Research Proposal

Analysis of Misinformation Cascades across Traditional and Social Media

Martino Mensio

March 2019

1 Introduction

Misinformation is a grand societal issue and a complex ecosystem, with several actors that create and spread content that can be genuine or have negative connotations, such as false content or intention to harm. This situation is facilitated by the hyper-connectivity that allows people to share ideas and retrieve information from a wide variety of content providers.

In this scenario, there are several efforts from the scientific and journalistic community that try to fight the spread of misinformation from different perspectives. From the side of the journalistic community, we have manual verification both at article-level with reviews published by fact-checkers,[1] and also manual assessment of news sources.[2][3] On the other side we have a whole set of computational approaches that provide predictive models based on the content (the articles themselves) or the context where the information is shared. Belonging to this category, different works analyse the diffusion of misinformation on social media, finding cascading patterns that can be used to differentiate between genuine news and not [17]. What we see is that there is a separation between these two perspectives, and what is actually missing is an analysis of the cascades that covers also traditional news media, in order to include public figures and actors that operate outside of social networks.

This PhD proposal aims to perform an investigation based on (mis)information cascades, extending them from the *social*-only environment to a much wider one that includes also traditional news outlets. The Research Questions are two: How do the cascading models behave when we consider this mixed environment? Can we build a model of credibility of the actors involved based on the cascading patterns?

These models will be important for the consumers of information (both on social media and on traditional media) to help recognising the credibility of the actors involved and avoid being unwillingly part of the propagation. With the patterns detected it will possible to estimate the credibility also for the nodes for which there is not a manual assessment.

2 Related Work

This section analyses how existing approaches try to *i)* detect misinformation *ii)* analyse the flow of propagation and *iii)* assess the credibility of the news sources.

Starting by the *misinformation detection*, there is a wide literature that makes use of different strategies, manual and automated. Starting with the manual, fact-checking organisations are publishing every day reviews of claims,[4] debunking fake contents that emerge online. And tools

[1] https://ifcncodeofprinciples.poynter.org/signatories

[2] http://www.opensources.co/

[3] https://www.newsguardtech.com/

[4] https://schema.org/ClaimReview

to search and navigate through this wide set of reviews are being built[5] and integrated in major websites.[6][7] Beyond manual efforts, there are also automated techniques that can be used to spot occurrences of misinformation that have not yet been fact-checked. As some surveys aggregate [8, 9, 13, 19], there are different families of automated methods that can provide clues about the presence of misinforming content making use of a wide variety of features. Such features can be the writing *style* [18], relying on the assumption that misinforming content is different on the linguistic surface because of its intent to deceive. Or can be the semantic content used by *knowledge-based* approaches, that does not correspond to actual facts present in ontologies (such as [4, 16]) or in other articles about the same topic [11]. A last family of approaches tries to detect misinformation using indicators of *credibility*, basing on article-level features such as headline clickbaitness [3], or source-level reputation [7], or even based on the comments and people sharing it over the web [6].

Moving to the characterisation of the *flow of propagation*, there is a wide variety of analyses that find specific features of the cascading diffusion [17, 5, 15]. These features, beyond being described in the relative studies, have also been used to train automated detection algorithms [10]. In this field there are also specific studies that measure the impact of bot nets in the spread of misinformation [14].

Another group of works instead focuses on establishing the *credibility* of the sources, usually identified by their domain name. This assessment is performed manually by reporters and journalists that build lists of domains and annotate their properties. Among these, we can consider lists like *opensources*[8] or companies that evaluate journalistic criteria (credibility and transparency)[9] or even the International Fact Checking Network that reviews the applications of fact-checkers based on five principles (non-partisanship and fairness, transparency of sources, transparency of funding and organisation, transparency of methodology, and open and honest corrections policy). Or there are also sources that consider reviews from the users.[10]

The current problem of the existing approaches is that there is not a good integration between the analysis of propagation and the assessment of the credibility of the sources. This is caused by the limitation of the analysis of information cascades just inside a single social network. Furthermore, to the best of our knowledge, there is not a corresponding analysis done on the cascading of the news articles that considers the mutual influence of different sources.

