



# FACADE: Fake Articles Classification and Decision Explanation

Erasmus Purificato<sup>1,2</sup>, Saijal Shahania<sup>1,3</sup>, Marcus Thiel<sup>1</sup>,  
and Ernesto William De Luca<sup>1,2</sup>

<sup>1</sup> Otto von Guericke University Magdeburg, Magdeburg, Germany  
{[erasmo.purificato](mailto:erasmo.purificato),[saijal.shahania](mailto:saijal.shahania),[marcus.thiel](mailto:marcus.thiel),[ernesto.deluca](mailto:ernesto.deluca)}@ovgu.de

<sup>2</sup> Leibniz Institute for Educational Media | Georg Eckert Institute,  
Brunswick, Germany

{[erasmo.purificato](mailto:erasmo.purificato),[deluca](mailto:deluca)}@gei.de

<sup>3</sup> German Centre for Higher Education Research and Science Studies,  
Hanover, Germany  
[shahania@dzhw.eu](mailto:shahania@dzhw.eu)

**Abstract.** The daily use of social networks and the resulting dissemination of disinformation over those media have greatly contributed to the rise of the *fake news* phenomenon as a global problem. Several manual and automatic approaches are currently in place to try to tackle and defuse this issue, which is becoming nearly uncontrollable. In this paper, we propose *Facade*, a fake news detection system that aims to provide a complete solution for classifying news articles and explain the motivation behind every prediction. The system is designed with a cascading architecture composed of two classification pipelines dealing with either low-level or high-level descriptors, with the overall goal of achieving a consistent confidence score on each outcome. In addition, the system is equipped with an explainable user interface through which fact-checkers and content managers can visualise in detail the features leading to a certain prediction and have the possibility for manual cross-checking.

**Keywords:** Fake news detection · Feature engineering · Explainability

## 1 Introduction

Over the last few years, the term *fake news* has become extremely popular to the point of making this phenomenon a worldwide issue [15, 17]. This concept gained traction following the emblematic 2016 US elections, in which the diffusion of misinformation on social networks has been used as a form of propaganda to get substantial political advantages [8]. The main characteristics of fake news, i.e. *volume*, *variety* and *velocity* [22], are sustained by the rapid spread of *web bots* [12] that make fabricated articles easy to publish and disinformation sources even more difficult to recognise and control. In this scenario, attention is being paid by fact-checkers [10] and content managers [18] in automatic detection systems for two main motivations: 1) manual detection by experts and organisations

is a time-consuming and expensive process, with a huge human-resources involvement to maintain it [11]; 2) the nature and composition of fake news are not the same for every fabricated article. Indeed, some news entries are blatant lies, while others hide their disinformation content among the facts. Furthermore, the outcomes have to be transparent to increase trust in such systems since the results must be cross-checked to be deemed false.

In this paper, we propose **Facade**<sup>1</sup>, an automatic system for *fake articles classification and decision explanation*. The system is designed with a cascading architecture composed of two classification pipelines. For each document to analyse, the detection process starts with a first classifier which exploits basic linguistic features (*low-level descriptors*) previously extracted from several fake news datasets. The second pipeline makes use of more complex features (*high-level descriptors*), such as sentiment, emotion, and attribution to known real or fake sources, computed by additional algorithms. We further present an explainable user interface (UI) which can help end users understand what parts of the investigated article are likely to be fake and for what reasons through the implementation of feature importance and post-hoc methods.

## 2 Existing Fake News Detection Systems

The early-stage detection systems started with manual fact-checking initiatives, and despite the enormous human effort required, some of them are nowadays still hugely reliable, such as *Truth-o-Meter* [1] and *Snopes* [2]. On the automatic detection front, many works, such as [19], shape their systems around the notion of linguistic similarity of the analysed content with known real or fake articles. Nevertheless, the state of the art is unsurprisingly dominated by machine learning and deep learning models, which usually rely on a supervised learning approach (e.g. [24, 26]). In a recent publication in this field, Zhang et al. [23] leveraged the relationship between the emotions portrayed in the news content and the end users' emotions expressed in the related comments. In most of the existing systems, however, the component of interpretability is almost overlooked. Due to the coexistence of fake and real news, it is necessary to incorporate the vision of experts and the audience in general [14, 25], and this can be achieved through an effective explainable UI. Only a few works, such as *dEFEND* [20] and *Xfake* [21], presented a solution having explainability as a fundamental part of the system.

## 3 The Facade System

The **Facade** system is designed with a cascading architecture composed of three main phases: 1) **Feature extraction**: low-level and high-level features are extracted from the adopted fake news datasets: *ISOT Fake News Dataset* [3],

<sup>1</sup> <https://github.com/dtdh/facade> (links to **demo video** and **live webapp** inserted inside the repository).

