Statistical Claim Checking: StatCheck in Action

Oana Balalau, Simon Ebel, Théo Galizzi, Ioana Manolescu, Quentin Massonnat
firstname.lastname@inria.fr
Inria and Institut Polytechnique de Paris
Palaiseau, France

Antoine Deiana, Emilie Gautreau, Antoine Krempf, Thomas Pontillon, Gerald Roux, Joanna Yakin
firstname.lastname@radiofrance.com
FranceInfo, Radio France
Paris, France

ABSTRACT

Fact-checking is a staple of journalists' work. As more and more important data is available in electronic format, computational fact-checking, leveraging digital data sources, has been gaining interest. A particular class of interesting data sources are *statistics*, that is, numerical data compiled mostly by governments, administrations, and international organizations.

We propose to demonstrate StatCheck, a fact-checking system specialized in the French media arena. StatCheck builds upon a prior pipeline [CMT17, CMT18, DCMT19] for fact-checking statistical claims against the INSEE national statistic institute's data and reports. In collaboration with factchecking journalists from France Info (Radio France), we have revisited and improved this pipelined, and enlarged its database by an order of magnitude by adding all Eurostat statistics. StatCheck also includes two novel statistic claim extraction modules, generalizing and improving over [DCMT19]. Based on the journalists' feedback, we built a new user interface, suiting better their needs. We will showcase StatCheck on a variety of scenarios, where statistical claims made in social media are checked against our statistic data. StatCheck has been featured in two recent publications [BEG$^+$22b, BEG$^+$22a].

1 INTRODUCTION

Professional journalism work has always involved verifying information with the help of trusted sources. In recent years, the proliferation of media in which public figures make statements, including traditional media available online, as well as social media, has lead to an explosion in the amount of content that may need to be verified in order to distinguish accurate from inaccurate, and even potentially dangerous, information.

Computational fact-checking is a growing, multidisciplinary field [CLL$^+$18, NCH$^+$21], with new meeting venues such as the "Global F<act" and "Truth and Trust Online". The main tasks of a fact-checking system are: identifying the claims made in an input document, finding the relevant evidence from a reference dataset, and optionally producing an automated verdict or if not, letting an end user decide on the truthfulness of the claim. Recent systems proposed in this area include [HZA$^+$17, KSPT20, PMYW18].

We propose to demonstrate StatCheck, a fact-checking system specialized in the French media arena. Differently from the above mentioned systems, StatCheck also includes a claim detection step, while assuming that an end user will decide if a claim is true or not based on the evidence retrieved. Another specificity is our focus on checking statistical claims, with the help of large public statistic databases (specifically, INSEE and Eurostat); the nature and organization of these databases raises specific challenges when searching for the statistic most appropriate to check a given claim.

Architecture and outline The overall architecture of our platform is presented in Figure 1, based on which we present the outline of this paper. To help journalists even more in their work to fact-check claims made in the public space, also ingest them in our system (by subscribing to media sources, or allowing users to upload their own content), and apply Natural Language Processing to identify, from their textual content, statistical claims.

2 SYSTEM DESCRIPTION

2.1 Statistic data acquisition and storage

INSEE publishes each statistic report as an HTML page that links to **statistic tables**, which may be in Excel (the most frequent case) or in HTML. The tables are not relational. On one hand, they have human-understandable header cells not only for each column (as is the case for a relational table), but also for each line. From this perspective, a statistic file resembles more a multidimensional aggregate query result. On the other hand, many tables feature *hierarchical (nested) headers*: for instance, a header cell corresponding to "Paris (75)" may appear as a child of another header cell corresponding to "Île-de-France".

While revisiting the platform, we re-crawled the INSEE Web site up to May 2022, leading to 60,002 Excel files and 58,849 HTML files. We then extract statistic tables from those files and index them in a custom compact format. Subsequently, we incrementally re-crawl the Web site each night to retrieve and ingest the possible studies published every day. We also added a new corpus of reference statistics, namely Eurostat: (*i*) 6,803 data tables; these are two-dimensional tables with line headers, row headers, and no nesting in the headers. Each header is a concatenation of a set of dimension values, e.g., EU15_NO as a line header corresponds to the value of metric "Union européenne - 15 pays (1995-2004) et Norvège" (*ii*) 580 dictionaries that map 243,083 statistical concepts codes into natural-language descriptions. The data files range from 2 lines, to 37 million lines (this largest file holds information about farmers and their agricultural properties across Europe). Together, the Eurostat data files total 414.908.786 lines. We store Eurostat data as plain files, and developed a specialized index to search in them [BEG$^+$22b, BEG$^+$22a].

