The COVID-19 Infodemic: Can the Crowd Judge Recent Misinformation Objectively

Kevin Roitero, **Michael Soprano**, Beatrice Portelli,
Damiano Spina, Vincenzo Della Mea, Giuseppe Serra,
Stefano Mizzaro, and Gianluca Demartini

29TH ACM INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT (CIKM 2020)

Context

[Fabrice Coffrini Getty Images]

- Misinformation is an ever increasing problem that has a negative impact on the society at large
- All of us have experienced **misinformation during the COVID-19 health emergency**
- WHO director Dr. Ghebreyesus chooses to target explicitly misinformation related problems
- **INFODEMIC!**

Aims

- **Investigate if crowd workers are able to identify and correctly classify (mis)information**
- Extensive crowdsourcing experiment
  - Each crowd worker is asked to fact check 60 COVID-19 related statements
  - **Data available** at https://github.com/KevinRoitero/crowdsourcingTruthfulness
- In a previous work we addressed political statements [Roitero et al. 2020]

[Roitero et al. 2020] Kevin Roitero, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2020. Can The Crowd Identify Misinformation Objectively? The Effects of Judgment Scale and Assessor's Background. In Proceedings of SIGIR '20. DOI:https://doi.org/10.1145/3397271.3401112

Research Questions

- **RQ1**: Suitability of crowd workers to detect and objectively categorize COVID-19 related misinformation
- **RQ2**: Aggregation of crowdsourced/expert judgments to improve the ability of workers
- **RQ3**: Effect of workers' political bias, background and CRT tests performances
- **RQ4**: Signals provided by the workers while performing the task
- **RQ5**: Sources of information that crowd workers use to identify online misinformation

Experimental Setup

- We sampled 10 statements for each of the six truthfulness level of **Politifact [Yang 2017]**
  - COVID-19 section[1]
- Each of 100 distinct crowd workers judges 6 + 2 (gold questions) = **8 statements**
  - 8*100=800 judgments collected
- Amazon Mechanical Turk crowdsourcing task
- 1 questionnaire with 7 questions to collect **worker background**
- 3 questions to measure **cognitive abilities (CRT tests)**
- The worker must provide:
  - a **truthfulness level** on a given six-level scale
  - **URL** which serves as **justification/source of evidence** for the judgments
  - a **textual motivation** for her/his response
- **Quality checks** to ensure quality of collected data

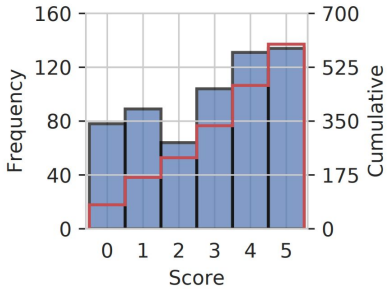
[Yang 2017] William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In ACL'17.

[1]https://www.politifact.com/coronavirus/

Descriptive Statistics

1153 US resident crowd workers participated to our study. 100 of them submitted successfully their work, 953 abandoned or failed:

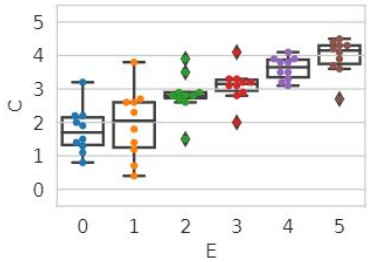
- **Workers' background**: the sample of workers is balanced along each questionnaire answer
- **CRT scores**: the 69% of workers answered correctly to at least 1/3 test question
- **Abandonment**: higher abandonment ratio w.r.t our previous work [Roitero et al. 2020]
- **Crowdsourced scores**: judgments are overall of a decent quality; evenly spaced steps in cumulative distribution

