Habilitation Abstract
Towards Explainable Fact Checking
Augenstein, Isabelle

Published in:
KI - Künstliche Intelligenz

DOI:
10.1007/s13218-022-00774-6

Publication date:
2022

Document version
Early version, also known as pre-print

Document license:
Unspecified

Citation for published version (APA):
Habilitation Abstract: Towards Explainable Fact Checking

Isabelle Augenstein

Abstract With the substantial rise in the amount of mis- and disinformation online, fact checking has become an important task to automate. This article is a summary of a habilitation (doctor scientiarum) thesis submitted to the University of Copenhagen, which was successfully defended in December 2021 [4]. The dissertation addresses several fundamental research gaps within automatic fact checking. The contributions are organised along three verticals: 1) the fact-checking sub-task they address; 2) methods which only require small amounts of manually labelled data; 3) methods for explainable fact checking, addressing the problem of opacity in the decision-making of black-box fact checking models.

Keywords Automatic Fact Checking · Explainable AI · Natural Language Understanding · Low-Resource Learning · Multi-Task Learning

1 Introduction

False information online is a substantial societal problem, which can have real-world consequences. These includes disinformation (intentionally false information) spread to e.g. influence political campaigns, and misinformation (unintentionally false information), e.g. about public health. As manual efforts to fact-check false claims cannot scale to the amount of newly emerging false information, automated methods are needed. These methods should be able to learn from limited amounts of annotated domain data, as current fact checking datasets are small and restricted to few domains. Moreover, to assist human decision making, it is important that these methods are transparent so that users can inspect them (model understanding), and that they can explain their reasoning for individual predictions (decision understanding).

2 Contributions

To address the above-mentioned challenges, this thesis makes several scientific contributions, which can be conceptualised along three axes: 1) the fact checking (FC) sub-task they address; 2) the method they present for dealing with limited and heterogeneous limited data (LLD); 3) the explainability method they propose. Table 1 indicates where along these three axes each of the ten papers that make up this thesis are located.

Fact Checking Subtask Figure 1 shows an example of a full fact-checking pipeline, starting with the detection of check-worthy claims, and ending with the verification of a claim’s veracity.
be to determine if this statement constitutes a claim or an opinion and, if it constitutes a claim, whether or not that claim is worth fact checking (claim check-worthiness detection). The latter is influenced by claim importance, which is subjective. Typically, only sentences unlikely to be believed without verification are marked as check-worthy. Furthermore, domain interests skew what is deemed check-worthy, which might e.g. only be certain political claims, or only celebrity gossip, depending on the application at hand. Lastly, claims can be very different in nature, e.g. numerical claims, claims about entity and event properties, position statements or quote verification. All of this already makes the first step in the fact checking pipeline, the detection of check-worthy claims, a surprisingly non-trivial task. A more in-depth investigation of these challenges can be found in [9].

Following this, if the statement indeed is a check-worthy claim, evidence documents which can be used to confirm or refute the claim are retrieved from the Web and ranked by their relevance to the claim (evidence retrieval and ranking). The source of documents for retrieval can be restricted to certain domains, e.g. Wikipedia, or the whole Web can be used as a source, in which case a search engine is often used in this part of the pipeline, as done for the MultiFC dataset [5].

The next step is to determine if the evidence documents retrieved agree with, disagree with or are neutral towards the claim (stance detection). There are many different labelling schemes for this task, ranging from simply ‘positive’ vs. ‘negative’ to a very fine-grained task with many different labels. This challenge is explored in more detail in [7]. Intuitively, this step is easier the more directly an evidence document discusses a claim – the less textual overlap there is between the two texts, the harder it to determine the stance automatically – which is discussed in [9].

Lastly, the overall veracity of the claim in question can be determined. As in stance detection, there exist many different labelling schemes for this step. In addition to fine-grained judgements about where on the scale from ‘completely true’ to ‘completely false’ a claim lies, labels such as ‘not enough information’ or ‘spins the truth’ are used. Solutions to this are discussed in great detail in [5].

Learning with Limited and Heterogeneous Labelled Data
A core methodological challenge addressed throughout this thesis is that most fact checking datasets are small in nature and apply different labelling schemes. Learning stable automatic fact checking models from those datasets thus presents a significant challenge. This thesis proposes several streams of methods to tackle this problem.

Automatic fact checking is, at its core, a classification task: each claim is categorised as belonging to one of a set of classes denoting the claim’s veracity. Hence, general text classification methods ought to be suitable for automatic fact checking as well. While they work well for most purposes, this formulation has several shortcomings, which can be addressed by studying different ways of modelling the output space, i.e. the way in which labels for the different fact checking sub-tasks are modelled and predictions are structured [5,11,73].

Another challenge concerns the generally low number of training instances of fact checking datasets. Two streams of research are explored in this thesis to tackle this problem – multi-task and transfer learning – both of which build on the general idea of obtaining more training data from other sources. Multi-task learning is the idea of, in addition to the target task, obtaining training data for so-called auxiliary tasks. The latter are tasks which are often related to the target task, be it in form (i.e. if the target task is a text classification task, those would also be text classification tasks); in domain (e.g. the target task and auxiliary task data could all be from the legal domain); or in the nature of the task (e.g. only taking into consideration different variants of sentiment analysis). The idea is then to train a model on all such tasks at once, but to only utilise the predictions of the target task. Multi-task learning for fact checking is investigated in [7,5,23].