*Fake News Dataset* [4], *Fake News Corpus* [5], *Multi-Perspective Question Answering Dataset* (MPQA) [6] and *Myers-Briggs Personality Type Dataset* (MBTI) [7]. 2) **Classification**: leveraging the low-level features, a first classifier is executed to the documents to produce the probability of how likely the analysed news is fake or real. 3) **Filtering**: based on the resulting probability and the related confidence level of the classifier that receives **low-level descriptors** (i.e. basic linguistic features extracted from the article texts and headlines, such as size, number of grammatical errors, parts of speech and term frequencies), each news is filtered and marked as *fake*, *real* or *uncertain*. For the latter group, a second classification is applied, making use of **high-level descriptors** (i.e. complex features detected from the news content with additional algorithms, like sentiment, entailment, attribution, syntactical structure, tones and latent topics). Both pipelines have different classifiers catering to the inputted features. The classifiers were selected based on specific evaluation metrics such as accuracy, recall, precision, F1 scores and other customised metrics, whose detailed discussion is out of the scope of this paper.

The explainability methods included in the system, constituting the basis for the UI, are *feature importance*, *partial dependence plots* and *SHAP*. **Feature importance** [9] is a widely used method for finding the attributes that contribute the most towards the classifier’s predictions. **Partial dependence plots** (PDP) [13] is a model-agnostic and global method, aiming to create a link between the target label (in our case, fake or real) and the attributes utilised by the classifiers (i.e. low-level and high-level descriptors). **SHAP** (SHapley Additive exPlanations) [16] is a state-of-the-art explainability technique and it is mainly used to figure out the effect of each attribute of a classifier’s prediction.

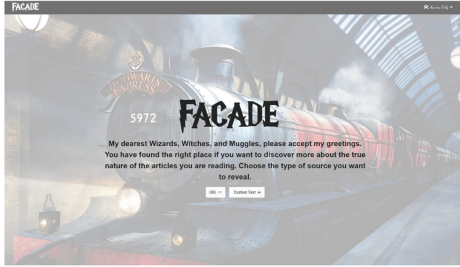
## 4 Demonstration

In this section, we will guide our readers in the exploration of the functionalities of Facade, whose UI has been designed with a *Harry Potter* style, resembling a wizard revealing the truth or the falsehoods of an investigated article.

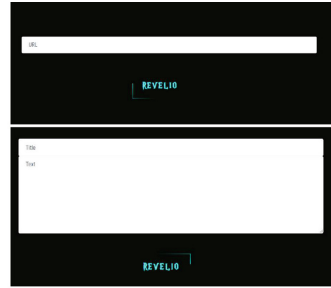
The initial page (Fig. 1) shows a welcome message (Fig. 1a) with two possible input options (Fig. 1b): insert the URL of a public article and manually type a custom text to analyse, useful to evaluate only a piece of news.

After the execution of the two pipelines, we land on the result page (Fig. 2a), where we can view the prediction (top left corner) and the related confidence score to the right. Optionally, we can highlight the sentences attributed by the system to real or fake sources, coloured in green and red, respectively. The colour gradient relates to the similarity score between a sentence and the attributed source. As displayed in Fig. 2b, we can also check the detailed explanations for attributions and features by hovering over the specific contributions. The list of the most influential features is on the right-hand side of the result page. The arrows next to each feature name indicate how strong the contribution of that feature towards the prediction is through their number.

Additionally, by browsing the *Explainer Dashboard*, we can visualise all the SHAP values and the partial dependence plots for a single prediction. With the



(a) Welcome page



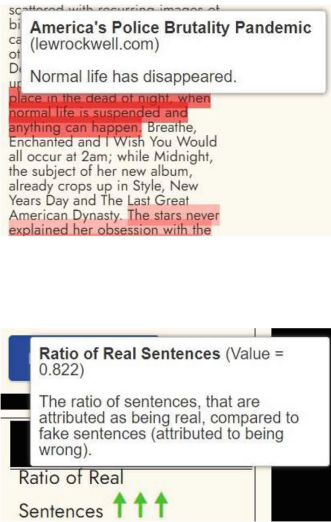
(b) Input options

Fig. 1. Initial page of Facade

“What If” module, we can adjust the feature importance scores to see how the prediction changes accordingly in a counterfactual scenario.



(a) Sentences highlights



(b) Feature explanations

Fig. 2. Result page of Facade with explanation details (Color figure online)

To summarise, the system is designed to deal with the needs of computer scientists and non-expert audiences. The more specific aspects, such as entering the URL or directly the news text to be questioned, and highlighting parts of the articles considered fake or real by the system in the second pipeline, mainly cater to the non-expert audiences. The highlighting is done in a realistic colour scheme so that it is easier for everyone to follow, irrespective of their

background or technical knowledge. The red and green colours are commonly used as a convention for wrong and right, respectively. Hence the same idea translates to them being associated with the fakeness or realness of the news articles. Moreover, highlighting and pop-up boxes are standard methods in UI design and might help in external validation by the user, who can check the reasoning behind the decision and be used to further improve the system in case of incorrect tagging. The design of the explainer dashboard is mainly done to ensure the technical information is communicated with accuracy and clarity, and it is openly addressed to computer scientists.

