CROWD TRUTH CONQUERS ALL: LEVERAGING HUMAN INTELLIGENCE TO FIGHT MISINFORMATION

Michael Soprano (michael.soprano@uniud.it)

My research interests consist in Information Retrieval and Crowdsourcing. In particular, the use of crowdsourcing based approaches to address the increasing amount of misinformation that is spreading online, a phenomenon for which the term ''infodemic'' has been created.[1] My long term goal is to build a human-in-the-loop system to cope with misinformation by measuring information truthfulness in real-time using crowd-powered data, human intelligence, and machine learning techniques.

Overview

The rise of (online) misinformation is a problem that harms society, and the information that we consume every day influences our decision-making process. Thus, the ability to distinguish between true and false claims becomes more and more important over time. A way to assess information truthfulness consists of relying on human experts who perform a fact-checking activity. However, the ever-increasing amount of misinformation is making this task more challenging, if not even impractical, over time. Furthermore, the limited amount of trained experts makes the whole process costly and non-scalable. A possible alternative is to rely only on machine learning systems to detect and assess the truthfulness level of information, but they require a great amount of data for the training phase and their reliability, effectiveness, and explainability are often not adequate. Therefore, to address the issues of such approaches one may think of relying on the large amount of non-expert people that consume information and ask them to perform the fact-checking activity. It is possible to use *crowdsourcing* [5] based approaches to collect truthfulness labels provided by non-expert people on statements [11, 12, 13, 14]. Such a decision leads to several opportunities but also difficulties; while it is possible to collect a large amount of data in a considerably short time, there is not any guarantee on the quality of the data collected. The solution that I intend to study is the integration of these three approaches, to cope with the aforementioned difficulties while preserving the benefits of the diverse methodologies.

Past Research

In past work, I investigated whether crowdsourcing can be reliably used to assess information truthfulness and to create large scale collections of truthfulness labels to train machine learning systems and models. In my first work [11], we ran a large crowdsourcing experiment where a crowd of workers were asked to fact-check statements given by politicians and search for evidence using a custom controlled search engine. We sampled a set of statements from two popular datasets, namely Politifact [15] and ABC [10]. Each dataset includes the truthfulness labels provided by expert fact-checkers for each statement. We recruited a crowd of workers and we asked each worker to perform the fact-checking task. We collected thousands of truthfulness labels on a six-level assessment scale and we compared them with the expert assessments. We found that crowdsourced truthfulness labels are indeed useful to single out true from false statements and that the personal background of each worker has a role in how he/she assess misinformation. We also found that workers put effort into finding reliable resource to justify their assessments. In my second work [13], we focused on a set of statements specifically targeting the COVID-19 pandemic. The goal was to understand how people address misinformation when it is related to a sensitive topic like health and when it is considerably recent. We manually sampled a set of COVID-19 related statements given by US politicians expressed on a six-level assessment scale. Similarly to my first work [11], we had the truthfulness label provided by expert fact-checkers for each statement of the dataset. In addition to searching a resource to justify their assessments, crowd workers were required to write a textual justification to motivate their choice. We wanted to investigate how they chose the justification resource and to understand if such justifications could be used to derive additional information. We found that workers were able to identify and objectively categorize misinformation related to the COVID-19 pandemic, that the provided labels show a high level of agreement with the expert labels when aggregated and that both expert and crowdsourced labels can be transformed to improve their quality. Furthermore, we found relations between justifications and assessments quality. We also collected answers from each worker to grasp their personal background and cognitive abilities, for both experiments.

[1] `https://www.who.int/dg/speeches/detail/munich-security-conference`

Current Research

I am currently working on two main lines of research. I submitted two articles [14, 12] to top tier journals (Scimago Q1 ranking) which are now under the second round of revision. My previous work [11, 13] showed that crowd workers can effectively and reliably identify and address information truthfulness. Crowd workers used a coarse-grained assessment scale to assess the overall truthfulness of a set of statements. However, a statement could be truthful but also imprecise, biased, difficult to read, and so on. Such differences may be nuanced and hard to spot. Therefore, a unidimensional notion of truthfulness may not suffice to take these differences into account. We proposed a multidimensional notion of truthfulness grounded on psychological studies and We asked crowd workers to assess statements along different dimensions. We sampled a set of statements from a dataset containing statements by US politicians, as I did in my previous work [11, 13]. We found that the assessments provided over the dimensions are sound and reliable and that each dimension measures a different facet of truthfulness. Moreover, We found that the whole set of dimensions is useful to understand the reasons behind the crowd workers assessments. We also found that the whole set of dimensions is useful to understand the reasons behind the crowd workers assessments and that worker quality and behavior could be leveraged to predict expert fact-checkers assessments by using machine learning. One may think of how the behavior and quality of a crowd of workers change when assessing information truthfulness at different time spans. We performed a longitudinal study where the experimental setup is the same as my COVID-19 pandemic related work [13]. We re-launched the same task multiple times with both novice and experienced workers. We found that time has a major effect on the quality of the assessments for both types of workers and that experienced workers spend more time assessing statements than novice workers.

Future Research

There are several lines of research that I plan to pursue to further understand how people address information truthfulness assessment.

Cognitive Biases More than 180 cognitive biases exists [3]. Such biases can affect human behavior and the reasoning process. I plan to define, characterize, and survey which cognitive biases affect a worker when assessing information truthfulness. This can be done by framing a situation where such biases can manifest to study their impact on worker quality and behavior. There are specific settings that can be employed when assessing information truthfulness that act as a countermeasure and prevent the bias from occurring or, at least, limit its effect to make the information truthfulness assessment activity as objective as possible. I aim to provide researchers with a framework of strategies and best practices to have an unbiased environment for the assessors.

