News Source Credibility in the Eyes of Different Assessors

Martino Mensio, Harith Alani
Knowledge Media Institute, The Open University, UK
{martino.mensio,h.alani}@open.ac.uk

Abstract

With misinformation being one of the biggest issues of current times, many organisations are emerging to offer verifications of information and assessments of news sources. However, it remains unclear how they relate in terms of coverage, overlap and agreement. In this paper we introduce a comparison of the assessments produced by different organisations, in order to measure their *overlap* and *agreement* on news sources. Relying on the general term of *credibility*, we map each of the different assessments to a unified scale. Then we compare two different levels of credibility assessments (source level and document level) by using the data published by various organisations, including fact-checkers, to see which sources they assess more than others, how much overlap there is between them, and how much agreement there is between their verdicts. Our results show that the *overlap* between the different origins is generally quite low, meaning that different experts and tools provide evaluations for a rather disjoint set of sources, also when considering fact-checking. For *agreement*, instead we find that there are origins that agree more than others on the verdicts.

1 Introduction

In a public sphere polluted by different shapes of misinformation, the most important role is played by the citizens (Cooke, 2017), with their ability to think critically and investigate while consuming a piece of news. In order to investigate, there is a wide variety of tools that can help, by providing indicators of credibility on different levels.

Checking the source that published a piece of news is one of the first steps often referred to in media literacy (Wineburg et al., 2016). Identifying who created the content and where it has been published can tell a lot about its credibility even before inspecting the news itself.

There are many efforts by journalists, fact-checkers and communities that annotate the relationship between the sources and misinformation. Various tools exist that rely on such annotations to notify users of the validity (or invalidity) of the information they are looking at.

It is inevitable that different verification organisations will sometime produce slightly different, or even conflicting, assessments of certain news articles or sources. It is also natural for verification sources to focus on news sources and not others. Such diversity is often needed, to reduce bias and to encourage further debate. This paper aims to reach a better awareness of such overlap and diversity, as a first step towards understanding how to relay such information to end users, and the potential impact on their final judgements.

To the end, this study compares the different information assessments available, looking for similarities and differences between them. In order to perform this comparison, however, there is a set of problems that need to be addressed. Each assessor uses a different set of labels and scores in their assessments. This variety stems from different criteria and methods of assessment.

Given the above, in this paper we investigate the following Research Questions:

1. How much do different assessments *overlap* when evaluating the same sources?
2. How often do different assessments, and assessors, *agree*? And which ones are more similar or different to each other?
3. Do fact-checkers check claims from sources that have been assessed at the source-level? And do the outcomes agree?

To answer these questions, this work presents two main contributions. First, after describing in Section 2 existing works that define credibility, in Section 3 we present our strategy to combine and *make comparable* different existing assessments,

mapping from their specific features to a common measure of credibility. The second is the measurement of *overlap* and *agreement* in Section 4, to express how much the considered assessments evaluate the same sources and whether or not they have the same outcomes. Then Section 5 presents the main challenges that we see for the next steps and Section 6 concludes the paper.

2 Related work

In this section, we first look into studies that define and formalise the concept of credibility. Then, after presenting the main assessments that are already available, we take a look at some methods for combining them together.

2.1 Credibility formalisation

One of the first works on source credibility (Hovland and Weiss, 1951) identified two main components (trustworthiness and expertise), and subsequent studies underlined the many different dimensions to be analysed (Gaziano and McGrath, 1986). Meyer (1988) identified two main components of credibility (*believability* and *community affiliation*), analysing their correlation with a set of items (bias, fairness, accuracy, patriotism, trust, security) retrieved by interviewing users. Later works (Abdulla et al., 2004; Yale et al., 2015; Kohring and Matthes, 2007) use this model and find variations on the number of factors and the importance of the items. A recent work (Prochazka and Schweiger, 2019) compares measures of credibility and trust, to obtain a generalised measure of trustworthiness of news media. From the point of W3C, the Credible Web Community Group[1] aims to standardise the assessments and credibility signals.

