the user with a set of characteristics about the article and a set of information about the domain hosting the article. Thus, the output generated does not consist of a binary classification of the truthfulness of an online newspaper article. Nonetheless, we are aware of the ethical issues surrounding the characterization and evaluation of online news.

First of all, the fallibility of NLP models must be taken into account, secondly some aspects concerning the world of information can have shades of subjectivity and be sensitive especially for some users.

Acknowledgments

This project has received funding from the European Union's Horizon Europe research and innovation program via the NGI Search grant agreement (ID: 101069364), linked with the Next Generation Internet Initiative.

References

[1] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, Science 359 (2018) 1146–1151. URL: https://www.science.org/doi/abs/10.1126/science.aap9559. doi:10.1126/science.aap9559.

[2] P. Nakov, D. P. A. Corney, M. Hasanain, F. Alam, T. Elsayed, A. Barr'on-Cedeno, P. Papotti, S. Shaar, G. D. S. Martino, Automated fact-checking for assisting human fact-checkers, in: International Joint Conference on Artificial Intelligence, 2021.

[3] S. I. Manzoor, J. Singla, Nikita, Fake news detection using machine learning approaches: A systematic review, in: 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019, pp. 230–234. doi:10.1109/ICOEI.2019.8862770.

[4] G. Ruffo, A. Semeraro, A. Giachanou, P. Rosso, Studying fake news spreading, polarisation dynamics, and manipulation by bots: A tale of networks and language, Computer science review 47 (2023) 100531.

[5] S. Vilella, A. Semeraro, D. Paolotti, G. Ruffo, Measuring user engagement with low credibility media sources in a controversial online debate, EPJ data science 11 (2022) 29.

[6] J. M. Kleinberg, Hubs, authorities, and communities, ACM computing surveys (CSUR) 31 (1999) 5.

[7] L. C. Freeman, D. Roeder, R. R. Mulholland, Centrality in social networks: Ii. experimental results, Social networks 2 (1979) 119–141.

[8] A. T. Cignarella, S. Frenda, V. Basile, C. Bosco, V. Patti, P. Rosso, Overview of the EVALITA 2018 task on irony detection in Italian tweets (IronITA), in: Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018) co-located with the Fifth CLiC-it, volume 2263, 2018, pp. 1–6.

[9] M. Sanguinetti, G. Comandini, E. Di Nuovo, S. Frenda, M. Stranisci, C. Bosco, T. Caselli, V. Patti, I. Russo, Haspeede 2@ evalita2020: Overview of the evalita 2020 hate speech detection task, in: Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online. CEUR. org, 2020.

[10] S. Frenda, V. Patti, P. Rosso, Killing me softly: Creative and cognitive aspects of implicitness in abusive language online, Natural Language Engineering (2022) 1–22. doi:10.1017/S1351324922000316.

[11] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations, 2022. URL: https://openreview.net/forum?id=nZeVKeeFYf9.

[12] V. L. Rubin, N. Conroy, Y. Chen, S. Cornwell, Fake news or truth? using satirical cues to detect potentially misleading news, in: Proceedings of the second workshop on computational approaches to deception detection, 2016, pp. 7–17.

[13] R. Kemm, The linguistic and typological features of clickbait in youtube video titles, Social Communication 8 (2022) 66–80.

[14] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, Science 359 (2018) 1146–1151. doi:10.1126/science.aap9559.

[15] R. Plutchik, A general psychoevolutionary theory of emotion, in: Theories of emotion, Elsevier, 1980, pp. 3–33.

[16] S. M. Mohammad, P. D. Turney, Crowdsourcing a word-emotion association lexicon, Computational Intelligence 29 (2013) 436–465. URL: https://doi.org/10.1111/j.1467-8640.2012.00460.x.

[17] A. Giachanou, P. Rosso, F. Crestani, Leveraging emotional signals for credibility detection, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19, Association for Computing Machinery, New York, NY, USA, 2019, p. 877–880. URL: https://doi.org/10.1145/3331184.3331285. doi:10.1145/3331184.3331285.

[18] V. Basile, M. Nissim, Sentiment analysis on Italian tweets, in: Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2013, pp. 100–107.

[19] R. Baly, G. Karadzhov, D. Alexandrov, J. Glass, P. Nakov, Predicting factuality of reporting and bias of news media sources, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 3528–3539. URL: https://aclanthology.org/D18-1389. doi:10.18653/v1/D18-1389.

[20] B. Ghanem, S. P. Ponzetto, P. Rosso, F. Rangel,

FakeFlow: Fake news detection by modeling the flow of affective information, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 679–689. URL: https://aclanthology.org/2021.eacl-main.56. doi:10.18653/v1/2021.eacl-main.56.

[21] M. J. Metzger, A. J. Flanagin, Psychological Approaches to Credibility Assessment Online, John Wiley & Sons, Ltd, 2015, pp. 445–466. doi:https://doi.org/10.1002/9781118426456.ch20.

[22] T. P. Peixoto, Bayesian stochastic blockmodeling, Advances in network clustering and blockmodeling (2019) 289–332.

Appendix - Centrality measures

Here are the details of Section 2.1. The hyperlink-induced topic search algorithm (also known as HITS or *hubs and authorities*) is defined for directed networks, and computes the authority centrality and the hub centrality, which quantify vertices' prominence in the two roles, as a "receiver" or as a "provider" of information. The hub vector $\vec{h}_t = (h_{1,t}, \ldots, h_{|V|,t})^\intercal$ and the authority vector $\vec{a}_t = (a_{1,t}, \ldots, a_{|V|,t})^\intercal$ in $t \in T$ of $G = (V, T, \varpi)$ are defined by the limit of the following set of iterations:

$$\vec{h}_t(x+1) = c_t(x) W_t \vec{a}_t(x+1) \tag{1}$$

and

$$\vec{a}_t(x+1) = d_t(x) W_t^\intercal \vec{h}_t(x), \tag{2}$$

where $c_t(x)$ and $d_t(x)$ are normalization factors to make the sums of all elements become unity, i.e., $\sum_{i=1}^{|V|} h_{i,t}(x+1) = 1$ and $\sum_{i=1}^{|V|} a_{i,t}(x+1) = 1$. The initial values of the scores are $h_{i,t}(0) = \frac{1}{|V|}$ and $a_{i,t}(0) = \frac{1}{|V|}$ for all $i \in V$.

The betweenness centrality measures the extent to which a node lies on paths between other vertices. We define the betweenness centrality of a node $i \in V$ at time $t \in T$ as

$$c_b(i,t) = \sum_{\substack{s,e \in V \\ i \neq s \neq e}} \frac{\sigma_{se,t}(i)}{\sigma_{se,t}}, \tag{3}$$

where $\sigma_{se,t}$ is the total number of shortest paths from node s to node e at time t, and $\sigma_{se,t}(i)$ is the number of such paths passing through node i.