

Automated Fact Checking in the News Room

Sebastião Miranda
Priberam Labs
Lisbon, Portugal
ssm@priberam.com

David Nogueira
Priberam Labs
Lisbon, Portugal
dan@priberam.com

Afonso Mendes
Priberam Informática, S.A.
Lisbon, Portugal
amm@priberam.com

Andreas Vlachos
Dep. of Computer Science and
Technology, University of Cambridge
Cambridge, United Kingdom
andreas.vlachos@cst.cam.ac.uk

Andrew Secker
BBC News Labs
London, United Kingdom
andrew.secker@bbc.co.uk

Rebecca Garrett
BBC Monitoring
London, United Kingdom
becky.garrett@bbc.co.uk

Jeff Mitchel
Psychology Dep., University of Bristol
Bristol, United Kingdom
jeff.mitchell@bristol.ac.uk

Zita Marinho
Priberam Labs
Institute of Systems and Robotics, IST
Lisbon, Portugal
zam@priberam.com

ABSTRACT

Fact checking is an essential task in journalism; its importance has been highlighted due to recently increased concerns and efforts in combating misinformation. In this paper, we present an automated fact checking platform which given a claim, it retrieves relevant textual evidence from a document collection, predicts whether each piece of evidence supports or refutes the claim, and returns a final verdict. We describe the architecture of the system and the user interface, focusing on the choices made to improve its user friendliness and transparency. We conduct a user study of the fact-checking platform in a journalistic setting; we integrated it with a collection of news articles and provide an evaluation of the platform using feedback from journalists in their workflow. We found that the predictions of our platform were correct 58% of the time, and 59% of the returned evidence was relevant.

CCS CONCEPTS

• **Computing Methodologies** → **AI**; *NLP*; Information Extraction; • **Human Centered Computing** → **HCI**; *HCI design and evaluation methodologies*; User studies; Usability testing; • **Information Systems** → **Information Retrieval**; *Evaluation of retrieval results*; Presentation of retrieval results; *Retrieval tasks and goals*; Question answering.

KEYWORDS

Fact Checking; Computational Journalism; Media Tools

ACM Reference Format:

Sebastião Miranda, David Nogueira, Afonso Mendes, Andreas Vlachos, Andrew Secker, Rebecca Garrett, Jeff Mitchel, and Zita Marinho. 2019. Automated Fact Checking in the News Room. In *Proceedings of the 2019 World Wide Web Conference (WWW '19)*, May 13–17, 2019, San Francisco, CA, USA. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3308558.3314135>

1 INTRODUCTION

Grounding information on reliable sources is a daunting experience, given the increasing amount of information circulating the web and other media platforms. Nevertheless, checking the veracity of claims is crucial for preserving trust in news sources. Manual verification of claims is a tedious task, that consumes a lot of time and effort from journalists and professional fact-checkers, and it typically requires searching for specific entities and content over large amounts of unstructured text [2, 4].

The rising interest in fact checking has led to the development of a number of approaches and tools automating the task or parts of it, with the motivation of facilitating the work of journalists, and interested readers more broadly [3, 9]. The most popular effort, ClaimBuster [5], proposed a fact checking platform which detects factual claims that are worth checking and then uses APIs to query search engines and databases (Google and Wolfram Alpha respectively). It also compares claims against previously fact checked ones in its database. The latter approach is also used by the system developed by Full Fact, Live, which is used to fact check repeated or paraphrased claims.¹ Neither of these approaches is able to fact check previously unchecked claims, while the queries through existing commercial APIs are not tailored to fact checking, thus retrieving information that is not necessarily relevant. Other approaches rely instead on the detection of rumours based on the spread of readership over social media [7]. However, a rumorous claim is not necessarily false, and vice versa [16]. There is also work that checks claims against tables, such as those released by

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

WWW '19, May 13–17, 2019, San Francisco, CA, USA

© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.

ACM ISBN 978-1-4503-6674-8/19/05.

<https://doi.org/10.1145/3308558.3314135>

¹<https://fullfact.org/blog/2017/jun/automated-fact-checking-full-fact/>

Claim:

Tesla builds car factory in **Shanghai**.

Evidence:

Electric carmaker **Tesla** has signed an agreement with Chinese authorities to build a factory in **Shanghai**. “We hope it will be completed very soon,” **Tesla** chief **Elon Musk** said.

Figure 1: Example of claim and evidence. Extracted named entities in bold: organizations (purple), locations (blue) and people (green).

the World Bank², using trained classifiers to select the appropriate tuple [10]. However such approaches are typically restricted to fact checking numerical claims against tabular sources. Finally, some approaches aim at providing a pipeline of tools for information retrieval [14], but do not go as far as to provide an actual fact check mechanism.