From the analysis of these studies, we have as objectives to define which specific declination of *credibility* we want to use and to understand the relationships of different components (bias, factuality, security, intent to deceive) with the selected measure of credibility. Of the two components identified in Meyer (1988), we are more interested in the *believability* component, because we want to have a measure of how much factual information a certain source provides and the quality of its journalism. Instead, the component of *community affiliation* is more related to opinions and points of view, and it may be related to political position, partisanship and a set of factors that go beyond the scope of evaluating credibility. Then there are mixed factors (e.g., bias) that can be generated by the community affiliation but affect the believability as well.

2.2 Real-world assessments

Here we present the existing assessments that provide credibility-related assessments, focusing mainly on human-generated evaluations that come from experts in journalism. We can identify two main levels of evaluations: source-level and document-level.

2.2.1 NG: NewsGuard

On the source-level assessments, we start with NewsGuard,[2] that provides ratings of reliability of online news brands through "nutrition labels". The ratings are generated by journalists that rate the website compliance to a set of criteria.[3] These criteria are divided into two groups: *credibility* (publishing false content, presenting information responsibly, regularly correcting errors, distinguishing news and opinions, avoiding deceptive headlines) and *transparency* (ownership and financing, clearly labelling advertising, management, providing names of content creators). The websites also receive an overall *score* (a sum of points for each criteria) and a *label*, that can be `Trustworthy`, `Negative`, `Satire` or `Platform` (user-generated contents, e.g., social networks or blogs).

2.2.2 MBFC: Media Bias/Fact Check

Another origin of ratings is Media Bias/Fact Check,[4] that provides reports assessing the factual level of the sources (`VERY LOW`, `LOW`, `MIXED`, `HIGH`, `VERY HIGH`) and the kind of bias (`CONSPIRACY-PSEUDOSCIENCE`, `LEAST BIASED`, `LEFT BIAS`, `LEFT-CENTER BIAS`, `PRO-SCIENCE`, `QUESTIONABLE SOURCE`, `RIGHT BIAS`, `RIGHT-CENTER BIAS`, `SATIRE`). The evaluations are produced by a team of editors with their own methodology.[5]

[1] https://credweb.org/

[2] https://www.newsguardtech.com/

[3] http://www.newsguardtech.com/ratings/rating-process-criteria/

[4] https://mediabiasfactcheck.com/

[5] https://mediabiasfactcheck.com/methodology/

2.2.3 WOT: My Web of Trust

My Web of Trust[6] is a crowdsourced reputation service that provides assessments of websites through a browser extension. It provides two components of *trustworthiness* ("How much do you trust this site?") and *safety* ("How suitable is this site for children?"), in terms of a score and a confidence measure.[7] Users can see the overall ratings and comments coming from the community and can provide their own rating.

2.2.4 OS: OpenSources

OpenSources[8] was a resource for assessing online information sources. The list was created looking for overall inaccuracy, extreme biases, lack of transparency, and other kinds of misinformation. The possible tags given to websites are the following: `fake news`, `satire`, `extreme bias`, `conspiracy theory`, `rumor mill`, `state news`, `junk science`, `hate news`, `clickbait`, `proceed with caution`, `political` and `reliable`. A domain can have up to three different tags. The list was recently taken down, possibly due to lack of updates, where many of the website they assessed have been shut down or have been renamed (usually because of *domain hopping*).[9]

2.2.5 IFCN: International Fact Checking Network

The International Fact Checking Network (IFCN) comes into play as domain-assessor considering that it evaluates the signatory compliance with a set of principles.[10] IFCN reviews twelve aspects grouped into five principles (organisation, non-partisanship and fairness, transparency of sources, transparency of funding and of organisation, transparency of methodology, correction policy) and the assessment lasts for one year from the certification date. IFCN is currently considered as the reference point for the credibility of fact-checkers, as third-party fact-checking initiatives arise on big platforms (e.g., Facebook,[11] Youtube[12] and Google Search[13]).

2.2.6 NTT: Newsroom Transparency Tracker

Newsroom Transparency Tracker[14] is an initiative part of The Trust Project, that shares the information published by media outlets with respect to four Trust Indicators (best practices, journalist expertise, type of work, diverse voices). Each source is characterised by 15 attributes, belonging to one of the indicators, that can be `full`, `partial` or `none` compliant.

