Computational Fact Checking: A Content Management Perspective

Sylvie Cazalens
Université de Lyon, CNRS
INSA-Lyon, LIRIS, UMR5205
Villeurbanne,F-69621, France
sylvie.cazalens@insa-lyon.fr

Julien Leblay
AIRC, AIST
Tokyo, Japan
julien.leblay@aist.go.jp

Philippe Lamarre
Université de Lyon, CNRS
INSA-Lyon, LIRIS, UMR5205
Villeurbanne,F-69621, France
philippe.lamarre@insa-lyon.fr

Ioana Manolescu
Inria and LIX, CNRS and
Ecole Polytechnique, France
ioana.manolescu@inria.fr

Xavier Tannier
Sorbonne Université, Inserm,
LIMICS, Paris, France
xavier.tannier@sorbonne-universite.fr

ABSTRACT

Data journalism designates journalistic work inspired by digital data sources. A particularly popular and active area of data journalism is concerned with fact-checking. The term was born in the journalist community and referred the process of verifying and ensuring the accuracy of published media content; since 2012, however, it has increasingly focused on the analysis of politics, economy, science, and news content shared in any form, but first and foremost on the Web (social and otherwise). These trends have been noticed by computer scientists working in the industry and academia. Thus, a very lively area of digital content management research has taken up these problems and works to propose foundations (models), algorithms, and implement them through concrete tools.

Our tutorial: (*i*) Outlines the current state of affairs in the area of digital (or computational) fact-checking in newsrooms, by journalists, NGO workers, scientists and IT companies; (*ii*) Shows which areas of digital content management research, in particular those relying on the Web, can be leveraged to help fact-checking, and gives a comprehensive survey of efforts in this area; (*iii*) Highlights ongoing trends, unsolved problems, and areas where we envision future scientific and practical advances.

PVLDB Reference Format:
S. Cazalens, J. Leblay, P. Lamarre, I. Manolescu, X. Tannier. Computational Fact Checking: A Content Management Perspective. *PVLDB*, 11 (12): 2110-2113, 2018.
DOI: https://doi.org/10.14778/3229863.3229880

Proceedings of the VLDB Endowment, Vol. 11, No. 12

DOI: https://doi.org/10.14778/3229863.3229880

1. OUTLINE

In Section 1.1, we provide a short history of journalistic fact-checking and presents its most recent and visible actors, from the media and/or NGO communities. Section 1.2 discusses the scientific content management areas which bring useful tools for computational fact-checking.

1.1 Data journalism and fact-checking

While data of some form is a natural ingredient of all reporting, the increasing volumes and complexity of digital data lead to a qualitative jump, where technical skills, and in particular data science skills, are stringently needed in journalistic work.

A particularly popular and active area of data journalism is concerned with fact-checking. The term was born in the journalist community; it referred to the task of identifying and checking factual claims present in media content, which dedicated newsroom personnel would then check for factual accuracy. The goal of such checking was to avoid misinformation, to protect the journal reputation and avoid legal actions. Starting around 2012, first in the United States (FactCheck.org[1]), then in Europe, and soon after in all areas of the world, journalists have started to take advantage of modern technologies for processing content, such as text, video, structured and unstructured data, in order to automate, at least partially, the knowledge finding, reasoning, and analysis tasks which had been previously performed completely by humans. Over time, the focus of fact-checking shifted from verifying claims made by media outlets, toward the claims made by politicians and other public figures. This trend coincided with the parallel (but distinct) evolution toward asking Government Open Data, that is: the idea that governing bodies should share with the public precise information describing their functioning, so that the people have means to assess the quality of their elected representation. Government Open Data became quickly available, in large volumes, e.g. through `data.gov` in the US, `data.gov.uk` in the UK, `data.gouv.fr` in France etc.; journalists turned out to be the missing link between the newly available data and comprehension by the public. Data journalism thus found