3 Approach

This section describes the steps that will be needed in order to build a model of the misinformation cascades over traditional and social media. This model will be used to assess the credibility of the involved actors by looking at their connections with genuine and misinforming cascades.

The first step, finding occurrences labelled as misinforming and genuine (as control group), is required to have a list of what needs to be tracked. We will base this step on known instances (fact-checking based) to start with high fidelity data instead of relying already on predictive fallible models. This step will provide a set of labelled articles and claims, retrieved through aggregators of fact-checking instances.[5]

[5] https://toolbox.google.com/factcheck/about

[6] https://www.blog.google/products/search/fact-check-now-available-google-search-and-news-around-world/

[7] https://www.poynter.org/fact-checking/2019/youtube-is-now-surfacing-fact-checks-in-search-heres-how-it-works/

[8] http://www.opensources.co/

[9] https://www.newsguardtech.com/

[10] https://www.mywot.com/

The second step is tracking the same instance of information in different contexts: news outlets may spread and replicate contents and stories, and users can share and post over social networks as well as other websites or personal blogs. Given in input the true/fake instances, we will search for different occurrences of the same article on different platforms, querying social media APIs and news aggregators. Different approaches will be used to recognise the different occurrences of the same information instance. First of all, a lot of occurrences can be found by performing URL matching. URLs can be contained in articles and posts to let the readers explore the related documents. When this is not enough, because content gets republished on different websites without explicit reference of the source, a headline matching can be used. This matching can be total or partial depending on how much it is done with the purpose of hiding the fact that the content is copied. But there could also be cases where the headlines are too different but the content are very similar. In this case there exists also methods for automatically finding matches, based on word-similarity [2] or semantic comparison [12]. This last group of methods is the only applicable to spoken replication of content by public figures.

The third step corresponds to finding mentions iteratively: the content of the articles and posts is analysed to find references to other posts and actors. In the social network environment this corresponds to mentions, retweets, likes and links in the content. In the traditional news instead there can be references in the form of URLs, or citations in many different formats to other news articles. Once that the mentions are found, a possible further step would be to include a stance detection step to determine whether the current article is supporting, denying, or simply related to the considered mention, as in the task A of the RumourEval comptetition.[11] The whole step can be applied again iteratively in order to build chains of mentions.

Once the chains are retrieved, the fourth step consists of putting them together and build a graph of diffusion. In this graph the nodes will be actors (intended as profiles of social media, news outlets and other public figures that are mentioned) and contents (social media posts and traditional news articles). The edges, depending on the types of nodes that they link, will be of type `creates` (an actor creating a content), `references` (a content mentioning an actor or another content) and `influences` (an actor influencing another actor). While building this graph it will be important to extract correctly the actors for each content. For newspapers and media outlets this can be done by looking at the domain name when available, or the name itself of the newspaper. For the public figures this is a bit different because not all of them have a domain name associated with their personal website, but they may just use personal blogs or social network profiles. And some of them may just be referenced by name, with all the problems of disambiguation and homonymy.

The fifth step is the analysis of this graph and the extraction of cascading patterns. This analysis will be performed by doing two different types of comparison. To start with, the cascading of the misinforming occurrences will be compared against the one of the genuine news, to understand if there are different patterns as was found in social media by [17]. Then the other comparison is done between the spread on social media and on traditional media, to understand if the dynamics of diffusion and influence are different and how they interface. This experimentation will target specifically the first research question.

The last step of the proposed pipeline will set up a credibility measure for the actors based on the patterns identified. Looking at the incoming and outcoming edges on the diffusion graph and with a predictive model built on top of the identified patterns, it will be possible to have an estimation of the credibility of public figures, traditional news outlets and social media profiles. This model targets the second research question.

[11] `https://competitions.codalab.org/competitions/19938`

4 Evaluation and time plan

The evaluation will cover the models built in the last two steps.

The cascading patterns quality will be evaluated by considering their ability to distinguish between misinforming and genuine occurrences. The baseline of comparison will be the one evidenced in previous studies [17, 10]. To compare with [10] we will follow the same evaluation on the social-only dataset. The prediction of genuine and misinforming labels will be done cascade-wise and content-wise and the performance will be quantified by computing the ROC AUC of the classification. The evaluation will also be performed for the traditional media cascades and on the overall cascades. In this case there are no external baselines and the comparison will be done with respect to the results achieved in the social media context.