In this work, we propose an automated fact checking platform that checks claims by identifying sentences providing evidence in a large document collection. These sentences are classified as supporting, refuting or only related to the claim, and then combined into a final verdict using a state-of-the-art neural network-based approach [15]. The model is trained on the recently released FEVER dataset [11], a large scale fact checking dataset derived from Wikipedia comprising 185K claims. Unlike previous work, our model is able to check novel claims without relying on a database of fact checks. Also, the evidence retrieved and classified as supporting or refuting provides a justification of the verdict, which is likely to be relevant not only to assess the overall correctness of the platform, but also as part of the fact checking research conducted by journalists. Our retrieval model considers more information about a specific claim than generic search engines by leveraging information from words and named entities present in the dataset to obtain the best matching evidence for each claim (see Figure 1).

We thoroughly evaluate the platform through user testing by journalists from the British Broadcasting Corporation (BBC) in the context of their workflow. Of the total 488 evidence passages retrieved by the system, the journalists reported that 58% were relevant and 59% were accurately classified as SUPPORTS/REFUTES/OTHER.

Our platform can be applied to document collections beyond the ones used in this paper and our findings should help inform future research in automated fact checking and computational journalism more broadly.

2 FACT CHECKING SYSTEM

Our fact-checking system comprises three main components: a document retrieval step, a sentence ranking, and a classification model, as shown in Figure 2.

Initially we retrieve documents from a collection of news articles, using a customized search engine based on inverse indexing and retrieval using the Okapi-BM25 algorithm [8]. Data is incrementally indexed from word and entity-level inverted indexes. The document retrieval component searches for documents whose features best match the claim. In the end of this step, we end up with approximately 10K documents related to the claim.

²<https://data.worldbank.org/>

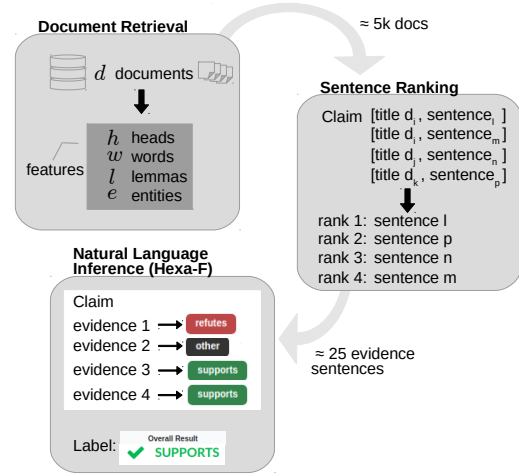


Figure 2: Fact-checking model comprised of three components: a) Document retrieval (top left), b) Sentence ranking (middle right), c) NLI prediction model (bottom).

Following this, we rank the sentences in these documents according to predefined token and feature matching rules. We compute the cosine similarity of the claim with each best candidate sentence, using word embeddings trained on the One Billion Word Benchmark corpus [1], and select those above a threshold tuned during development. This step aims at providing only sentences with high relevance to the claim, reducing the number of potential evidence that may be only marginally related with the claim.

In the final component, we classify each of the extracted candidate evidence in terms of whether they support, refute or are just related to the claim (other). We employ the natural language inference (NLI) model from the Hexa-F system [15] (one of the best performing systems in the FEVER shared task [12]) to classify the relation between the selected evidence sentences and the claim, one of SUPPORTS/REFUTES/OTHER, and a similar label which expresses whether the combined set of evidence sentences SUPPORTS, REFUTES or is simply related to the claim.

2.1 Engineering Considerations

In the first document retrieval component, we use inverted indexes to extract relevant documents, in Figure 3. We select all documents that match features in the claim, such as lemmas, words and extracted named entities [6]. This document retrieval step takes about 50 ms to retrieve around 5k documents out of a dataset with about 445k documents.

In the second component, we rank all the sentences from the 5k documents based on how well each sentence feature $\phi(s_i)_j$ matches the claim $\phi(c)_j$, using a positional ranking approach. We use a ranking score based on ordered distances between N matching features in the sentence i : $S_1(s_i, c) = \sum_{j=1}^N \exp(-d_{i,j})$, where $d_{i,j} = \text{pos}(\phi(s_i)_j) - \text{pos}(\phi(s_i)_{j-1})$ represents the word distance between the position (pos) of each j -th feature. This score is maximized if all words in the sentence match the claim exactly; it decreases exponentially with the word distance between matches. This part takes on average 336 ms. Next, we filter the sentences based on

the following rules: we keep only those sentences with length < 500 words; those that contain all the entities mentioned in the claim, as well as novel words that were not previously seen in previously selected sentences (less than 90% overlap with all words previously encountered). The novelty filter increases the diversity in the evidence passed to the entailment step.