[1] `http://factcheck.org`

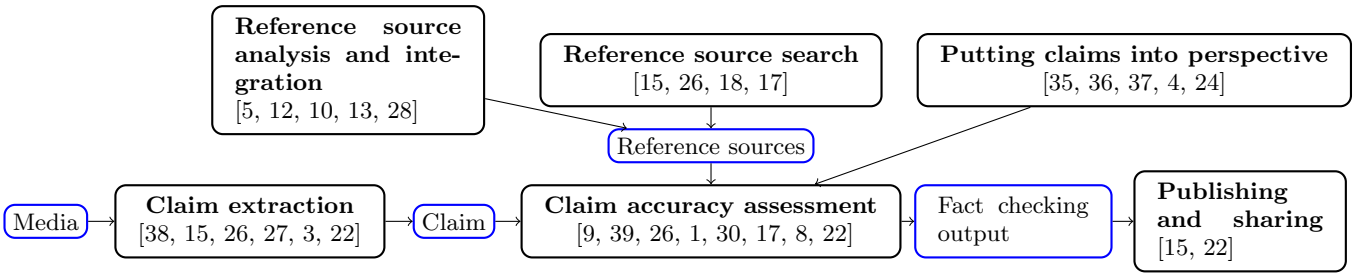

Figure 1: Fact checking tasks, ingredients, and relevant works: an overview.

one of its most useful incarnations in fact-checking based on digital content and tools; there are natural connections with investigative journalism, which also needs to identify, analyze and exploit complex databases. This has been illustrated most visibly in recent years by the Panama Papers[2] and Paradise Papers[3], investigations into tax evasion across the world. Beyond journalists, concerned citizens, NGOs such as FactCheck.org, and scientists such as those running `climatefeedback.org` also joined the discussion; this has enlarged the scope of journalistic fact-checking, beyond politics, to issues related to health (medical scandals), the environment (pollution through dangerous pesticides, or the controversy over climate change, studied in particular by ClimateFeedback mentioned above) and many others. Another parallel development is the massive production of fake news and influence steering through bot-generated content. While (typically false) propaganda information is not novel, the Web and the social media, amplified by the so-called "echo chamber" and "filter bubble" effects, have taken its scale to a higher order of magnitude; fake news production is quasi industrial[4]).

These aspects being noticed by computer scientists who, as citizens, are also eager to contribute to the way modern society works. An active research area has taken up these problems and works to propose foundations (models), algorithms, and implement them through concrete tools. The efforts have been many but scattered. Google, in particular, has recognized the usefulness and importance of fact-checking efforts, by making an effort to index and show them next to links returned by the users[5].

1.2 Related scientific areas

While a fully automatic approach to fact-checking is beyond reach (and probably not even desirable), several areas of data science contribute useful concepts and tools:

Data management, in the sense of persisting data and querying it: journalists need it both for the claims made (typically publicly through the Web) and for the reference data sources which they can use in their verification (such as reference statistic datasets published by government agencies). Yet, our interactions with journalists and fact-checkers highlight that establishing repositories of persistent data is not an obvious thing for them, especially that they may want to store data files, but also links, establish interconnections, annotate the data etc. We will briefly review the kind of data sources they have to deal with, and existing techniques which data management (and in particular Web data management) may have to offer.

Data integration allows exploiting together datasets of different origins and often independently produced. This plays a central role in analyses like the Panama and Paradise paper[6]. We will review the data integration architectures [14] (mostly focusing on data warehouses, mediators, data spaces [16] and data lakes [20]) and comment on their applicability to fact-checking scenarios we encountered. Still in a data integration scenario, a very relevant task is the selection of the best information sources to answer a specific query (in the classic scenarios) [7, 31], or to check a specific claim (in a modern fact-checking scenario) [22]. In a related field, truth discovery attempts to quantity the veracity of data when collected and merged from many, possible disagreeing, sources [11, 12].

Text analysis and information extraction, in particular through automated classification and learning, is gaining momentum as a way to cope with the huge number of documents published in social or mainstream media. In the context of the web, these techniques allow to go from unstructured or poorly structured text to structured knowledge bases which lend themselves more easily to fact-checking answers. Text analysis can be used to detect trends in news, extract the source of claims [23, 34, 33] or recognize rumors [6]. There have been some attempts at creating end-to-end fact validation systems, collecting and monitoring facts from online data, and looking for evidences to input claims [26, 21]. News analysis has established itself as a research topic on its own, covering news clustering over time, causality detection between news events or credibility analysis.

Natural Language Processing has many related subfields: textual entailment–comparing two portions of text and deciding whether the information contained in the first one can be implied from the second [9, 30]; stance detection–determining from a text whether it is in favor of a given target or against it [2]; entity linking–connecting an entity mention that has been identified in a text to one of the known entities in a knowledge base [29, 32].