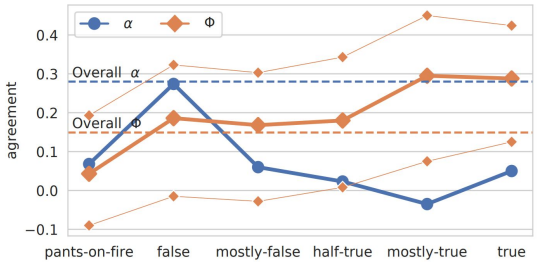


[Roitero et al. 2020] Kevin Roitero, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2020. Can The Crowd Identify Misinformation Objectively? The Effects of Judgment Scale and Assessor's Background. In Proceedings of SIGIR '20. DOI:https://doi.org/10.1145/3397271.3401112

RQ1: Crowd Accuracy

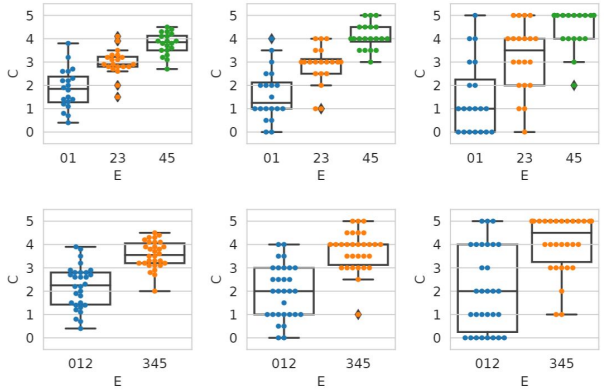
- **External agreement**: individual judgments in agreement with expert labels
  - Workers are capable of recognizing and classifying misinformation statements related to COVID-19 pandemic
- **Internal agreement**: agreement among the workers
  - Higher agreement levels for `mostly-true` and `true` categories
  - Workers are most effective in identify and categorizing statements with a higher truthfulness level





RQ2: Merging Assessment Levels

- We group adjacent categories to check if looking at data on a more coarse-grained ground truth improves the results
  - six Politifact categories in either three (01, 23, 45) or two (012, 345) new categories
  - if we merge categories together **the crowd can effectively detect and classify misinformation statements related to COVID-19 pandemic**



RQ3: Worker Background and Bias

- **Questionnaire**: CEM^{ORD} effectiveness metric for ordinal classification [Amigó et al. 2020]
  - Very conservative workers provide lower quality label (statistically significant)
  - Other answers to questionnaire do not provide additional insights
- **CRT Test**: there is some variation in both accuracy and CEM^{ORD}
  - Never statistically significant
  - The number of correct answers to the CRT test is not correlated with worker quality

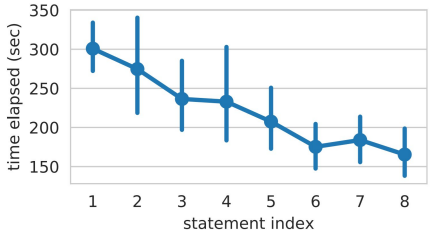
		Correctly classified statements								Acc	CEM^{ORD}
		0	1	2	3	4	5	6	Sum		Mean
CRT	0	5	11	9	4	0	1	1	31	.14	.48
correct	1	5	10	12	6	1	0	0	34	.22	.53
answers	2	1	6	1	6	3	1	0	18	.21	.51
	3	1	1	6	4	3	2	0	17	.15	.47
	Sum	12	28	28	20	7	4	1	100		

	Correctly classified statements								Acc	CEM^{ORD}
	0	1	2	3	4	5	6	Sum		Mean
Very conservative	4	3	1	0	0	0	1	9	.13	.46
Conservative	0	9	2	3	1	0	0	15	.21	.51
Moderate	6	6	6	7	0	1	0	26	.20	.50
Liberal	2	8	13	4	4	2	0	33	.16	.50
Very Liberal	0	2	6	6	2	1	0	17	.21	.51
Sum	12	28	28	20	7	4	1	100		

[Amigó et al. 2020] Enrique Amigó, Julio Gonzalo, Stefano Mizzaro, and Jorge Carrillo-de-Albornoz. 2020. An Effectiveness Metric for Ordinal Classification: Formal Properties and Experimental Results. In ACL'20. DOI: 10.18653/v1/2020.acl-main.363

