A Modality-level Explainable Framework for Misinformation Checking in Social Networks

Vítor Lourenço
Universidade Federal Fluminense
Dell Technologies
Niterói, Rio de Janeiro, Brazil
`vitorlourenco@id.uff.br`

Aline Paes
Universidade Federal Fluminense
Niterói, Rio de Janeiro, Brazil
`alinepaes@ic.uff.br`

Abstract

The widespread of false information is a rising concern worldwide with critical social impact, inspiring the emergence of fact-checking organizations to mitigate misinformation dissemination. However, human-driven verification leads to a time-consuming task and a bottleneck to have checked trustworthy information at the same pace they emerge. Since misinformation relates not only to the content itself but also to other social features, this paper addresses automatic misinformation checking in social networks from a multimodal perspective. Moreover, as simply naming a piece of news as incorrect may not convince the citizen and, even worse, strengthen confirmation bias, the proposal is a modality-level explainable-prone misinformation classifier framework. Our framework comprises a misinformation classifier assisted by explainable methods to generate modality-oriented explainable inferences. Preliminary findings show that the misinformation classifier does benefit from multimodal information encoding and the modality-oriented explainable mechanism increases both inferences' interpretability and completeness.

1 Introduction

The paper-free and web-based shift in the traditional press (*e.g.*, CNN, Globo, The New York Times) has contributed to democratizing access to daily news. Another recent behavior shift of digital press is to leverage social media vehicles (*e.g.*, Reddit, Twitter, Facebook, WhatsApp, TikTok) to spread information to time-critical news (*e.g.*, life-threatening situations). For example, Pew Research Center's 2021 news consumption report [17] indicates that 48% of Americans have social networks as a source of news, while Statista's News Consumption in Latin America Dossier [15] indicates that approximately two-thirds of people interviewed in Argentina, Brazil, Chile, and Mexico have social platforms as their news source.

With the advent of digital media, another observed trend is that everyone can share a piece of information. For example, the reports mentioned above do not separate the consumption of news coming from official presses or any other user. This way, the ease and agility with which information is disseminated on social networks come with their pitfalls, such as misleading information to totally false information. The widespread of false or misleading information (referred to as *misinformation* in this paper) is a rising concern with critical social impact, as seen during the past elections and the COVID-19 pandemic [12, 13, 18, 10, 11]. Accordingly, the Statista's Dossier, Latin America is the most affected region with misinformation, varying from health and politics to products and services.

In recent years, fact-checking organizations emerged (*e.g.*, Lupa [1], Aos Fatos [2], Chequeado [3]) to mitigate misinformation dissemination. These organizations usually rely on multiple sources and

---

[1] `https://lupa.uol.com.br/`   [2] `https://www.aosfatos.org/`   [3] `https://chequeado.com/`
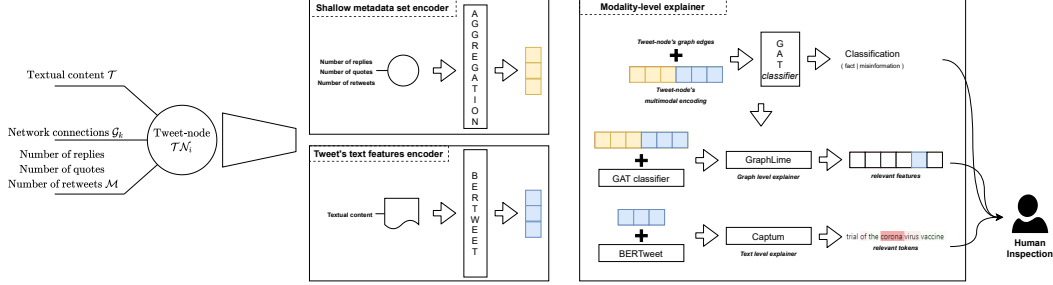
Figure 1: The proposed modality-level explainable framework. Red and green elements illustrate graph-based and text-based feature vectors, respectively.

documents – *i.e.*, multimodal, multilingual, and multi-topical data –, and human-driven verification to classify social media claims and news articles as trustworthy or misinformation, leading to a time-consuming task and, mostly, a bottleneck to overcome to have checked trustworthy information. To address the lack of scalability of traditional manual fact-checking, several work [18] propose methods for automatic fact-checking social media content. For instance, Shang et al. [10] propose a multimodal approach for COVID-19-related misinformation detection on TikTok, Shu et al. [12, 13] and Nielsen and McConville [8] propose the `FakeNewsNet` and `MuMiN`, respectively, which both are resources that links claims with Twitter posts.