In the end of the second component, we re-rank the sentences by averaging the feature matching score of each sentence S_1 and the cosine similarity between the claim and the sentence embeddings $S_2 = \cos(s_i, c)$. We obtain these embeddings via a weighted average of all words in the sentence, weighted by term-frequency/inverse document frequency. To improve the performance of the sentence re-ranking component, we added the title of the document to each sentence and considered the weighted sum of all words in the title and sentence combined. This part takes about 652 ms, although it could be easily parallelizable. On average, 76 sentences are retrieved per claim after these steps.

We further removed all sentences with an averaged similarity score below a given threshold $((S_1 + S_2) / 2 < 0.6)$, to ensure high quality evidence. In the end, about 25 sentences on average are selected. After re-ranking, all selected sentences are used as input in the NLI model (third component) which takes about 738 ms to predict labels for each of the sentences and the overall label for the claim.

3 USER INTERFACE

The interface allows end-users to input a claim (black bar at the top of Figure 3), and receive a set of evidence sentences as output. The evidence is displayed in three columns: the top five ranking sentences that are in favour of the claim on the left, the top five that are against the claim in the middle, and the top five other sentences related to the claim on the right. In the bottom, a final overall label is presented to the user as either other, supporting or refuting the claim (label at the bottom). The interface allows users to directly evaluate it, providing feedback for evidence sentence w.r.t. the correctness of the label ("correct label?") and its relevance ("relevant?"), as well as the correctness of the overall prediction. We provide a video with an example of a user interacting with the platform via the interface in <https://vimeo.com/309336679>.

4 USER EVALUATION DESIGN

11 BBC journalists provided feedback on the overall system and on the classification model. The journalists were asked to interact with the system by providing factual claims and evaluating the output of the model. For each claim, up to 15 evidence sentences were presented to them, 5 per category, each classified as SUPPORTS, REFUTES or OTHER. For each sentence, the journalists provided feedback on two aspects: relevance and correctness.

To assess correctness, for each evidence sentence returned by the system, the journalist inputs the label which, according to his/her research on the subject, would be the correct one via the buttons in the "correct label" box (see Figure 3). This measures primarily the accuracy of the entailment component of the system, assuming that the sentences returned are all related. For the final classification of the model (see Figure 3 (bottom)) they also assessed the overall prediction, to whether the claim was globally supported or refuted considering all the retrieved evidence.

Precision	Class			
	SUPPORTS	REFUTES	OTHER	ALL
Relevant	71%	69%	49%	59 %
Evidence Correctness	48%	27%	70%	58 %
Global Correctness	56%	26%	31%	42 %

Table 1: User evaluation on the full dataset.

To assess relevance, the journalists were also asked to provide feedback as to whether they found each sentence returned relevant (see the buttons in the "relevant ?" box in Figure 3). This part of the feedback aims to evaluate the quality of the retrieved evidence: whether it helps the journalists fact-check the input claim, regardless of the classification label attributed by the system. It also serves as a proxy to measure the precision of the retrieval component, as all sentences shown to the journalists should be relevant (ideally).

The journalists also had access to (i) a Question Answering (QA) [13] tool that could serve as an additional source of information, (ii) the full document's text with annotated entities [6]. The QA implemented in the platform was used by the journalists, but was not evaluated in this paper. We provide an example of the additional information, available for each extracted evidence sentence in Figure 4.

5 RESULTS

Table 1 summarizes the results of the system is the user evaluation conducted. We show precision for each class (row:Relevant) measuring the proportion of the retrieved evidence sentences that were deemed relevant by the journalists, and precision for the predictions of the platform for the relation of each evidence sentence to the claim (row:Evidence) and the global prediction for the claim taking all the evidence into account (Overall result presented in Figure 3) (row:Global). The platform was evaluated on 67 claims in total, with a total of 488 evidence sentences retrieved: 30% supporting, 14% refuting and 56% related to the claim, as classified by our platform.

In 71% of the claims checked by the journalists, it was reported that the evidence shown in the SUPPORTS column was relevant, and so was for 69% of the evidence in the REFUTES columns. We consider these results to be very encouraging, given the difficulty of the task. Retrieving evidence that contradicts a given claim, is not usually as simple as retrieving related evidence by feature matching. Introducing different information, e.g. entities, dates, actions, etc., that could refute the claim requires more complex language understanding methods. We further observe that the precision in predicting evidence as SUPPORTS is 48% in the full dataset and increases to 67% in the subset of evidence deemed relevant (not shown in the table). The same trend is observed for the REFUTES label from 27% to 39%. This result suggests that the retrieval component still requires some improvement, especially for retrieving evidence refuting the claim. Strategies beyond feature matching are needed to improve the retrieval of relevant but opposing arguments. The precision of the classifier predicting the global label of the claim given the evidence also requires further improvement.