2.2.7 FC: Fact-checkers

On the document and claim level we have fact-checkers, that manually review and debunk stories by using different rating methods. In the last years there has been a standardisation process towards a `ClaimReview` schema,[15] in order to publish the reviews in a structured format. However, only some fact-checkers adopted the standard,[16] and those who did sometimes differ in how they use some of the `ClaimReview` schema fields.

2.3 Combining assessments

Looking for ways to combine existing assessments, we found studies spanning several ideas. Trust Networks (Golbeck et al., 2003) are built for user-user ratings of trust, creating a graph where users are nodes and trust scores are directed edges. Using rules of network capacity (the maximum amount of trust between a source and a sink is limited by the smallest edge weight along the path) it is possible to infer the trust between an arbitrary couple of nodes. A different model of trust propagation is presented in (Bistarelli and Santini, 2007), also accounting for confidence values. The authors combine different evaluations on the same node by taking the one with the higher value of confidence, expressed by the source node. Similarly, the theory of *belief functions* (Beynon et al.,

[6] https://www.mywot.com/

[7] https://www.mywot.com/wiki/index.php/API

[8] http://www.opensources.co/ down from 2019

[9] https://www.buzzfeednews.com/article/craigsilverman/publishers-are-switching-domain-names-to-try-and-stay-ahead

[10] https://ifcncodeofprinciples.poynter.org/know-more/what-it-takes-to-be-a-signatory

[11] https://www.facebook.com/help/publisher/182222309230722

[12] https://www.poynter.org/fact-checking/2019/youtube-is-now-surfacing-fact-checks-in-search-heres-how-it-works/

[13] https://www.blog.google/products/search/fact-check-now-available-google-search-and-news-around-world/

[14] https://www.newsroomtransparencytracker.com/

[15] https://schema.org/ClaimReview

[16] https://reporterslab.org/a-better-claimreview-to-grow-a-global-fact-check-database/

2000) models decisions with uncertainty and multiple agents. However, it lacks a weighting mechanism to include the credibility of the source nodes providing the evaluations. We discuss in Section 5 the requirements for a decision criterion when the outcomes of evaluations disagree.

3 Approach

In order to combine the different assessments, we build a graph in which the nodes represent registered domain names (referred to as "sources" throughout this document), and the edges instead correspond to the assessments retrieved from the origins listed earlier. An origin can be different from the source, when we retrieve a set of evaluations from a third party. For example, the data for the fact-checkers coming from DataCommons[17] contains as source nodes the fact-checkers that created the corresponding `ClaimReview`. In this work we are keeping as nodes only the sources, without also representing the documents, in order to simplify this preliminary study. Therefore, if the assessments point to a specific document (identifiable with a URL), the tail of the edge goes to the source extracted from the URL. Multiple parallel edges can exist between the same couple of nodes.

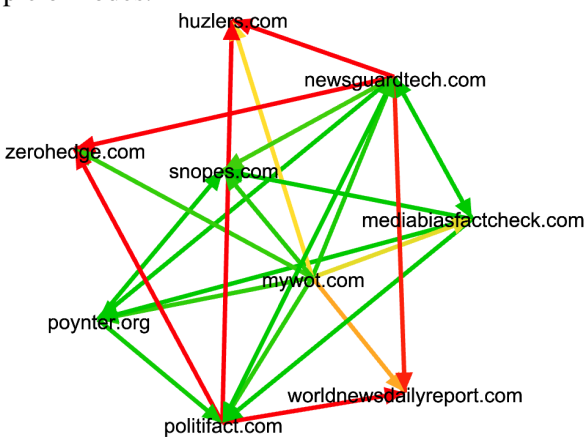

Figure 1: A small sample of the source-level credibility graph. The colours of the edges represent the value of the credibility expressed: red stands for negative assessments, green for positive ones, passing gradually through shades of yellow.

In Figure 1 we can see a small sample of the graph, representing some connections that come from the sources of data available. In this small sample we can already see some disagreement on the evaluation of the domain `zerohedge.com`, as will be analysed in Section 4.