## 5 Conclusion

We presented a novel fake news detection system which includes a set of capabilities able to overcome the limitations of the existing systems by exploiting both linguistic features extracted from benchmarking fake news datasets to analyse an article's text and complex features (e.g. sentiment, topic, attribution) computed for enriching the range of descriptors and enhance the classification performance. In addition, through the implementation of an explainable UI, we aim to provide fact-checkers and content managers with a reliable tool for cross-checking the validity of the system results. In the next steps, we plan to improve the system's response time and perform a user study to evaluate the overall user satisfaction in interacting with Facade and its UI.

## References

1. <https://www.politifact.com/truth-o-meter/>
2. <https://www.snopes.com/fact-check/>
3. <https://www.uvic.ca/ecs/ece/isot/datasets/fake-news/index.php>
4. <https://www.kaggle.com/datasets/jruvika/fake-news-detection>
5. <https://github.com/several27/FakeNewsCorpus>
6. <https://mpqa.cs.pitt.edu/corpora/mpqacorpora/>
7. <https://www.kaggle.com/datasets/datasnaek/mbti-type>
8. Allcott, H., Gentzkow, M.: Social media and fake news in the 2016 election. *J. Econ. Perspect.* **31**(2), 211–36 (2017)
9. Altmann, A., Tološi, L., Sander, O., Lengauer, T.: Permutation importance: a corrected feature importance measure. *Bioinformatics* **26**(10), 1340–1347 (2010)
10. Clayton, K., et al.: Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Polit. Behav.* **42**(4), 1073–1095 (2020). <https://doi.org/10.1007/s11109-019-09533-0>
11. Dale, R.: NLP in a post-truth world. *Nat. Lang. Eng.* **23**(2), 319–324 (2017)
12. Ferrara, E., Varol, O., Davis, C., Menczer, F., Flammini, A.: The rise of social bots. *Commun. ACM* **59**(7), 96–104 (2016)
13. Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E.: Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. *J. Comput. Graph. Stat.* **24**(1), 44–65 (2015)

14. Ha, L., Andreu Perez, L., Ray, R.: Mapping recent development in scholarship on fake news and misinformation, 2008 to 2017: disciplinary contribution, topics, and impact. *Am. Behav. Sci.* **65**(2), 290–315 (2021)
15. Loomba, S., de Figueiredo, A., Piatek, S.J., de Graaf, K., Larson, H.J.: Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nat. Hum. Behav.* **5**(3), 337–348 (2021)
16. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems*, vol. 30 (2017)
17. McGonagle, T.: “Fake news” False fears or real concerns? *Neth. Q. Hum. Rights* **35**(4), 203–209 (2017)
18. Molina, M.D., Sundar, S.S., Le, T., Lee, D.: “Fake news” is not simply false information: a concept explication and taxonomy of online content. *Am. Behav. Sci.* **65**(2), 180–212 (2021)
19. Pérez-Rosas, V., Kleinberg, B., Lefevre, A., Mihalcea, R.: Automatic detection of fake news. arXiv preprint [arXiv:1708.07104](https://arxiv.org/abs/1708.07104) (2017)
20. Shu, K., Cui, L., Wang, S., Lee, D., Liu, H.: defend: explainable fake news detection. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 395–405 (2019)
21. Yang, F., et al.: XFake: explainable fake news detector with visualizations. In: *The World Wide Web Conference*, pp. 3600–3604 (2019)
22. Zhang, X., Ghorbani, A.A.: An overview of online fake news: characterization, detection, and discussion. *Inf. Process. Manag.* **57**(2), 102025 (2020)
23. Zhang, X., Cao, J., Li, X., Sheng, Q., Zhong, L., Shu, K.: Mining dual emotion for fake news detection. In: *Proceedings of the Web Conference 2021*, pp. 3465–3476 (2021)
24. Zhou, X., Jain, A., Phoha, V.V., Zafarani, R.: Fake news early detection: a theory-driven model. *Digit. Threats Res. Pract.* **1**(2), 1–25 (2020)
25. Zhou, X., Zafarani, R., Shu, K., Liu, H.: Fake news: fundamental theories, detection strategies and challenges. In: *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pp. 836–837 (2019)
26. Zuo, C., Karakas, A., Banerjee, R.: A hybrid recognition system for check-worthy claims using heuristics and supervised learning. In: *CEUR Workshop Proceedings*, vol. 2125 (2018)