Additionally, we received textual feedback from the journalists about the overall quality of the platform. They mentioned that it is helpful for fact checking, despite not being always accurate, both in the retrieval of relevant evidence and in their evaluation. An interesting remark mentioned possible improvements for handling

Claim

Russia meddled with US elections

Q

BBC (EN)

Supports

Refutes

Other

Evidence

✓ Not climate change or trade, but why didn't John Podesta give a server that wasn't his to the CIA." Several American intelligence agencies have concluded Russia meddled in last year's US elections, but Mr Trump has as recently as this week said he thinks other countries could have been responsible. [2017-07-07] BBC

correct label ?

relevant ?

supports

refutes

other

yes

no

✓ Kosachev: "No evidence" Russia meddled in US election [2017-09-26] BBC

correct label ?

relevant ?

supports

refutes

other

yes

no

✓ Trump had come under fire for defending Russia against claims that they meddled in the 2016 US presidential elections. [2018-07-19] BBC

correct label ?

relevant ?

supports

refutes

other

yes

no

✓ Meanwhile, a US Senate committee announced it would investigate claims Russia meddled in the election. " [2017-01-14] BBC

correct label ?

relevant ?

supports

refutes

other

yes

no

✓ Questions surrounding Russia's role in last year's US elections and allegations of collusion involving Donald Trump's campaign team have dogged his presidency. [2017-11-12] BBC

correct label ?

relevant ?

supports

refutes

other

yes

no

Evidence

✗ Vice-presidential nominee Mike Pence quickly stated there will be "serious consequences" if Russia attempts to influence US elections. [2016-07-27] BBC

correct label ?

relevant ?

supports

refutes

other

yes

no

✗ The new US ambassador to the UK, Woody Johnson, has said Theresa May was right to air her grievances about Russia's potential influence in the US elections. [2017-11-14] BBC

correct label ?

relevant ?

supports

refutes

other

yes

no

✗ The arrest also came days after the justice department charged 12 Russian intelligence officers with hacking Democratic officials in the 2016 US elections. [2018-07-16] BBC

correct label ?

relevant ?

supports

refutes

other

yes

no

✗ Facebook says it has removed 32 accounts and pages believed to have been set up to influence the mid-term US elections in November. [2018-07-31] BBC

correct label ?

relevant ?

supports

refutes

other

yes

no

✗ Mr Coats said Russia was not the only country that was trying to intervene in US elections, without elaborating. [2018-08-02] BBC

correct label ?

relevant ?

supports

refutes

other

yes

no

Evidence

✗ Russia meddled on Twitter after UK terror attacks, study says [2017-12-18] BBC

correct label ?

relevant ?

supports

refutes

other

yes

no

Overall Result

Feedback

✓ SUPPORTS

correct label ?

supports

refutes

other

Figure 3: Example of the fact checking interface. Claim: "Russia meddled with US elections" (top). Five maximum evidence sentences for each column: SUPPORTS/REFUTES/OTHER (middle). Example of final system decision and feedback buttons (bottom).

Trump: I hold Putin responsible over election meddling

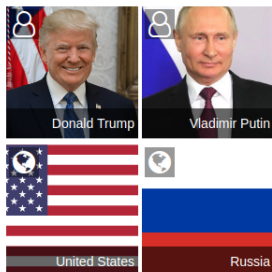
Summary:

- US President Donald Trump has said he holds Russian leader Vladimir Putin personally responsible for Russian interference in the 2016 US presidential election.
- In an interview broadcast with CBS News, he said Mr Putin must be held responsible for what happens in Russia.
- He said he had made clear to Mr Putin at their summit in Finland this week that Russian meddling had to stop.
- Mr Trump's assertions about Russia since he returned to the US are in sharp contrast to his comments at a joint news conference in Helsinki with Mr Putin.

ITPC Topics:

politics diplomacy international relations

Entities:



English

US President **Donald Trump** has said he holds Russian leader **Vladimir Putin** personally responsible for Russian interference in the 2016 **US** presidential election. In an interview broadcast with **CBS News**, he said **Mr Putin** must be held responsible for what happens in **Russia**. He said he had made clear to **Mr Putin** at their summit in **Finland** this week that Russian meddling had to stop. **Mr Trump's** assertions about **Russia** since he returned to the **US** are in sharp contrast to his comments at a joint news conference in **Helsinki** with **Mr Putin**.