Instead for the evaluation of the models for actor credibility, the performances will be measured as the ability to predict indicators for both the news outlets and social media profiles. For the news outlets the gold values correspond to the manually annotated lists of source credibility, paying attention at how the mapping of terminology is performed along different dimensions (truthiness, credibility, transparency). Instead for the social media profiles the we will have a baseline established by other works in credibility assessment [1] and the performances will be evaluated by looking at measurements like F-measure and ROC AUC of the prediction.

The research plan is shown in Table 1

Time period	Activities
M1-M2	Step 1: misinformation collection
M3-M6	Step 2: build models to track the misinformation on different platforms
M5-M8	Step 3: build models to recognise mentions
M9-12	Paper 1: techniques for tracking spreading information on different channels
M9-M12	Step 4: model and build the graph of references
M13-M18	Step 5: Identify patterns of information and misinformation
M19-M24	Evaluation and comparison of the patterns
M19-M24	Paper 2: cascading patterns also outside social media
M25-M30	Step 6: build models for actor credibility
M27-M32	Evaluation of the credibility model
M27-M32	Paper 3: credibility estimation
M27-M32	Refinement of the models
M31-M36	Thesis writing

Table 1: The plan

References

[1] M. Alrubaian, M. Al-Qurishi, M. Al-Rakhami, M. M. Hassan, and A. Alamri. Reputation-based credibility analysis of twitter social network users. *Concurrency and Computation: Practice and Experience*, 29(7):e3873, 2017.

[2] D. Bär, T. Zesch, and I. Gurevych. Text reuse detection using a composition of text similarity measures. *Proceedings of COLING 2012*, pages 167–184, 2012.

[3] P. Bourgonje, J. M. Schneider, and G. Rehm. From clickbait to fake news detection: an approach based on detecting the stance of headlines to articles. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 84–89, 2017.

[4] G. L. Ciampaglia, P. Shiralkar, L. M. Rocha, J. Bollen, F. Menczer, and A. Flammini. Computational fact checking from knowledge networks. *PloS one*, 10(6):e0128193, 2015.

[5] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559, 2016.

[6] S. Dungs, A. Aker, N. Fuhr, and K. Bontcheva. Can rumour stance alone predict veracity? In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3360–3370, 2018.

[7] D. Esteves, A. J. Reddy, P. Chawla, and J. Lehmann. Belittling the source: Trustworthiness indicators to obfuscate fake news on the web. *arXiv preprint arXiv:1809.00494*, 2018.

[8] M. Fernandez and H. Alani. Online misinformation: Challenges and future directions. In *Companion of the The Web Conference 2018 on The Web Conference 2018*, pages 595–602. International World Wide Web Conferences Steering Committee, 2018.

[9] L. Graves. Understanding the promise and limits of automated fact-checking. *Factsheet*, 2:2018–02, 2018.

[10] F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein. Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673*, 2019.

[11] A. T. Nguyen, A. Kharosekar, M. Lease, and B. Wallace. An interpretable joint graphical model for fact-checking from crowds. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[12] V. Oleshchuk and A. Pedersen. Ontology based semantic similarity comparison of documents. In *14th International Workshop on Database and Expert Systems Applications, 2003. Proceedings.*, pages 735–738. IEEE, 2003.

[13] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*, 2017.

[14] C. Shao, G. L. Ciampaglia, O. Varol, A. Flammini, and F. Menczer. The spread of fake news by social bots. *arXiv preprint arXiv:1707.07592*, pages 96–104, 2017.

[15] C. Shao, P.-M. Hui, L. Wang, X. Jiang, A. Flammini, F. Menczer, and G. L. Ciampaglia. Anatomy of an online misinformation network. *PloS one*, 13(4):e0196087, 2018.

[16] B. Shi and T. Weninger. Fact checking in heterogeneous information networks. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 101–102. International World Wide Web Conferences Steering Committee, 2016.

[17] S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[18] W. Y. Wang. ”liar, liar pants on fire”: A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 422–426, 2017.

[19] X. Zhou and R. Zafarani. Fake news: A survey of research, detection methods, and opportunities. *arXiv preprint arXiv:1812.00315*, 2018.