Although automatic fact-checking improves scalability, the aforementioned approaches lack a key aspect towards providing an ultimate tool for specialists or fact-checking agencies: explainability. Explanatory factors are decisive for users' confidence and to address legal concerns. Like so, recent work explore explainable methodologies to meet this requirement. As examples, Shu et al. [14] and Kou et al. [6] propose explainable frameworks for fake news detection, and Kou et al. [5] and Shang et al. [11] design solutions for multimodal explainable misinformation checking.

In this work, we address social networks' misinformation checking from a modality-level explainable perspective. Different from related work that do not tackle explainability from both *interpretable* and *complete* aspects [1], our work combines two modality-oriented explainable methods to overcome the lack of interpretability from general multimodal solutions while keeping explanatory factors understandable (*i.e.*, complete). In detail, the called *modality-level explainable framework* aims to combine social networks' graph-based structure with their semantics and multimodal content targeting to classify social networks' misinformation posts while pointing out the explanatory factors that led to each classification. To do so, in our framework, we combine Graph Attention Network (GAT) classifier [16], a graph-based classier model, with GraphLime [2], a graph-based explainable model, and Captum [4], a text-based explainable tool, to the explained misinformation classification. Our preliminary findings show that the combination of different data modalities improves overall classification and a qualitative inquiry over provided explanations shows that they do contribute to overall classification understanding.

2 Methodology

This section presents our modality-level explainable framework. Our goal with this framework is to provide a solution to detect misinformation from multimodal, multilingual, and multi-topical Twitter posts (also known as *tweets*) and identify core features able to explain the classification, empowering human interpretation. We, first, state the misinformation tweet classification problem from a multimodal and explainable facet. Subsequently, we present our framework.

2.1 Problem statement

Given a tweet-node $\mathcal{TN}_i = (\mathcal{T}, \mathcal{G}_k, \mathcal{M}, \mathcal{C})$, where $\mathcal{T}$ is the tweet itself (textual content); $\mathcal{G}_k$ defines the local k-hop network connections of the tweet, such as replies, quotes, and retweets (*i.e.*, re-post of the same tweet); $\mathcal{M}$ is the tweet's shallow metadata set, which includes, as examples, the tweet's user owner, the number of likes, the location, among others; and $\mathcal{C}$ being the tweet's multimodal content, that might include images, videos, and audios. Our goal is to classify the tweet as *misinformation*

Table 1: GAT's performance score in the misinformation classification task regarding each input feature. We report the F1-score average and standard deviation values from five independent executions.

GAT's input tweet representation	F1-score
Graph-based features *only*	0.9225 ± 0.0260
Text-based features *only*	0.8942 ± 0.0097
Multimodal features	$\mathbf{0.9444 \pm 0.0052}$

(*i.e.*, $y_i = 0$) or *fact* (*i.e.*, $y_i = 1$). After, identify the most representative features within $\mathcal{TN}_i$ that led the classification.

2.2 Modality-level explainable framework

As illustrated in Figure 1, our framework comprises two main steps: tweet-node encoding; and misinformation detection and classification explainability. The tweet-node encoding is divided twofold to encode the tweet-node's shallow metadata set $\mathcal{M}$, and the tweet's textual content $\mathcal{T}$. In this work, the shallow metadata features are a result of the aggregation of the number of replies, quotes, and retweets, while the text-based features are generated by encoding the textual content. After, both feature vectors are concatenated to form a unique multimodal tweet-node's vector representation.