Data and graph mining methods applied to structured and regular data enable the analysis of (static and streamed) information coming from the media; related work focus on very specific types of queries (e.g. checking that criminality rate has decreased during the mandate of M. X's as a mayor [21, 35]) or on tracking exceptional events [4]. The

[2] `https://panamapapers.icij.org/`
[3] `https://www.icij.org/investigations/paradise-papers/`
[4] See e.g., `http://cnnmon.ie/2GqfWX8`
[5] `https://developers.google.com/search/docs/data-types/factcheck`
[6] `http://bit.ly/2Drp4aJ`

context in which an information item is produced may hold valuable hints toward the trustworthiness of that information. Existing research on social network analytics may help identify communities of fake news producers, identify rumor spreaders, etc.

Machine learning is frequently leveraged to help classify published content according to their topic, to their likely trustworthiness or to their "checkworthiness" [6, 19, 22]. Journalists in particular strongly appreciate automated help to narrow the documents on which they should focus their verification effort. Fake news detection is now a very active field mobilizing a growing numbers of researchers, and is now the focus of international challenges[7,8].

Temporal and spatial aspects of the above: the news arena is by definition one of perpetual movement, and many areas of reality follow the same pattern; time is a natural dimension of all human activity. Facts can be true during a period of time and then become false. Also, many hoaxes are spread periodically, and many "news" can be false just because the fact they relate happened years ago.

Image and video processing and classification: pictures and videos are a very common way to disseminate fake news, either by lying about their provenance or date or creation, or by doctoring their content. *audio, image and video processing* have been very dynamic on this subject, notably through the field of multimedia forensics, leading to verification systems and services such as RevEye, Tineye or InVID; however, their techniques are very specific and we will not be able to cover them in a 3-hour tutorial.

2. TUTORIAL ORGANIZATION

The three-hour tutorial will be organized and structured following a set of stages involved in fact-checking work; these stages are outlined in Figure 1, which we borrow from [25]. Tasks involved in fact checking are shown in black boxes, together with their inputs and outputs (shown in blue); main relevant works from the literature appear in their respective tasks.

The central task is to assess the accuracy of a claim accuracy, based on reference sources; this takes as input the claim, and outputs a fact check result or analysis. The claim may have to be extracted from a text source, made available through some media, such as newspapers, social media, political or government communication etc. Reference source search may be needed to identify the reference sources most suited in order to check a given claim. An active area of fact checking work is concerned with putting claims into perspective by analyzing how claim validity is impacted by a slight change in the claim statement. Finally, content management techniques are also called upon to facilitate publishing and sharing fact checking outputs.

Each stage will be the topic of a dedicated section of the tutorial, where we highlight the challenges, outline solutions in the area, and point to remaining open problems.

3. ORGANIZERS

Sylvie Cazalens is an associate professor at INSA Lyon, LIRIS Lab since 2013 within the Database research group, her main research interests are on data integration and semantic interoperability.

Julien Leblay is a Research Scientist at the National Institute of Advanced Industrial Science and Technology (AIST) in Tokyo, Japan. His research interests cover data management and query processing in general, with a particular focus on applications to Web data, i.e., data typical found on web services and Open Data.

Ioana Manolescu is a senior researcher at Inria Saclay and Ecole Polytechnique. She is the lead of the CEDAR INRIA team focusing on rich data analytics at cloud scale. Her main research interests include data models and algorithms for fact-checking, performance optimizations for semistructured data and the Semantic Web, and distributed architectures for complex large data.

Philippe Lamarre is a professor at INSA Lyon since 2011. He leads the Database group of the LIRIS laboratory. His work is centered on data and knowledge base management in open distributed systems with special interest in query evaluation and autonomy of participants.

Xavier Tannier is a professor at Sorbonne Université and researcher at the LIMICS lab since 2017. He was associate professor at University Paris-Sud and researcher at LIMSI-CNRS from 2007 to 2017. His main field of research lies in natural language processing and text mining in large collections of documents.

Expertise on the topic J. Leblay and I. Manolescu have been among the first researchers considering fact-checking Web content from a data and knowledge management perspective, publishing a demonstration paper called "Fact-checking and analyzing the Web" in the ACM SIGMOD conference in 2013 [18]. The demonstration was subsequently shown at the Computation+Journalism conference held in New York, in October 2014. X. Tannier has many years of experience on conducting text analysis and NLP projects together with major media actors, notably AFP (Agence France Presse). More generally, he has also worked extensively on event extraction from journalistic content, text classification etc.

