Employing Transformers and Humans for Textual-Claim Verification

Mohammed SAEED (EURECOM)

1 Context

From politicians to advertisers, from advocacy groups to enterprises — everyone who seeks to persuade others has an incentive to distort, exaggerate or obfuscate the facts. The topic of *fake news* has experienced a substantial resurgence of interest in our society. Indeed, we have started, nowadays, to catch a glimpse of the dire effects fake news could have on several angles. For examples, hundreds of people died because of coronavirus misinformation[1]. Thus, it is not a surprise that massive digital misinformation has been designated as a major technological and geopolitical risk by the 2013 report of the World Economic Forum, and an 'infodemic' by the World Health Organization.

Journalistic Fact-Checking. In the pre-automation period, checking for fake news was performed by trained journalists, who were employed to proofread and verify claims made in written or spoken language. This type of fact-checking verifies the solidity of the stated claims *before publication*; acting as a core part of journalistic work [1]. Such fact-checking procedure is a vital component in the news reporting process. In this thesis, we tackle the second type of fact-checking, which happens not before something is published but rather after a claim becomes public. This form of *a posteriori* fact-checking is often performed by specialists in active NGOs such as PolitiFact, FullFact, and Les Décodeurs.

Manual Fact-Checking is not Enough. With the struggles of manual fact-checkers to keep up with the ever-increasing surge of fake news, we are facing a widening partition between the growing amount of data on one hand, and the shrinking body of trained journalists on the other. Alas, relying solely on manual fact-checkers for the fight is not enough. One possible direction to narrow this partition is through the efficient use of computing algorithms and resources. Computing methods would, ideally, attempt to imitate a fact-checker as much as possible. This would not only ease some manual aspects of the fact-checking process, but could help, to a certain extent, with the enormous volume of fake news produced every day. For such reason, it is not a surprise that *Computational Fact-Checking* has sparked interest across academic labs and industries, and has emerged as an aspiring research field among academics [16].

2 State of the Art

Despite the numerous fact-checking systems currently present [6], we are yet far from fully automating the fact-checking pipeline due to several ordeals. We identify four main obstacles that hinder the scope, efficiency, and applicability of fact-checking systems:

1. While there exists many fact-checking systems that attempt to integrate tabular information [5], the majority of such systems has been applied on well-curated datasets, disregarding the challenge posed by domain-specific claims that require costly annotations of domain experts. How does one go about verifying numerous claims without relying excessively on expert fact-checkers for manual labeling, while coping with the imperfections of machine learning models?

[1] `https://www.bbc.com/news/world-53755067`

2. Despite that pre-trained language models (PLMs) can be utilized for claim verification, PLMs still lack in terms of logical reasoning, as they cannot perform deductive reasoning using logical rules. They are also severely inconsistent [4] in terms of negation and symmetry. However, even for some line of work that augments PLMs with logical rules [2, 15], the kind of supported rules is limited. This is not sufficient for real-world first-order logic. How can a PLM "reason" with facts, rules, and hypothesis in natural language with probabilistic outputs ?

3. As PLMs store vast factual and common-sense information [10], one cannot always expect to arrive at the desired output when querying them. Being able to direct the PLM when querying to match a desired criterion is required for better retrieval.While approaches that search for optimal prompts through means of learning algorithms [17], or rely on external resources to enrich the input context improve performance [8], they do not take the desired output type into account on one hand, and require some form of training on the other. How can we enforce the desired type, knowing that PLMs encode latent concepts such as city and year [3]?

4. As automating the entire fact-checking pipeline is currently unfeasible, including humans in the pipeline is inevitable. Considering that expert fact-checkers are scarce, a certain methodology remedies this by relying on a larger number of unprofessional humans to perform fact-checking; or what is known as the 'wisdom of the crowd'. While a series of works analyze such crowdsourcing approach to fact-checking [7, 11], it has been only done for controlled environments, and there is no clear vision of how such an approach would behave in real uncontrolled settings. Is such an approach effective? And how does it compare to experts and computational methods ?

