Assessing the Credibility of Claims on the Web

Kashyap Popat
Supervised by Prof. Dr. Gerhard Weikum
Max Planck Institute for Informatics
Saarland Informatics Campus, Saarbrücken, Germany
kpopat@mpi-inf.mpg.de

ABSTRACT

In my doctoral research, I plan to address the problem of assessing the credibility of arbitrary claims made in natural-language text — in an open-domain setting. Automatic credibility assessment is a complex task depending upon many factors. To start with, we propose three factors which can help in assessing the credibility of textual claims: (i) the reliability of the web sources talking about the claim, (ii) the language style of the articles reporting the claim and, (iii) their stance (i.e., support or refute) towards the claim. In addition, we also focus on extracting user-interpretable explanations as evidence supporting the verdict of the assessment.

Keywords

Credibility Analysis; Rumor Detection; Text Analytics

1. PROBLEM

Motivation: The web has been a vast resource of valuable information. However, in recent times, the spread of false claims in social media, other web-sources, and even in news has become hazardous for information credibility on the web. The World Economic Forum identified "the rapid spread of misinformation online" as one of the top 10 challenges the world faces [5]. Recently, Facebook enlisted five fact-checking organizations to review stories that are flagged by users as potentially fake [8] and Google News also introduced fact check features [6] to combat this critical problem.

With the increasing number of hoaxes and rumors, truth-checking websites like *snopes.com*, *politifact.com*, *truthorfiction.com* and others have become popular. These websites compile articles written by experts who manually investigate contentious claims by determining their provenance and authenticity from various sources; and provide a verdict (true or false) with supporting evidence. The work in my research aims to replace this manual verification/falsification with a robust automated system.

Determining the credibility of a claim automatically is an extremely challenging task. As studied in [9], even humans sometimes cannot easily distinguish hoax articles in Wikipedia from authentic ones, and quite a few people have mistaken satirical articles (e.g., from theonion.com) as truthful news.

WWW 2017 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3053379

Claim: Facebook soon plans to charge monthly subscription fees to users of the social network.
Assessment: False
Explanation: The rumor that Facebook will suddenly start charging users to access the site has become one of the social media era's perennial chain letters.

Table 1: A sample claim with assessment and explanation.

Limitations of State of the Art: Prior approaches for credibility assessment (e.g., [4, 11, 12, 14]) are limited to the structured data – resolving conflict amongst multi-source data. Other credibility assessment approaches like [3, 9, 20] work only in restricted social media settings and rely heavily on platform specific features like "number of edits", "followers", "retweets" etc. None of these approaches make an attempt to address the problem of assessing credibility of arbitrary *textual claims*, expressed freely in an *open-domain* setting, without making any assumptions on the structure of the claim or characteristics of the community or website where the claim is made. Also, no prior approaches consider providing user-interpretable explanations in the form of evidence — supporting the verdict of credibility assessment.

Problem Statement: Given a textual claim, assess its credibility and decide if it is true or false and also provide supporting evidence explaining the assessment.

Table 1 shows an example for the input and output of our method. For the given example, our system assesses its credibility as *false*, and provides user-interpretable explanation in the form of informative snippets automatically extracted from an article published by a reliable web-source refuting this claim.

2. STATE OF THE ART

Our work draws motivation from the following research:

Truth discovery: Truth discovery approaches [4, 11, 14, 29–31] aim to resolve conflicts in multi-source data (see [13] for a survey). They assume input data to have a structured representation: an entity of interest (e.g., a person) along with its potential values provided by different sources (e.g., the person's birthplace). It is assumed that the conflicting values are already available. To resolve conflicts for a particular entity, these approaches exploit that reliable sources often provide correct information.

Advancing further, work in [17] proposes a method to generate conflicting *values* or *fact candidates* from Web contents. They use linguistic features to detect the objectivity of the source reporting the fact. However, this work still operates on structured input in the form of Subject-Predicate-Object (SPO) triples for the fact candidates, obtained by applying Open Information Extraction to Web pages.