Subsequently, the authors have been working on a French research project called ContentCheck[9] dedicated to content management models, algorithms and tools for journalistic fact-checking, in collaboration with Les Décodeurs[10], a data journalism and fact-checking part of Le Monde, France's leading national newspaper. The present tutorial proposal is issued of ContentCheck joint research.

Inria and JSPS (Japan Society for the Promotion of Science) support this collaboration through a joint international associate team, WebClaimExplain.

4. ACKNOWLEDGMENTS

This work is partially supported by the French National Research Agency grant ANR-15-CE23-0025-01 (ContentCheck project), by the JSPS Bilateral Joint Projects fund and Inria's Associated Team program (WebClaimExplain team). Julien Leblay is supported by the KAKENHI grant number 17K12786.

[7] http://www.fakenewschallenge.org/
[8] https://herox.com/factcheck
[9] https://team.inria.fr/cedar/contentcheck/
[10] http://www.lemonde.fr/les-decodeurs/

5. REFERENCES

[1] E. Agirre, D. Cer, M. Diab, and A. Gonzalez-Agirre. *Sem 2013 Shared Task: Semantic Textual Similarity. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics and the Shared Task: Semantic Textual Similarity*, pages 32–43, 2013.

[2] I. Augenstein, T. Rocktäschel, A. Vlachos, and K. Bontcheva. Stance Detection with Bidirectional Conditional Encoding. In *EMNLP*, pages 876–885, 2016.

[3] M. Babakar and W. Moy. The state of automated factchecking. https://fullfact.org/media/uploads/full_fact-the_state_of_automated_factchecking_aug_2016.pdf, 2016.

[4] A. Belfodil, S. Cazalens, P. Lamarre, and M. Plantevit. Flash points: Discovering exceptional pairwise behaviors in vote or rating data. In *PKDD*, pages 442–458, 2017.

[5] A. Borodin, G. O. Roberts, J. S. Rosenthal, and P. Tsaparas. Link analysis ranking: Algorithms, theory, and experiments. *ACM Trans. Internet Technol.*, 5(1):231–297, Feb. 2005.

[6] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on Twitter. In *WWW*, pages 675–684, 2011.

[7] B. Catania, G. Guerrini, and B. Yaman. Context-dependent quality-aware source selection for live queries on linked data. In *EDBT*, pages 716–717, 2016.

[8] 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.

[9] I. Dagan, O. Glickman, and B. Magnini. The PASCAL Recognising Textual Entailment Challenge. In *PASCAL Challenges Workshop for Recognizing Textual Entailment*, 2005.

[10] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang. Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 601–610, New York, NY, USA, 2014. ACM.

[11] X. L. Dong, L. Berti-Équille, and D. Srivastava. Integrating conflicting data: The role of source dependence. *PVLDB*, 2(1):550–561, 2009.

[12] X. L. Dong, L. Berti-Equille, and D. Srivastava. Truth discovery and copying detection in a dynamic world. *PVLDB*, 2(1):562–573, 2009.

[13] X. L. Dong, E. Gabrilovich, K. Murphy, V. Dang, W. Horn, C. Lugaresi, S. Sun, and W. Zhang. Knowledge-Based Trust: Estimating the Trustworthiness of Web Sources. *PVLDB*, 8(9):938–949, 2015.

[14] X. L. Dong and D. Srivastava. Big data integration. In *ICDE*, pages 1245–1248, 2013.

[15] R. Ennals, B. Trushkowsky, and J. M. Agosta. Highlighting disputed claims on the web. In *Proceedings of the 19th international conference on World wide web*, pages 341–350. ACM, 2010.

[16] M. Franklin, A. Halevy, and D. Maier. From databases to dataspaces: A new abstraction for information management. *SIGMOD Record*, 34(4):27–33, Dec. 2005.

[17] D. Gerber, D. Esteves, J. Lehmann, L. Bühmann, R. Usbeck, A.-C. N. Ngomo, and R. Speck. DeFacto—temporal and multilingual deep fact validation. *Web Semantics: Science, Services and Agents on the World Wide Web*, 35:85–101, 2015.