3.1 Credibility definition

We use the term *credibility* to capture the different assessments provided by the various originators listed in Section 2. Credibility is a measure that can be positive or negative. It represents how the assessor judges a given source, which tends to be based on factors of factuality (objective factors) and believability (subjective factors) (Meyer, 1988).

Credibility value is relative to the origin that gave the assessment. This means that it is also dependent on the credibility values of the assessors themselves.

A credibility measure is characterised by a *value*, that indicates whether it is positive, negative or neutral with a range of $[-1; +1]$. It also has a measure of *confidence* that expresses the level of certainty of the assessment, in the range $[0; +1]$.

3.2 Mapping the vocabulary

From the theories analysed in Section 2.1 we can score importance of the factors according to our definition of credibility: first indications of factuality and adherence to journalistic standards and transparency, then the ones about extremism (e.g., conspiracy theories) and intent to deceive (e.g., clickbait, misleading) (Volkova and Jang, 2018). To translate the values provided by the various origins described in Section 2.2, we used the method described in Table 1. These values are obtained by taking indicators of credibility that are relevant for the believability component and not for the community affiliation. In most cases, the mapping is just a selection of indicators coming from the origin and a translation to the desired interval for credibility and confidence. The ones that are most opinionated are the values used for OS due to the big variety of labels that it provides. However, the qualitative comparison in the following is made on the original values and not on our mapping values, in order to be able to discuss and optimise these parameters later.

4 Analysis and Results

In this section, after presenting some statistics about the data that have been collected, we describe how we measure the overlap and agreement between two assessment origins, showing the re-

[17] https://datacommons.org/factcheck/download

origin	credibility formulation	confidence formulation
NG	linear $score[0;100] \rightarrow cred[-1;1]$ exception `Platform`, `Satire` $\rightarrow cred = 0$	$conf = 1$ exception `Platform`, `Satire` $\rightarrow conf = 0$
MBFC	$factuality\{LOW, MIXED, HIGH\} \rightarrow cred\{-1, 0, 1\}$	$conf = 1$ if $factuality$, otherwise $conf = 0$
WOT	$trust.score[0;100] \rightarrow cred[-1;1]$	$trust.conf[0;100] \rightarrow conf[0;1]$
OS	`fake` $\rightarrow -1$, `reliable` $\rightarrow 1$ `conspiracy`, `junksci` $\rightarrow -0.8$ `clickbait`, `bias` $\rightarrow -0.5$ `rumor`, `hate` $\rightarrow -0.3$ all other tags $\rightarrow 0$	$conf = 1$ when credibility is not null otherwise $conf = 0$
IFCN	starting from $cred = 1$ apply penalties for `partially` (0.05) and `none` (0.1) compliant with lower bound $cred = 0$	if expired signatory $conf = 0.5$ otherwise $conf = 1$
NTT	proportionally to the number of indicators satisfied, `partial` compliance counts half	$conf = 1$
FC	if `ratingValue` in the `reviewRating`[18] $cred = \frac{\texttt{ratingValue-worstRating}}{\texttt{bestRating-worstRating}} * 2 - 1$ otherwise map the `alternateName` with a known list (e.g. `true` $= 1.0$)	if the mapping is successful $conf = 1$ otherwise $conf = 0$

Table 1: The mapping to obtain credibility and confidence values for each of the origins considered.

sults of both measures on the considered subsets of information. Then we show a similar analysis comparing the source-level assessments with the data coming from the fact-checkers, thus providing a response to the third Research Question.

4.1 Data statistics

For collecting the data used in this paper, we first took the first 1 million entries from the list contained in the Open PageRank dataset,[19] reducing it to 762194 sources by removing the subdomains. This list has been used in order to retrieve the assessments An overall view of the data retrieved from each of the origins listed in Section 2.2 can be seen in Table 2. Looking at the second column, we can see that the biggest origin of ratings is WOT, covering 40% of the sources used as input. Then all the other rating origins have a size that is orders of magnitude smaller. Given these sizes, it is clear that many domains only have one or two assessments available. Together with this measure of coverage, we also show the average value of credibility assigned by using our mappings, that gives an idea of the average verdict coming from the considered origin.