2.2 Statistic Search

Our goal is to find the most relevant datasets for a given keyword query $Q = k_1, k_2, \ldots, k_n$. We also aim to return the finest level of granularity: a row, column, or cell, if it can be found to exactly

Infox sur Seine, March 22–23, 2023, Paris, France
2018. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

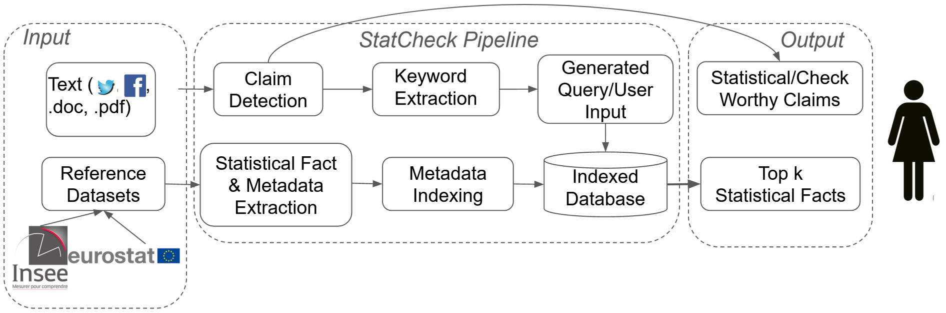

Figure 1: Statistical fact-checking architecture overview.

answer the query. Results of the query are usually numbers, thus we need to interpret them through metadata, found in: the title, column headers, row headers, or comments associated to the file.

For each dataset, we split its metadata into tokens and use a Word2Vec model to find the 50 closest semantically related tokens for each term in the metadata. We store these tokens in a term-location index I_{TL}. To determine the relevant datasets, we look up the query keywords in the index and rank the datasets based on a relevance score introduced in [CMT18]. The score is calculated based on the semantic distance between the query keywords and the datasets' metadata and their locations in the metadata.

2.3 Claim Detection

A claim is a statement to be validated, that is we aim to establish if it is true of false. The validation is achieved by finding related statements, called evidence, which back up or disprove the claim. In our work, the claims are detected in an input text, while the evidence is retrieved from a set of trusted sources, our reference datasets. Our platform detects claims from text stored in *.txt*, *.odt*, *.docx* or *.pdf* files, and from the Twitter posts of public figures. For posts, our platform retrieves regularly the most recent updates of a predefined group of users.

2.3.1 Statistical Claim Detection. In [DCMT19], the authors introduced a statistical claim detection method that given an input set of statistical entities (e.g. chômage, coefficient budgétaire) and a sentence, it retrieves all the *statistical statements* of the form ⟨`statistical entity, numerical value and unit, date`⟩ present in the sentence. *The statistical statement, if present, represents the statistical claim to be verified.* The statistical entities and units are retrieved using exact string matching, while the date is extracted using HeidelTime [SG10], a time expression parser. If no date is found by the parser, the posting timestamp is used. More context about the claim to be verified is found using a Named Entity Recognition (NER) model, which returns organizations and locations. We note, however, that the organization and location are optional, while a statistical statement is not complete without one of its three elements. The initial statistical entity list is constructed from the reference datasets by taking groups of tokens from the headers of tabels, we refer to [DCMT19] for more details.

2.3.2 Check-worthy Claim Detection. To complement the statistical claim detection model, we developed a model that is not conditioned on a set of initial statistical entities. The model classifies a sentence as check-worthy or not, where check-worthiness is defined as *sentences containing factual claims that the general public will be interested in learning about their veracity* [AHLT20].