The second main step is misinformation detection with its explainability. Aiming to accurately identify misinformation, we encode the tweet-node's graph connections $\mathcal{G}_k$. After, we trained the classifier to label tweets between fact or misinformation, coupled with the multimodal tweet-node's vector representation. Finally, to provide explanations over the associated label, we employ graph-based and text-based explainable methods. The graph-based method is able to identify the most important features within the tweet-node's multimodal representation, while the text-based method highlights textual aspects that lead to a better understating of the associated label.

3 Experiments

We conduct experiments on the MuMiN [8] dataset, drawing preliminary insight into the proposed modality-level explainability of misinformation classification.

The MuMiN dataset is a public misinformation graph dataset that contains multimodal information from Twitter. Specifically, MuMiN associates multi-topical and multilingual tweets with fact-checked claims, and it also includes textual and visual content from tweets. We used the `MuMiN-small` version, which contains 2183 claims and 7202506 tweets. From the dataset, we filtered the English written tweets, four entities type (*Claim*, *Tweet*, *Reply*, and *User*), and six relations (*Posted*, *Mentions*, *Retweeted*, *Quote_Of*, *Reply_To*, and *Discusses*) to our analyses. Furthermore, we aim our efforts at the characterization of textual and relational features representation.

We developed our framework[4] using PyTorch [9]. As our misinformation classifier, we opt to use Graph Attention Networks (GAT) [16], which is a widely used GNN model. The classifier was trained using a single Nvidia RTX2060 GPU, with a learning rate of 0.005, 16-dimensional hidden layer, and Adam optimizer [3] for a total of 800 epochs. The textual features were encoded using HuggingFace's[5] pretrained BERTweet model [7], followed by a linear mapping of its original 768-dimensional space vector to a 3-dimensional space vector to match the dimensions of the shallow metadata set representation. The dataset split used was the same proposed in the MuMiN dataset.

3.1 Preliminary Findings

In the following, we present our preliminary findings. First, we quantitatively evaluate our misinformation classifier. To do so, we subject it to three different input scenarios: graph-based features *only*, text-based features *only*, and multimodal features, which is the concatenation of both graph-based and text-based features; and measure the obtained F1-score. In the second experiment, we perform a qualitative analysis of the identified explanatory features.

---

[4] `https://github.com/vitornl/MeLLL-MuMiN-explainable`
[5] `https://huggingface.co/docs/transformers/model_doc/bertweet`

Table 2: Report of two inferred cases with their classification, features, and explanatory factors identified by our framework.

Text	Shallow metadata set	Classification	GraphLime most representative features	Human interpretation
Great news! Carona virus vaccine ready. Able to cure patient within 3 hours after injection. Hats off to US Scientists. Right now Trump announced that Roche Medical Company will launch the vaccine next Sunday, and millions of doses are ready from it !!! VIA: @wajih79273180 https://t.co/BZJCLtwuXq	Number of retweets 26 Number of replies 42 Number of quotes 7	Misinformation	Number of retweets Number of replies	Users who retweet tend to spread misinformation
17,000 anti-vaxxers, anti-science, far-right & neo-Nazi organizations attend a protest against #coronavirus restrictions in #Berlin & defy #publichealth precautions. Let's see what happens to #Germany's #COVID19 case counts in the next 2-3 weeks. https://t.co/TD5xIoT5sV	Number of retweets 11 Number of replies 9 Number of quotes 2	Fact	Text	Word importance illustrated in Figure 2.

Figure 2: Word importance for text explainability of the second tweet from Table 2.

The results of the former experiment are shown in Table 1 and are the average and standard deviation values from five independent executions. It demonstrates that the multimodal feature vector enables the GAT model to better generalize the domain and, as result, achieves better overall performance in misinformation detection, which confirms the results obtained by Nielsen and McConville [8].