3 Scientific Results Obtained

To tackle the above challenges, this thesis makes progress in the direction of automated fact-checking systems by (i) extending neural networks for better fact-retrieval and reasoning, (ii) investigating if crowdsourcing is an efficient approach for fact-checking, and (iii) combining both neural networks and humans for claim verification. We put forward the following main results:

1. We present a fact-checking system, SCRUTINIZER, that verifies statistical claims in natural language by exploiting the synergy between humans and algorithms [9]. The system is based on two components: a claim translation component and a question planning component. The former is responsible for automated translation from text to query elements (such as datasets and attributes) needed to fact-check the claim, done through the use of trained classifiers. However, as data is not always available, the system needs to solicit feedback from experts for labeling the data. A question planning component interacts with human domain experts to optimize verification tasks for maximal benefit. This is done by modeling claim selection as an integer linear programming problem that allows to select the most optimal claims to be labeled. We apply SCRUTINIZER to two use-cases: one domain-specific related to energy, and the other to COVID-19 coronavirus, with an online demo[2] that has been used to verify more than 25k claims.

2. We explore how to emulate reasoning with PLMs using first-order soft logic rules [12]. We extend previous work by (i) harnessing PLMs with logical rules containing binary predicates, as opposed to unary predicates, and (ii) incorporating soft rules during training. For (i), we develop a data generation algorithm that encompasses logical aspects such as symmetry, which is crucial when dealing with binary predicates, and negation. For (ii), we propose to modify the objective function to integrate the weights of rules. We test our system RULEBERT on single rules of various confidences, multiple rules including rules with conflicting conclusions, and chained rules. We finally test RULEBERT on several external datasets showing improvements in deductive reasoning and in logical notions such as negation and symmetry. We also release this novel dataset, comprising 3.2M examples derived from 161 logical rules.

[2] `https://coronacheck.eurecom.fr`

3. We propose the idea of Type Embeddings [13], additional input embeddings that guide the model into the direction of a certain type, for example by steering the outputs to years instead of locations. We present a method to compute TEs based on removing the first singular vector of a token embedding matrix. To test the effectiveness of this embedding, we offer a suite of tests. Finally, we show that PLMs equipped with TEs provides an increase in performance on fact-retrieval datasets and text generation.

4. We analyze the Birdwatch program [14], the first large-scale community-based fact-checking initiative by Twitter. We inspect how the crowd attempts to fact-check tweets in practice, while comparing with human experts and automated fact-checking tools. We focus on three aspects related to claim selection, evidence retrieval, and claim verification. Our insights show that the crowd could be effective in verifying truthfulness of claims in a faster pace compared to expert fact-checkers, however more care should be taken into profiling these workers, as in the uncontrolled environment we study (Twitter), since malicious users could manipulate the verification process in their favor. We also release the matched dataset of 11.9k tweets with Birdwatch checks and identify 2.2k tweets verified both by Birdwatch users and experts.

4 Publications

Within the thesis, 11 papers have been published in top-tier conferences and journals. Two demonstration papers got an award, 5 were published in A* venues, and 3 in A venues.

1. **Scrutinizer: A Mixed-Initiative Approach to Large-Scale, Data-Driven Claim Verification.** *Georgios Karagiannis*, Mohammed Saeed*, Paolo Papotti, Immanuel Trummer.* [Proceedings of the Very Large Data Bases Endowment (PVLDB) 2020 Long Paper]

2. **Scrutinizer: Fact checking statistical claims.** *Georgios Karagiannis*, Mohammed Saeed*, Paolo Papotti, Immanuel Trummer.* [VLDB 2020 Demo Paper + **Best Demo Paper Award** at BDA 2020]

3. **Fact-Checking Statistical Claims with Tables.** *Mohammed Saeed, Paolo Papotti.* [IEEE Data Engineering Bulletin Volume 44 Issue 3 2021]

4. **RuleBERT: Teaching Soft Rules to Pre-Trained Language Models.** *Mohammed Saeed, Naser Ahmadi, Preslav Nakov, Paolo Papotti.* [Empirical Methods in Natural Language Processing (EMNLP) 2021 Long Paper (Main)]

5. **You Are My Type! Type Embeddings for Pre-trained Language Models.** *Mohammed Saeed, Paolo Papotti.* [EMNLP 2022 Long Paper (Findings)]

6. **Crowdsourced Fact-Checking at Twitter: How Does the Crowd Compare With Experts?** *Mohammed Saeed, Maelle Nicolas, Nicolas Traub, Gianluca Demartini, Paolo Papotti.* [Conference on Information and Knowledge Management (CIKM) 2022 Long Paper]