Time-bound assessment Limiting the time available to a worker to express an assessment affect its quality but it can also increase the quality of the collected data [7]. It is an interesting facet to study if the same effect occurs in the context of information truthfulness assessment. I plan to launch a crowdsourcing experiment concerning information truthfulness where the time available for crowd workers to provide each label is constrained. Since they are paid with respect to an estimate of the time required to complete the task, a time-bound assessment, if effective, would allow optimizing the cost of a crowdsourcing task by finding the sweet spot where the worker is not under pressure but also not allowed to multitask.

Magnitude estimation In my previous work [11], I used three different pointwise rating mechanisms to collect assessments concerning information truthfulness but their final quality was similar. There exist other mechanisms which are worth studying such as magnitude estimation [9]. Such a mechanism is a psychophysical scaling technique for the measurement of sensation, where observers assign numbers to stimuli in response to their perceived intensity. It has many applications and it has been used also in Information Retrieval [8]. I plan to launch a crowdsourcing experiment concerning information truthfulness where crowd workers are required to assess information truthfulness using magnitude estimation to understand if such a technique provides improvements in worker quality and behavior.

Pairwise assessment People are rarely exposed to one information element from a single source of information at a time [1]. Also, people able to evaluate diverse information in a short amount of time. My previous experiments [11, 13] are performed in a pointwise fashion, where each worker address one information element at a time, sequentially. One may think about allowing workers to address multiple elements at the same time since pairwise comparison is a promising alternative [4]. Therefore, I plan to expose the crowd workers to (at least)

two statements in a pairwise fashion to study the impact on worker quality and behavior. I also plan to make workers produce a ranked list of statements. Machine learning systems usually use pairwise or listwise approaches to learn how to rank a set of items [6]. This means that there is a discrepancy between how crowdsourcing-based approaches and machine learning techniques collect and aggregate labels. I aim to leverage crowd-powered data using machine learning techniques to learn how to rank information elements when their truthfulness is considered.

Conversational crowdsourcing The usage of many conversational agents such as smart speakers, assistants, and chatbots is increasing over time [2]. The COVID-19 pandemic is pushing this trend even further. For example, people who are working from home are more likely to ask for updates on news and information [**?**]. Allowing people to address information truthfulness using such conversational agents could provide several benefits. The possibility to perform such a task by using a smart speaker in a voice-only fashion could keep people more engaged and interested. This is reasonable also when considering chatbots since people would perform the task using an interface to which they are already used. My goal is to implement and use such interfaces to launch a crowdsourcing experiment concerning information truthfulness to understand if they provide improvements in worker quality, behavior, satisfaction, and engagement.

Conclusions I am distributing and making available to the research community the software that I have designed and implemented to conduct each crowdsourcing experiment along with adequate documentation and the data collected. I believe that the aforementioned future research lines and my previous and current research will help in understanding how to leverage crowd-powered human intelligence to build a human-in-the-loop system to fight the spread of misinformation.

References

[1] J. Allen, B. Howland, M. Mobius, D. Rothschild, and D. J. Watts. Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, 2020.

[2] K. Bitterly. 1 in 4 Americans Own a Smart Speaker. What Does That Mean for News? *New York Times*, 2019.

[3] J. Caverni, J. Fabre, and M. Gonzalez. *Cognitive biases*. Elsevier, 1990.

[4] A. Checco and G. Demartini. Pairwise, Magnitude, or Stars: What's the Best Way for Crowds to Rate?, 2016.

[5] J. Howe. The Rise of Crowdsourcing, Wired 14, 2006.

[6] H. Li. *Learning to Rank for Information Retrieval and Natural Language Processing*. Synthesis Lectures on Human Language Technologies. 2014.

[7] E. Maddalena, M. Basaldella, D. De Nart, D. Degl'Innocenti, S. Mizzaro, and G. Demartini. Crowdsourcing Relevance Assessments: The Unexpected Benefits of Limiting the Time to Judge. In *4th AAAI HCOMP*, 2016.

[8] E. Maddalena, S. Mizzaro, F. Scholer, and A. Turpin. On Crowdsourcing Relevance Magnitudes for Information Retrieval Evaluation. 2017.

[9] H. R. Moskovitz. Magnitude Estimation: Notes On What, How, When, And Why To Use It. *Journal of Food Quality*, 1977.

[10] RMIT University. RMIT ABC Fact Check. `https://apo.org.au/collection/302996/rmit-abc-fact-check`.

[11] K. Roitero, **M. Soprano**, S. Fan, D. Spina, S. Mizzaro, and G. Demartini. Can The Crowd Identify Misinformation Objectively? The Effects of Judgment Scale and Assessor's Background. In *43st ACM SIGIR*, 2020.

[12] K. Roitero, **M. Soprano**, B. Portelli, M. De Luise, D. Spina, V. Della Mea, G. Serra, S. Mizzaro, and G. Demartini. Can the Crowd Judge Recent Misinformation on the COVID-19 Infodemic Objectively? A Longitudinal Study. *Information Systems*. Under revision.

[13] K. Roitero, **M. Soprano**, B. Portelli, D. Spina, V. Della Mea, G. Serra, S. Mizzaro, and G. Demartini. The COVID-19 Infodemic: Can the Crowd Judge Recent Misinformation Objectively? In *29th ACM CIKM*, 2020.

[14] **M. Soprano**, K. Roitero, D. La Barbera, D. Ceolin, D. Spina, S. Mizzaro, and G. Demartini. The Many Dimensions of Truthfulness: Crowdsourcing Misinformation Assessments on a Multidimensional Scale. *Information Processing & Management*. Under revision.

[15] W. Y. Wang. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *55th Annual Meeting of ACL*, 2017.