RQ4: Worker Behavior

- **Time**: the amount of time spent on average by workers on each statement decreases while the statement position increases ⟶ **learning effect**
- **Worker Signals**: workers provide many signals that correlate with the quality of their work that can be exploited to aggregate individual judgments
  - individual scores aggregated using weighted mean with (political views and CRT performances as weights)
  - No noticeable increase in external agreement
- **Queries**: the amount of queried issued on average by workers on each statement lowers while the position increases
  - Fatigue, boredom, learning effect
  - More than one query for each statement position: **query reformulation**

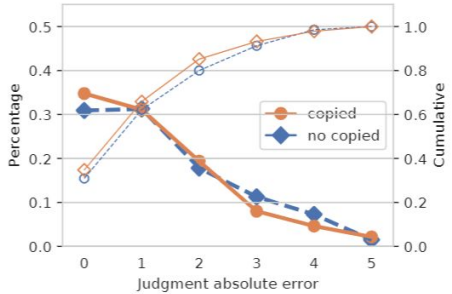


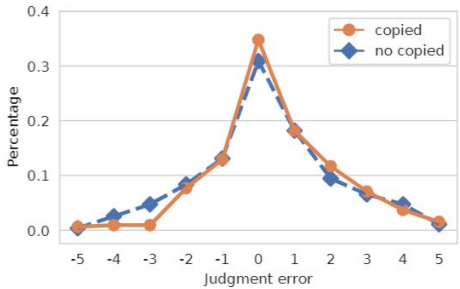
Statement Position	1	2	3	4	5	6	7	8	Sum	Mean
Number of Queries	352 16.8%	280 13.4%	259 12.4%	255 12.1%	242 11.6%	238 11.3%	230 11.0%	230 11.4%	2095	261.9
Statement as Query	22 9%	32 13%	31 12.6%	33 13.5%	34 13.9%	30 12.2%	29 11.9%	34 13.9%	245	30.6

RQ5: Sources of Information

- **URL Analysis:** many fact checking websites among the top-10 URLs
  - **Workers tend to identify trustworthy information sources to support their judgments**
- **Justifications**: we correlate how the workers provide their justification with their quality.
  - Justifications with text copied from the selected webpage or not, with or without “free text” generated by each worker
  - **Workers of high quality tend to read the text** from the selected web page and to report in within the justification box
  - **Workers which directly quote the text do n the truthfulness of the statement**

URL	Percentage%
snopes.com	11.79%
msn.com	8.93%
factcheck.org	6.79%
wral.com	6.79%
usatoday.com	5.36%
statesman.com	4.64%
reuters.com	4.64%
cdc.gov	4.29%
mediabiasfactcheck.com	4.29%
businessinsider.com	3.93%





Conclusions / Take Home Messages

- **RQ1:** workers are able to detect and objectively categorize (mis)information related to the COVID-19 pandemic
- **RQ2**: both crowdsourced and expert judgments can be transformed and aggregated to improve label quality
- **RQ3**: workers' political background is indicative of label quality
- **RQ4:** there are several behavioral signals related with worker quality
- **RQ5:** workers use multiple sources of information and there are relations between justification provided by workers and their quality

Thank You!

Contacts:

- Università degli Studi di Udine, Udine, Italy
  - Kevin Roitero - roitero.kevin@spes.uniud.it
  - **Michael Soprano - soprano.michael@spes.uniud.it**
  - Beatrice Portelli - portelli.beatrice@spes.uniud.it
  - Vincenzo Della Mea - vincenzo.dellamea@uniud.it
  - Giuseppe Serra - giuseppe.serra@uniud.it
  - Stefano Mizzaro - mizzaro@uniud.it
- University of Queensland, Brisbane, Australia
  - Gianluca Demartini - g.demartini@uq.edu.au
- RMIT University, Melbourne, Australia
  - Damiano Spina - damiano.spina@rmit.edu.au