Figure 4: Example of additional information for each evidence. Original document (right) with entities in bold.

time-frames and dates. For instance, claims using the present tense should refer to current events, while those mentioned in the past tense together with dates should refer to that specific time-period only. Also they suggested that presenting the evolution of results

over time would be very helpful to substantiate the claim. Handling time constraints in the retrieval process is a very interesting and challenging research direction. On the whole, the journalists reported that the system has a lot of potential to help their work.

This user testing was extremely useful both for BBC journalists to experiment with state-of-the-art technology and for us to receive feedback to improve our platform in the future.

6 CONCLUSIONS AND FUTURE WORK

This paper introduces a novel fact checking platform aimed to assist journalists in their investigative work-flow. Our platform can be used for search of supporting and refuting evidence regarding factual claims. We evaluated using on a journalistic corpus with testing by eleven journalists, which found it to yield relevant results in 59% of the retrieved evidence. The performed user study provided very fruitful feedback to direct future work in automated fact checking. Suggested improvements such as handling temporal remarks, pose an interesting issue that we found very relevant to advance research in the field of information retrieval for fact checking.

ACKNOWLEDGMENTS

The authors would like to thank James Thorne for insightful discussion and the BBC journalists for performing the user testing and for their valuable feedback. This work is supported by the EU H2020 SUMMA project (grant agreement № 688139), and by Lisbon Regional Operational Programme (Lisboa 2020), under the Portugal 2020 Partnership Agreement, through the European Regional Development Fund (ERDF), within project INSIGHT (№ 033869).

REFERENCES

- [1] Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. *One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling*. Technical Report. Google.
- [2] Sarah Cohen, Chengkai Li, Jun Yang, and Cong Yu. 2011. Computational Journalism: a call to arms to database researchers. *CIDR 2011, California, USA*. January (2011), 148–151.
- [3] Lucas Graves. 2018. *Understanding the Promise and Limits of Automated Fact-Checking*. Technical Report. Reuters Institute, University of Oxford.
- [4] Naeemul Hassan, Bill Adair, James T Hamilton, Chengkai Li, Mark Tremayne, Jun Yang, and Cong Yu. 2015. The Quest to Automate Fact-Checking. In *Proceedings of the 2015 Computation+Journalism Symposium*.
- [5] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, et al. 2017. ClaimBuster: the first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment* 10, 12 (2017), 1945–1948.
- [6] Afonso Mendes, David Nogueira, Samuel Broscheit, Filipe Aleixo, Pedro Balage, Rui Martins, Sebastião Miranda, and Mariana Almeida. 2017. SUMMA at TAC Knowledge Base Population Task 2017. In *Proceedings of the Text Analysis Conference KBP 2017*. NIST, USA.
- [7] Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. 2011. Rumor has it : Identifying Misinformation in Microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 1589–1599.
- [8] Stephen E. Robertson, Steve Walker, and M Beaulieu. 2000. Experimentation as a way of life: Okapi at TREC. *Information processing & management* 36, 1 (2000), 95–108.
- [9] James Thorne, , and Andreas Vlachos. 2018. Automated Fact Checking: Task Formulations, Methods and Future Directions. In *Proceedings of the 27th International Conference on Computational Linguistics*. ACL, 3346–3359.
- [10] James Thorne and Andreas Vlachos. 2017. An extensible framework for verification of numerical claims. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 37–40.
- [11] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *NAACL-HLT*.
- [12] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Workshop on Fact Extraction and VERification (FEVER). In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. ACL.
- [13] Dirk Weissenborn, Pasquale Minervini, Isabelle Augenstein, Johannes Welbl, Tim Rocktäschel, Matko Bosnjak, Jeff Mitchell, Thomas Demeester, Tim Dettmers, Pontus Stenetorp, et al. 2018. Jack the Reader. In *Proceedings of the 56th of ACL system demonstrations*. 25–30.
- [14] Gregor Wiedemann and Andreas Niekler. 2014. Document Retrieval for Large Scale Content Analysis using Contextualized Dictionaries. In *Terminology and Knowledge Engineering 2014*. Berlin, Germany, 10 p.
- [15] Takuma Yoneda, Jeff Mitchell, Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Four Factor Framework For Fact Finding (HexaF). In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. ACL, 97–102.
- [16] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and Resolution of Rumours in Social Media: A Survey. *ACM Comput. Surv.* 51, 2, Article 32 (Feb. 2018), 36 pages.