7. **Transformers for Tabular Data Representation: A Survey of Models and Applications.** *Gilbert Badaro, Mohammed Saeed, Paolo Papotti.* [Transactions of the Association for Computational Linguistics (TACL) Journal Volume 10 2022]

8. **Automatic Verification of Data Summaries.** *Rayhane Rezgui, Mohammed Saeed, Paolo Papotti.* [International Natural Language Generation Conference (INLG) 2021 Systems Paper]

9. **Neural Re-rankers for Evidence Retrieval in the FEVEROUS Task.** *Mohammed Saeed, Giulio Alfarano, Khai Nguyen, Duc Pham, Raphaël Troncy, Paolo Papotti.* [4th Fact Extraction and VERification (FEVER) Workshop Systems Paper]

* denotes equal contribution.

10. **Pythia: Unsupervised Generation of Ambiguous Textual Claims from Relational Data.** *Enzo Veltri, Donatello Santoro, Gilbert Badaro, Mohammed Saeed, Paolo Papotti.* [IEEE International Conference on Data Engineering (ICDE) 2023 Full Paper]

11. **Pythia: Unsupervised Generation of Ambiguous Textual Claims from Relational Data.** *Enzo Veltri, Donatello Santoro, Gilbert Badaro, Mohammed Saeed, Paolo Papotti.* [**Best Demo Paper Award** at Special Interest Group on Management of Data (SIGMOD) 2022]

References

[1] S. Cazalens, P. Lamarre, J. Leblay, I. Manolescu, and X. Tannier. A content management perspective on fact-checking. In *Companion Proceedings of the TheWebConf*, WWW '18, page 565–574, 2018.

[2] P. Clark, O. Tafjord, and K. Richardson. Transformers as soft reasoners over language. In C. Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3882–3890, 2020.

[3] F. Dalvi, A. R. Khan, F. Alam, N. Durrani, J. Xu, and H. Sajjad. Discovering latent concepts learned in BERT. In *International Conference on Learning Representations*, 2022.

[4] Y. Elazar, N. Kassner, S. Ravfogel, A. Ravichander, E. Hovy, H. Schütze, and Y. Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021.

[5] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko. Revisiting deep learning models for tabular data. *ArXiv preprint*, abs/2106.11959, 2021.

[6] Z. Guo, M. Schlichtkrull, and A. Vlachos. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, 2022.

[7] N. Hassan, M. Yousuf, M. Mahfuzul Haque, J. A. Suarez Rivas, and M. Khadimul Islam. Examining the roles of automation, crowds and professionals towards sustainable fact-checking. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 1001–1006, New York, NY, USA, 2019. Association for Computing Machinery.

[8] Z. Jiang, F. F. Xu, J. Araki, and G. Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020.

[9] G. Karagiannis*, M. Saeed*, P. Papotti, and I. Trummer. Scrutinizer: A mixed-initiative approach to large-scale, data-driven claim verification. *Proc. VLDB Endow.*, 13(12):2508–2521, July 2020.

[10] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, and A. Miller. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[11] K. Roitero, M. Soprano, B. Portelli, M. D. Luise, D. Spina, V. D. Mea, G. Serra, S. Mizzaro, and G. Demartini. Can the crowd judge truthfulness? a longitudinal study on recent misinformation about covid-19. *Personal and Ubiquitous Computing*, pages 1 – 31, 2021.

[12] M. Saeed, N. Ahmadi, P. Nakov, and P. Papotti. RuleBERT: Teaching soft rules to pre-trained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1460–1476, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics.

[13] M. Saeed and P. Papotti. You are my type! type embeddings for pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Online and Abu Dhabi, UAE, Dec. 2022. ACL.

[14] M. Saeed, N. Traub, M. Nicola, G. Demartini, and P. Papotti. Crowdsourced fact-checking at twitter: How does the crowd compare with experts? In *31st ACM International Conference on Information and Knowledge Management*, Online and Atlanta, Georgia, USA, Oct. 2022.

[15] S. Saha, S. Ghosh, S. Srivastava, and M. Bansal. PRover: Proof generation for interpretable reasoning over rules. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 122–136, Online, Nov. 2020. Association for Computational Linguistics.

[16] A. Vlachos and S. Riedel. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA, June 2014. Association for Computational Linguistics.

[17] Z. Zhong, D. Friedman, and D. Chen. Factual probing is [MASK]: Learning vs. learning to recall. In *Proceedings ACL*, pages 5017–5033, Online, 2021. Association for Computational Linguistics.