4.2 Overlap

The first of our Research Questions is about the overlap, defined as the condition when two origins evaluate the same source. Given the different number of ratings from each origin, we want to have a measure that expresses in a meaningful way the relationship between the sets of sources with respect to the origins. One measure is the Jaccard index (Jaccard, 1901), that is specific to measuring the overlap of sets. However, when one of the sets is much bigger than the other, the index values become too insignificant. A related measure is Szymkiewicz-Simpson coefficient that, while accounting for size scale problem, is still a symmetric measure considering as denominator the smaller size between the two sets (Vijaymeena and Kavitha, 2016). For these reasons, we selected an asymmetrical measure of overlap (Yoshida et al., 2001) that considers the ratio of the intersection of two sets over the size of the second set:

$$overlap(A \rightarrow B) = \frac{|A \cap B|}{|B|}$$

In this way, we can establish how much of the first set belongs to the second one. Figure 2 shows the pairwise values of this measure.

Looking at the first column, we can see that all the origins have a high overlap with WOT, meaning that it almost covers the entirety of domains under analysis. Nevertheless, there are domains that have not been rated by WOT, which are

origin	#sources	avg. credibility
WOT	308155	0.4264
NG	2795	0.5433
MBFC	2404	0.3874
OS	811	-0.6618
IFCN	82	0.8786
NTT	52	0.4256

Table 2: The top origins of assessments, in terms of number of rated sources.

[18] https://schema.org/Rating

[19] https://www.domcop.com/openpagerank/what-is-openpagerank

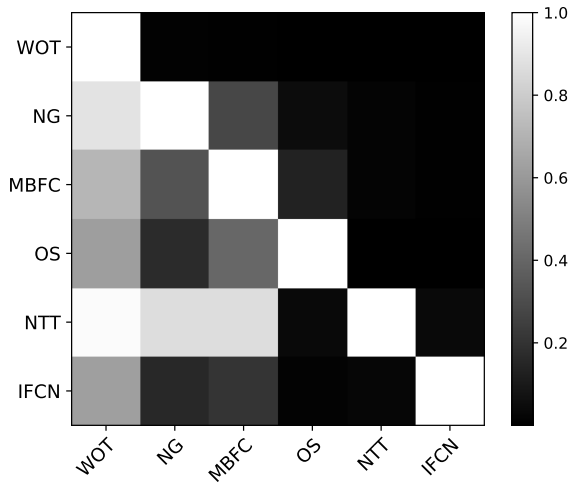

Figure 2: The directed overlap coefficient between each pair of origins of credibility assessment. The value of each cell shows the percentage of how many sources evaluated from the assessor on the row have also been evaluated by the assessor on the column. Darker colours mean lower overlap.

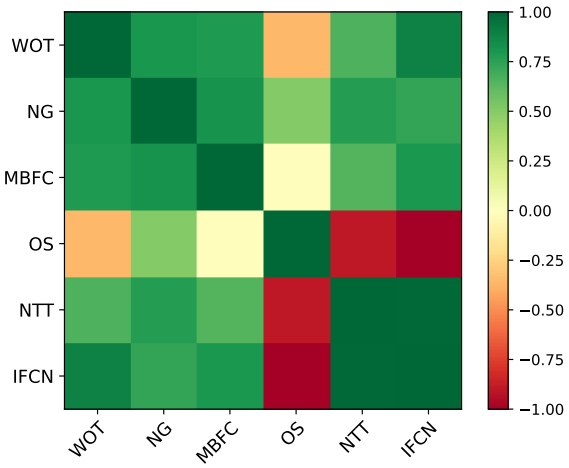

Figure 3: The pairwise cosine similarity between the different origins of credibility assessments, calculated from the nodes that they both assess. Green values correspond to high similarity, and red to disagreement.

mainly recent websites created after 2016.[20] We can also observe that many credibility origins have a high overlap ratio with MBFC and NG, even though the two sets have a low overlap themselves. Overall, we see that the overlap is quite low, with an overall average value of 0.276.