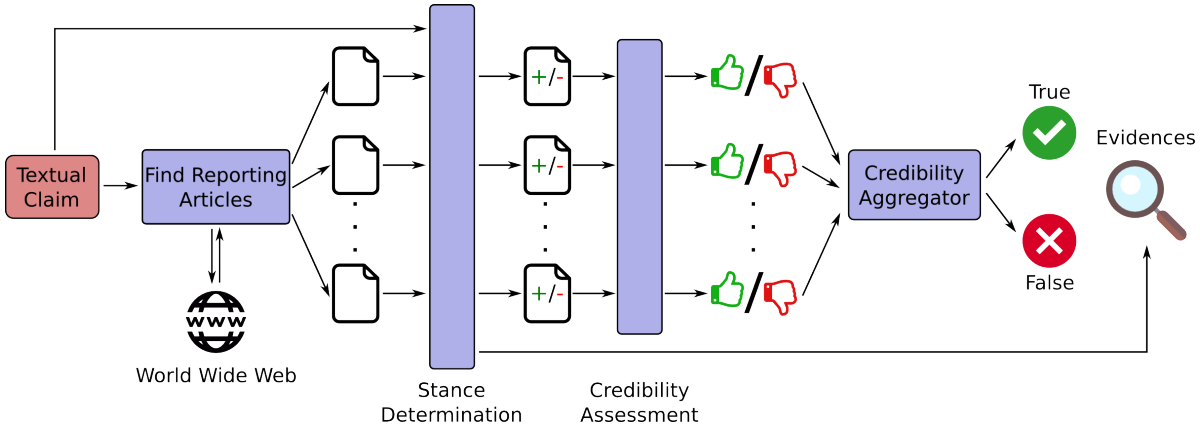

Figure 1: System framework for credibility assessment (+/- labels for articles indicate the stance i.e support/refute towards the claim).

A classic example is: *"Obama is born in Kenya"* viewed as a triple ⟨*Obama, born in, Kenya*⟩ where *"Kenya"* is the critical value. The assumption of such a structure is crucial in order to identify alternative values for the questionable slot (e.g., *"Hawaii"*, *"USA"*, *"Africa"*), and is appropriate when checking facts for tasks like knowledge base curation. However, these approaches are limited in their coverage, and cannot handle many kinds of claims found on news and social media, which are often in the form of long sentences or entire paragraphs.

The method proposed in [12] supports credibility assessment of statements. However, this approach relies on the *user* providing the *doubtful* portion of the input statement (e.g., the birthplace). Based on this doubtful unit, alternative statements (e.g., alternative birthplaces) are generated via web search and ranked to identify the correct statement. Note that these approaches can handle only input statements for which alternative facts or values are given or can be retrieved a priori.

These approaches are limited to resolving conflicts amongst alternative fact candidates (or, multi-source data) in structured datasets. Our work focuses on addressing these limitations and proposing a general approach to process unstructured natural-language claims without requiring any alternative claims.

Credibility analysis within communities and social media: A study in [16] focuses on credibility analysis within online health communities. They use a probabilistic graphical model to jointly infer user trustworthiness, language objectivity, and statement credibility. A similar approach is used to identify credible news articles, trustworthy news sources, and expert users in [15] and to extract *Adverse Drug Reactions* from social media in [28].

Wikipedia hoaxes are studied in [9]. The impact of hoax articles is assessed by measuring (i) how long they survive before being debunked, (ii) how many page-views they receive, and (iii) how heavily they are referred to by documents on the web. This work also proposes a model which can determine if a Wikipedia article is hoax or not using features specific to Wikipedia.

Approaches for credibility assessment of social media posts exploit *community-specific* features for detecting rumors, fake, and deceptive content [3, 20, 26, 27]. Temporal, structural, and linguistic features were used to detect rumors on Twitter in [10]. Study in [7] addresses the problem of detecting fake images in Twitter based on influence patterns and social reputation.

All these approaches are limited to online communities and social media, relying heavily on community-specific characteristics. In contrast, we study credibility in an open domain setting without relying on such explicit signals.

An algorithm for propagating trust scores in a heterogeneous network of claims, sources, and articles is proposed in [23]. However, it does not analyze the claims in free-text form, the language style or the stance. It also requires weak supervision at the evidence level in the form of human judgment on the trustworthiness of articles.