For the latter experiment, we qualitatively analyze the provided most important features and their capability of indeed correlating the tweet-node with its associated label. Our investigations are shown in Table 2, where the tweet-node text and shallow metadata set are displayed along with its GAT's classification, the GraphLime's [2] most representative features, and an interpretation of the explanatory factors. On the first entry, our GAT model classifies the tweet-node being as misinformation and GraphLime indicates that the most representative features are the number of retweets and the number of replies. After, analyzing both retweets and replies we identify that users that reply or retweet have other tweets classified as misinformation as well, suggesting a misinformation spread trend. The second entry was classified as being a fact while having the text embedding feature part being GraphLime's most representative feature. Like so, we further explore Captum's [4] textual factors that assist the label explanation. We observe that the "protest against #coronavirus" and "Let's see what happens to #Germany" parts corroborate the correct classification.

4   Conclusion

In this paper, we devised the modality-level explainable framework, a solution to detect misinformation from Twitter posts and provide explanatory factors by identifying the most relevant features that lead to the classification. Our framework comprises two main steps: encoding tweet-node information and detecting possible misinformation posts while providing relevant features that guide the classification explanation. Our preliminary findings show that the proposed framework confirms previous misinformation classification results by achieving the best when exposed to the multimodal vector space. Besides, our qualitative analysis demonstrates how the modality-level does contribute to overall classification understanding. Namely, the graph-based explainer is able to identify the most relevant features within the text-node representation vector, amid the text-based explainer enhances the classification understating at the tweet level.

Our perception of future work is twofold. At first, our goal is to enhance and further explore the proposed modality-level frameworks by incorporating other modality-driven explainable methods. We hypothesize that leveraging more modality might enhance the misinformation classifier and, mainly, contribute to improving both explanations' interpretation and completeness. Secondly, we target to explore how each modality influences topic-specific subjects (*e.g.*, COVID-19 pandemic, and elections).

References

[1] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. Explaining explanations: An overview of interpretability of machine learning. *Proc. - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018*, pages 80–89, 5 2018.

[2] Q. Huang, M. Yamada, Y. Tian, D. Singh, D. Yin, and Y. Chang. Graphlime: Local interpretable model explanations for graph neural networks. 1 2020. ISSN 2331-8422.

[3] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

[4] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch. *CoRR*, abs/2009.07896, 2020.

[5] Z. Kou, D. Yue Zhang, L. Shang, and D. Wang. Exfaux: A weakly supervised approach to explainable fauxtography detection. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 631–636, 2020.

[6] Z. Kou, L. Shang, Y. Zhang, and D. Wang. Hc-covid: A hierarchical crowdsource knowledge graph approach to explainable covid-19 misinformation detection. *Proc. ACM Hum.-Comput. Interact.*, 6 (GROUP), jan 2022.

[7] D. Q. Nguyen, T. Vu, and A. T. Nguyenh. Bertweet: A pre-trained language model for english tweets. pages 9–14. Association for Computational Linguistics (ACL), 11 2020.

[8] D. S. Nielsen and R. McConville. Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. *SIGIR 2022 - Proc. of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3141–3153, 7 2022.

[9] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[10] L. Shang, Z. Kou, Y. Zhang, and D. Wang. A multimodal misinformation detector for covid-19 short videos on tiktok. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 899–908, 2021.

[11] L. Shang, Z. Kou, Y. Zhang, and D. Wang. A duo-generative approach to explainable multimodal covid-19 misinformation detection. *Proc. of the ACM Web Conference - WWW 2022*, pages 3623–3631, 4 2022.

[12] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.

[13] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *CoRR*, abs/1809.01286, 2018.

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

[15] Statista. News consumption in latin america | statista, 2022. URL https://www.statista.com/topics/8083/news-in-latin-america.

[16] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph Attention Networks. *International Conference on Learning Representations*, 2018.

[17] M. Walker and K. E. Matsa. News consumption across social media in 2021 | pew research center, 2021. URL https://www.pewresearch.org/journalism/2021/09/20/news-consumption-across-social-media-in-2021/.

[18] X. Zhou and R. Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*, 53(5), sep 2020.