4.3 Agreement

For measuring the agreement of the credibility assessments, the vocabulary mapping described in 3.2 was used to convert the assessments to the common scale of credibility. To answer the second Research Question, we use as a measure of agreement the *cosine similarity* of the evaluations provided by pairs of origins on the sources that they both rate. Input to the similarity computation is two arrays of scores of credibility with an element for each of the sources that have been evaluated by both origins of assessments. We can interpret the cosine similarity as a measure of correlation between the two origins. Cosine similarity gives a value that is bound in the interval $[-1; +1]$, with $+1$ being an indication of total agreement and -1 of total disagreement: the angle formed between the two multidimensional vectors of ratings (features) gives an indication of discrepancy between the two origins considered.

In Figure 3 we can see the values of agreement highlighting groups of origins that agree more than others. For example, the group (WOT, NG, MBFC) has very high values of agreement. The high values in the row of IFCN instead shows that in most cases, the sources agree on giving high credibility to the IFCN signatories.

There are however a few values that demonstrate some disagreement, that need to be taken into account together with the corresponding overlap of each cell. In the specific case of intersection between IFCN and OS, the red colour is caused by the only domain `weeklystandard.com` that they both rate: IFCN rates it with a positive score, since it is a signatory (at the moment with an expired assessment),[21] while OS tags this source as `political` and `extreme bias`, that has a generally negative connotation. Another example is the intersection between NTT and OS on the domain `ijr.com` that satisfies 10 criteria out of 15 for NTT, while OS tags it as `political`, `bias` and `unreliable`. This domain is positively reviewed also by NG and MBFC. Even between origins that generally agree, there are some interesting cases: `zerohedge.com` is rated as 90% trustworthy by WOT, while NG gives a completely negative score, saying that it "*severely violates basic standards of credibility and transparency*". The website is tagged as `conspiracy` also by OS. Similarly, there is a big group of sources[22]

[20] https://www.pcworld.com/article/3139814/web-of-trust-browser-extensions-yanked-after-proving-untrustworthy.html

[21] https://ifcncodeofprinciples.poynter.org/profile/tws-fact-check

[22] `freedomsphoenix.com`, `rawstory.com`,

that OS tags as `conspiracy`, `clickbait` and `bias`, but on WOT they are considered as trustworthy.

In general, we see many cases where supposedly high-quality origins do not agree. Such disagreement might be due to difference in opinion, to the use of different assessment methods, or simply to the difference in what exactly is being assessed.

4.4 Fact-checking vs source assessments

In order to analyse the relationship between the assessments of the sources with the detailed evaluations done by fact-checkers (RQ3), we considered two different groups of origins: source-level assessments on the one side and fact-checkers on the other. For the latter, the choice of fact-checkers was done by the availability of annotated data: a subset of IFCN signatories using `ClaimReview` with the addition of the data from EuVsDisinfo.[23] Table 3 shows how many different sources have been fact-checked by each of the fact-checkers. The low values are often due to the lack of proper annotation of the Claim `appearance`.[24]

Fact Checker	#sources of claims
politifact.com	367
euvsdisinfo.com	244
factcrescendo.com	27
factcheckni.org	27
aap.com.au	20
factly.in	18
tempo.co	10
lemonde.fr	10

Table 3: Fact-checkers sorted by decreasing number of unique sources checked.

We evaluated overlap and agreement in the same way as in previous sections, with the only difference that in this case we are comparing the fact-checkers with regard to the source-level assessments. Figure 4 shows both overlap and agreement measures for this comparison.

Starting with overlap, the fact-checker data that we have collected shows that only a small portion of the claims reviewed appear on sources known by the domain-level origins, with the exception of

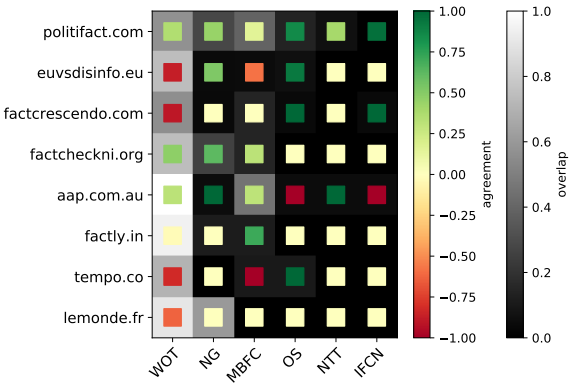

Figure 4: The overlap and agreement measures between fact-checkers and source-level assessors. For overlap, the value of each cell shows how much the domains reviewed by the fact-checkers have also been evaluated by the origins on the columns. The agreement is evaluated on sources that have both ratings.