[18] F. Goasdoué, K. Karanasos, Y. Katsis, J. Leblay, I. Manolescu, and S. Zampetakis. Fact checking and analyzing the web. In *SIGMOD*, pages 997–1000, 2013.

[19] C. Guggilla, T. Miller, and I. Gurevych. CNN- and LSTM-based claim classification in online user comments. In *International Conference on Computational Linguistics: Technical Papers (COLING 2016)*, pages 2740–2751, Dec. 2016.

[20] A. Halevy, F. Korn, N. F. Noy, C. Olston, N. Polyzotis, S. Roy, and S. E. Whang. Goods: Organizing Google's datasets. In *SIGMOD*, pages 795–806. ACM, 2016.

[21] N. Hassan, A. Sultana, Y. Wu, G. Zhang, C. Li, J. Yang, and C. Yu. Data in, fact out: Automated monitoring of facts by factwatcher. *PVLDB*, 7(13):1557–1560, 2014.

[22] N. Hassan, G. Zhang, F. Arslan, J. Caraballo, D. Jimenez, S. Gawsane, S. Hasan, M. Joseph, A. Kulkarni, A. K. Nayak, et al. Claimbuster: The first-ever end-to-end fact-checking system. *PVLDB*, 10(12):1945–1948, 2017.

[23] R. Kessler, X. Tannier, C. Hagège, V. Moriceau, and A. Bittar. Finding salient dates for building thematic timelines. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 730–739, 2012.

[24] J. Leblay. A declarative approach to data-driven fact checking. In *AAAI*, pages 147–153, 2017.

[25] J. Leblay, I. Manolescu, and X. Tannier. Computational fact-checking: problems, state of the art, and perspectives (tutorial). In *The Web Conference*, 2018.

[26] J. Lehmann, D. Gerber, M. Morsey, and A.-C. N. Ngomo. DeFacto-deep fact validation. In *International Semantic Web Conference*, pages 312–327. Springer, 2012.

[27] R. Levy, Y. Bilu, D. Hershcovich, E. Aharoni, and N. Slonim. Context Dependent Claim Detection. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, pages 1489–1500, Dublin, Ireland, 2014. Dublin City University and Association for Computational Linguistics.

[28] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han. A survey on truth discovery. *Acm Sigkdd Explorations Newsletter*, 17(2):1–16, 2016.

[29] X. Ling, S. Singh, and D. Weld. Design Challenges for Entity Linking. *Transactions of the Association for Computational Linguistics (TACL)*, 3:315–328, 2015.

[30] A. Lotan, A. Stern, and I. Dagan. TruthTeller: Annotating Predicate Truth. In *Proceedings of NAACL-HLT 2013*, page 752–757, Atlanta, USA, June 2013.

[31] T. Rekatsinas, A. Deshpande, X. L. Dong, L. Getoor, and D. Srivastava. Sourcesight: Enabling effective source selection. In *SIGMOD*, pages 2157–2160, 2016.

[32] W. Shen, J. Wang, and J. Han. Entity Linking With a Knowledge Base: Issues, Techniques, and Solutions. *Transactions on Knowledge and Data Engineering*, 2015.

[33] X. Tannier and F. Vernier. Creation, Visualization and Edition of Timelines for Journalistic Use. In *"Natural Language meets Journalism" Workshop next to IJCAI*, New York, USA, July 2016.

[34] G. B. Tran, E. Herder, and K. Markert. Joint graphical models for date selection in timeline summarization. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 1598–1607, 2015.

[35] Y. Wu, P. K. Agarwal, C. Li, J. Yang, and C. Yu. Toward computational fact-checking. *PVLDB*, 7(7):589–600, 2014.

[36] Y. Wu, P. K. Agarwal, C. Li, J. Yang, and C. Yu. Computational fact checking through query perturbations. *ACM Transactions on Database Systems (TODS)*, 42(1):4, 2017.

[37] Y. Wu, J. Gao, P. K. Agarwal, and J. Yang. Finding diverse, high-value representatives on a surface of answers. *PVLDB*, 10(7):793–804, 2017.

[38] H. Yu and V. Hatzivassiloglou. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 129–136, 2003.

[39] Álvaro Rodrigo, A. Peñas, and F. Verdejo. Overview of the Answer Validation Exercise 2008. In *Proceedings of the 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008*, pages 296–313, 2008.