Finally, this thesis examines transfer learning as a way of increasing the amount of training data for a task [9,10,8]. Transfer learning can be seen as as special form of multi-task learning where a model is trained on several tasks, but instead of it being trained on several tasks at once, it is trained on several tasks sequentially. The last one of these tasks is typically the target task, unless there is no training data available for the target task at all, in which case one typically speaks of unsupervised or zero-shot transfer learning. The other tasks are typically referred to as the source tasks.

Explainable Natural Language Understanding
The last vertical of relevance to fact checking is how to make fact checking models more transparent, such that end users can understand what a model as a whole has learned, as well as why a model produces a certain prediction for a certain instance.

The inner workings of deep neural networks are, as already mentioned, relatively complex; especially since modern models have too many parameters to inspect them individually. One solution to this is to generate natural language explanations, based on the assumption
that the easiest explanations to understand for users are those written in natural language. The overall aim is to produce free text (typically a sentence or a paragraph) that succinctly explains how the model has arrived at a certain prediction. Technically, this is achieved by training a model for both the main task and to generate a textual explanation. In [2], we approach this using multi-task learning.

Another way of interpreting what a model has learned is to try to reveal systemic vulnerabilities of the model. Sometimes, usually because a model is trained on biased and/or small amounts of training data, it learns spurious correlations resulting in features that are red herrings – which a model has only seen a handful of times at training time, which were always associated with only one label, but which, in reality, are not indicative of the label. An example from the fact checking domain could be if a model is only exposed to false claims mentioning certain people, then it will very likely learn to always predict the veracity of ‘false’ whenever this name occurs in a claim. The goal of generating adversarial examples is to identify such features, sometimes also called ‘universal adversarial triggers’, and use them to automatically generate instances which a fact checking model would predict an incorrect label for. This not only tells a user what a model would likely struggle with, but the automatically generated adversarial instances can in turn be used to improve models. In [3], we explore how to generate adversarial claims for fact checking.

Lastly, another explainability technique explored in this thesis is post-hoc explainability. Unlike the approaches presented for generating natural language explanations and discovering universal adversarial triggers, post-hoc explainability methods are methods that can be applied after a model for a certain application task has already been trained. The general idea is to find regions of the input which best explain the predicted label for the corresponding instance. These regions of the input are typically called ‘rationales’, and are portions of the input text – words, phrases or sentences – which are salient for the predicted label given the trained model. [4] addresses the highly challenging task of automatically evaluating such post-hoc explainability methods.

3 Summary of Findings

This thesis presents the first thorough comparison of multiple claim check-worthiness detection tasks – rumour detection on Twitter, check-worthiness ranking in political debates and speeches and ‘citation needed’ detection on Wikipedia [9]. Experiments show that a unified approach to check-worthiness detection, using a novel extension of positive unlabelled learning, leads to substantial improvements over the state of the art. The thesis further proposes methods for improving out-of-domain application, and finds that unsupervised domain adaptation using mixture-of-expert methods leads to significant gains for rumour detection [10].

For stance detection, the thesis presents several approaches to better model the relationships between tasks using multi-task and transfer learning, leading to improvements in results in low-resource settings [6,7,8]. Moreover, methods are proposed to model the output space by learning label embeddings, which capture the semantic relationships between heterogeneous labels drawn from different datasets [7]. Another direction shown to improve results is to model the conversational structure of texts – specifically for tweets, which can be organised in threads [11]. Lastly, the thesis performs an in-depth study of explainability methods based on rationale selection, which highlight how salient parts of the input are for the predicted label [1]. It then shows how these explanations can be evaluated automatically without the need for human annotations, proposing a suite of diagnostic properties to do so.

Lastly, the overall veracity of a claim can be determined. As in stance detection, there exist many different labelling schemes for this step. Successful methods to approach this again include multi-task learning and label space modelling [5]. Moreover, prediction models can overfit to spurious correlations observed in data. The thesis thus studies ways to reveal systemic vulnerabilities of the model, and generate new, fluent claims using adversarial learning. Training on these additional claims, conversely, improves the robustness of fact checking models [2]. Lastly, the thesis presents the first study on generating natural language explanations for real-world claims [3]. This is achieved by framing the task as a summarisation problem, where, provided with elaborate fact checking reports, a model has to generate veracity explanations close to the human justifications. Experiments show that this is a highly challenging task, though that optimising for veracity prediction alongside explanation generation leads to improvements in terms of coverage and overall quality.

4 Conclusions and Future Work

This thesis presents core contributions within automatic fact checking, presenting methods that rely less on manually labelled training data, as well as first approaches to generate explanations for fact checking. Overall, explainability is a very challenging task. To make substantial progress, large, human-annotated datasets for different explanation types are required. Research on
methodology development could then focus on exploiting relationships between different explainability methods in a joint framework.

Acknowledgements The research described in the thesis was partially funded by several grants from the European Commission: the PHEME FP7 project (grant No. 611233), the Marie Skłodowska-Curie project under grant agreement No 801199, ERC Starting Grant Number 313695 and QUARTZ (721321, EU H2020 MSCA-ITN).

References