WOT. NG and MBFC have still significant overlap with some of the fact-checkers.

Considering the agreement instead, we can see that the column of NG has always non-negative similarities (meaning that the outcomes of fact-checks in most cases match with the label given to the source by the assessment in the column), and this could indicate a match between the criteria evaluated and the effective history of publishing genuine news. Looking row-wise instead, we see that PolitiFact is the fact-checker that mostly agrees with domain-level assessments.

The only row to disagree with the IFCN source-level assessment (meaning that a fact-checker has been debunked) is `aap.com.au`, that in our dataset debunked 3 times `abc.net.au`. Always on the AAP row, the disagreement with OS is because of `dailytelegraph.com.au`, that has been fact-checked several times with opposing labels (some stories are true and some are false) resulting in an average of 0.2 credibility (still positive), while OS tags it as `bias/rumour`. In this case this comparison shows the importance putting a higher weight on fact-checks with negative results (publishing false content is more penalising).

Inspecting the other rows, we find different types of disagreement. The first group relates to the `platforms`, such as Facebook, Twitter, YouTube, Blogspot and Instagram, that have been fact-checked on a more detailed level (post/profile/blog) by the different fact-checkers for which we have the data. On sources with user-generated content, NG does not provide an evaluation itself, but it seems that WOT instead gen-

`informationclearinghouse.info`, `antiwar.com`, `allnewspipeline.com`, `americablog.com`

[23] https://euvsdisinfo.eu/

[24] https://schema.org/Claim

erally gives a positive score. This highlights the need for credibility assessments at a more granular level, because the domain-level is sometimes too generic. We can have credibility on the source itself, but having different values for parts of it.

In addition to the above, some domains are websites that are just mediators to reaching the original content. In this category we can see URL shorteners (e.g., `bit.ly`) and web archives (e.g., `archive.org`). The annotated claim appearances should point directly to the origin of the content, and this set of examples underlines the need to add relationships with different semantics (e.g., `sameAs`) to indicate redirects.

5 Open challenges

From the results of our analysis, we see a set of open challenges that we aim to point out.

First of all, *how to handle disagreement*? Why should one rely on one origin of information instead of another? To answer these questions, it is not enough to look at the confidence expressed by the origin itself, because this only expresses their beliefs. We might need to adapt the Dempster-Shafer theory (Beynon et al., 2000) in order to include also an external weight that measures the trust that a certain origin deserves. To obtain this trust factor we need to compute the credibility that the origin under analysis has. Between the score of credibility (in the range $[-1; +1]$) and the trust factor (that is a weight in the range $[0; 1]$) we need to define a relationship function that is able to have higher trust of the origins that have a higher credibility, but still takes into consideration origins that have a neutral (or also negative) credibility. This process can be applied recursively until we reach a set of *a-priori* origins that we (or the users, allowing customisation) believe in. Overall, the decision on disagreement cases needs to be taken with a multi-weighted voting mechanism, accounting for *i)* confidence of the assessment itself *ii)* trust in the origin and *iii)* granularity level.

A second challenge is how to present this kind of results to the public. In order to avoid critics (as happened to NG[25] or to the Poynter list of Unreliable News Sources[26]), the direction that needs to be taken is to provide a *risk of misinformation* measure instead of saying that a certain source is not good or credible. And we need to provide support for the score provided, by linking it to the original assessments, delegating to them their own statements. Then the explanation needs to be complete also in the case of multi-step assessments, when also the origins of the assessments have been evaluated. In this case we need to provide the score and the evidence also for the origins involved.

6 Conclusions

This paper presented a comparison of different origins of credibility assessments.

The comparison has been done using measurements of overlap, in order to see whether different origins evaluate the same domains, showing a quite low average overlap. This underlines the importance of combining different origins, in order to achieve better coverage over more domains.