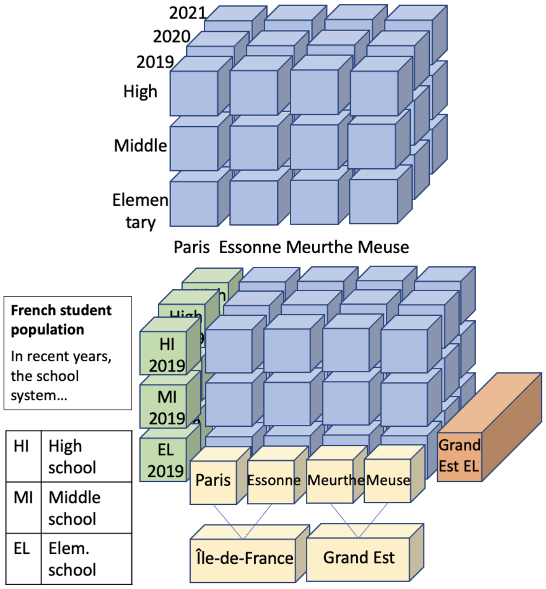

Figure 2: Multidimensional statistic data: conceptual view (top), structure of actual published dataset (bottom).

3 DEMONSTRATION SCENARIOS

Our system is developed in Python and deployed on a Unix server, accessible via the Web. Demonstration attendees will be able to: (*i*) Ask queries in the statistic search interface, and inspect the results, at the level of cell, line, or column, together with their metadata from the original statistic site (INSEE or Eurostat); (*i*) Visualize incoming social media messages (as they arrive in real-time and are stored and analyzed by our platform), in order to see the statistical mentions and claims deemed potentially check-worthy, identified in these messages. Note that the system also proposes candidate search queries for the statistic search interface. (*iii*) Select various options (restrict to numerical claims or not, include statements about the future or not, include first-person texts or not, etc.) and see their impact on the claim extraction output. (*iv*) Write their own text and/or suggest other content to be processed by our analysis pipeline (Section 2.3).

ACKNOWLEDGEMENTS

We thank Radio France, France Info, and in particular Antoine Krempf and his *Le vrai du faux* team, for their valuable feedback. This work is partially funded by AI Chair SourcesSay project (ANR-20-CHIA-0015-01).

REFERENCES

[AHLT20] Fatma Arslan, Naeemul Hassan, Chengkai Li, and Mark Tremayne. A Benchmark Dataset of Check-worthy Factual Claims. In *14th International AAAI Conference on Web and Social Media*. AAAI, 2020.

[BEG+22a] Oana Balalau, Simon Ebel, Théo Galizzi, Ioana Manolescu, Quentin Massonnat, and al... Fact-checking Multidimensional Statistic Claims in French. In *TTO 2022*, Truth and Trust Online, 2022.

[BEG+22b] Oana Balalau, Simon Ebel, Théo Galizzi, Ioana Manolescu, Quentin Massonnat, and al... Statistical Claim Checking: StatCheck in Action (demonstration). In *CIKM 2022*, 31st ACM International Conference on Information and Knowledge Management, 2022.

[CLL+18] Sylvie Cazalens, Philippe Lamarre, Julien Leblay, Ioana Manolescu, and Xavier Tannier. A content management perspective on fact-checking. In *WWW (Companion Volume)*, pages 565–574. ACM, 2018.

[CMT17] Tien Duc Cao, Ioana Manolescu, and Xavier Tannier. Extracting linked data from statistic spreadsheets. In *International Workshop on Semantic Big Data*, International Workshop on Semantic Big Data, pages 1 – 5, 2017.

[CMT18] Tien-Duc Cao, Ioana Manolescu, and Xavier Tannier. Searching for Truth in a Database of Statistics. In *WebDB*, pages 1–6, 2018.

[DCMT19] Tien Duc Cao, Ioana Manolescu, and Xavier Tannier. Extracting statistical mentions from textual claims to provide trusted content. In *NLDB*, 2019.

[HZA+17] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, Vikas Sable, Chengkai Li, and Mark Tremayne. Claimbuster: The first-ever end-to-end fact-checking system. *PVLDB*, 10(12):1945–1948, 2017.

[KSPT20] Georgios Karagiannis, Mohammed Saeed, Paolo Papotti, and Immanuel Trummer. Scrutinizer: Fact checking statistical claims. *Proc. VLDB Endow.*, 13(12):2965–2968, 2020.

[NCH+21] Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. Automated fact-checking for assisting human fact-checkers. In *IJCAI*, pages 4551–4558, 2021. Survey Track.

[PMYW18] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In *EMNLP*, pages 22–32, 2018.

[SG10] Jannik Strötgen and Michael Gertz. Heideltime: High quality rule-based extraction and normalization of temporal expressions. In *Int'l. Workshop on Semantic Evaluation*, pages 321–324, 2010.