Our measurements of agreement have shown groups of origins that agree more and some that agree less. This sets the importance of effectively combining the different assessments with credibility measures of the origins themselves.

Furthermore, this paper compared the data coming from source and document level, looking at discrepancies between the two levels. This underlined the importance of being able to be more specific than only labelling the domain with a specific value of credibility.

We acknowledge the many limitations of this study, particularly due to the limited availability of fact-checking data and to the oversimplification of just considering the source level. In particular, having just one unified measure of credibility, without doing a multidimensional comparison (e.g., factuality, bias, intention), seems to be limiting and the citizens should be informed about each component of the score provided.

In future work we aim to investigate the challenges identified in Section 5, in order to account for disagreeing origins and to have a way to integrate the different measures of credibility for the sources involved. Additionally, we plan to extend the nodes definition not only to news sources but also allowing document and claim level nodes, in the direction of creating a Credibility Knowledge Graph that can be explored and used to improve the inference at different levels.

[25] https://www.niemanlab.org/2019/01/newsguard-changed-its-mind-about-the-daily-mails-quality-its-green-now-not-red/

[26] https://www.poynter.org/letter-from-the-editor/2019/letter-from-the-editor/

Acknowledgments

This research was supported by the EU Project Co-Inform which has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 770302.

References

Rasha A Abdulla, Bruce Garrison, Michael B Salwen, Paul D Driscoll, and Denise Casey. 2004. Online news credibility. In *Online news and the public*, pages 167–184. Routledge.

Malcolm Beynon, Bruce Curry, and Peter Morgan. 2000. The dempster–shafer theory of evidence: an alternative approach to multicriteria decision modelling. *Omega*, 28(1):37–50.

Stefano Bistarelli and Francesco Santini. 2007. Sclp for trust propagation in small-world networks. In *International Workshop on Constraint Solving and Constraint Logic Programming*, pages 32–46. Springer.

Nicole A Cooke. 2017. Posttruth, truthiness, and alternative facts: Information behavior and critical information consumption for a new age. *The Library Quarterly*, 87(3):211–221.

Cecilie Gaziano and Kristin McGrath. 1986. Measuring the concept of credibility. *Journalism quarterly*, 63(3):451–462.

Jennifer Golbeck, Bijan Parsia, and James Hendler. 2003. Trust networks on the semantic web. In *International workshop on cooperative information agents*, pages 238–249. Springer.

Carl I Hovland and Walter Weiss. 1951. The influence of source credibility on communication effectiveness. *Public opinion quarterly*, 15(4):635–650.

Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579.

Matthias Kohring and Jörg Matthes. 2007. Trust in news media: Development and validation of a multidimensional scale. *Communication research*, 34(2):231–252.

Philip Meyer. 1988. Defining and measuring credibility of newspapers: Developing an index. *Journalism quarterly*, 65(3):567–574.

Fabian Prochazka and Wolfgang Schweiger. 2019. How to measure generalized trust in news media? an adaptation and test of scales. *Communication Methods and Measures*, 13(1):26–42.

MK Vijaymeena and K Kavitha. 2016. A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal*, 3(2):19–28.

Svitlana Volkova and Jin Yea Jang. 2018. Misleading or falsification: Inferring deceptive strategies and types in online news and social media. In *Companion of the The Web Conference 2018 on The Web Conference 2018*, pages 575–583. International World Wide Web Conferences Steering Committee.

Sam Wineburg, Sarah McGrew, Joel Breakstone, and Teresa Ortega. 2016. Evaluating information: The cornerstone of civic online reasoning. *Stanford Digital Repository. Retrieved January*, 8:2018.

Robert N Yale, Jakob D Jensen, Nick Carcioppolo, Ye Sun, and Miao Liu. 2015. Examining first-and second-order factor structures for news credibility. *Communication Methods and Measures*, 9(3):152–169.

Hideyuki Yoshida, Takehiko Shida, and Toshiki Kindo. 2001. Asymmetric similarity with modified overlap coefficient among documents. In *2001 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (IEEE Cat. No. 01CH37233)*, volume 1, pages 99–102. IEEE.