Stance Determination: Opinion mining methods for recognizing a speaker's stance in online debates are proposed in [21, 24]. Structural and linguistic features of users' posts are harnessed to infer their stance towards discussion topics in [22]. However, these approaches are all tailored for debate forums.

Evidence Detection: Approaches for Evidence Retrieval aim to find entire documents which can be used as evidence for a claim [1, 2]. In contrast, we aim to extracts informative textual snippets that support or refute a claim, instead of retrieving entire documents.

3. PROPOSED APPROACH

In order to achieve the automatic credibility assessment, we propose an end-to-end generic framework. Given a claim in the form of a sentence or a paragraph, it first uses a search engine to identify documents from multiple web sources referring to the claim. For these potential sources of evidence and counter evidence, interplay between several factors like the *language* (e.g., bias, subjectivity, etc.) of the retrieved articles, the *reliability* of the web sources where the articles appeared, and the *stance* of these article towards the claim (i.e., whether it supports or refutes the claim) is analyzed. Figure 1 gives a pictorial overview of our proposed framework.

Consider a set of textual claims in the form of sentences or short paragraphs, and a set of web sources hosting articles that report on the claims. Credibility label of each claim can be *True* or *False*. Given the labels of a subset of the claims, our objective is to predict the credibility label of the remaining unlabeled claims. To achieve this, we propose a distant supervision based credibility assessment model.

In this process of automatic credibility assessment, we need to (a) determine the stance (supporting or refuting) of the reporting articles towards the claims. Using this information, we (b) assess the reliability of the underlying sources. Then, we (c) compute the credibility opinion score of each article and finally, (d) aggregate scores from all sources, weighed by their reliabilities, to obtain the overall credibility label of target claims. Figure 1 depicts this flow between the various steps.

4. METHODOLOGY

We propose three factors for assessing the credibility of a textual claim. The following sections explain these factors and how we use them in our model.

4.1 Language Style

The language style in which a claim is reported in an article plays a crucial role in understanding its credibility. Objective and unbiased reporting language increases the credibility of the information. On the other hand, highly subjective or a sensationalized style of reporting brings down the credibility. This hypothesis is validated in [17] through an experiment using Amazon Mechanical Turk.

In order to capture the linguistic style of the reporting articles to model the above hypothesis, we use the set of lexicons from [15], in particular the following types of stylistic features:

Assertive verbs: capture the degree of certainty to which a proposition holds

Factive verbs: presuppose the truth of a proposition in a sentence

Hedges: soften the degree of commitment to a proposition

Implicatives: trigger presupposition in an utterance

Report verbs: emphasize the attitude towards the source of the information

Discourse markers: capture the degree of confidence, perspective, and certainty in the set of propositions made

Subjectivity and bias: a list of positive and negative opinionated words, and an affective lexicon to capture the state of mind (like attitude and emotions) of the writer while writing an article

4.2 Understanding Stance

In order to assess the credibility of a claim, it is important to understand whether the articles reporting the claim are supporting it or not. For example, an article from a reliable source like *iflscience.com* refuting the claim *"Solar panels drain the sun's energy, experts say"*, will make the claim less credible.

In order to understand the stance of an article, we divide the article into a set of snippets, and extract the snippets that are strongly related to the claim. This set of snippets helps in determining the overall score with which the article refutes or supports the claim. We compute both the support and refute scores, and use them as two separate features in our model. We train a classifier based on bag-of-words features which given a snippet, gives the probability of how likely the text refutes or supports a claim. In the later stage, we can also use these snippets as evidence explaining the result of our credibility assessment.

4.3 Credibility-driven Source Reliability

Apart from the reporting style of the article and its stance about the claim, the reliability of the web source hosting the article also has a significant impact on the credibility of the claim. For instance, one should not believe a claim reported by an article from the "The UnRreal Times" website[1], as opposed to a claim on the "World Health Organization" website.

For each web source, we determine the stance of its articles (regarding the respective claims) using the *Stance Classifier* explained above. A web source is considered reliable if it contains articles that refute false claims and support true claims.

4.4 Classification using Distant Supervision

Credibility labels are available *per-claim*, and not per-reporting-article. Thus, we use *Distant Supervision* for *training* — whereby we use the (observed) credibility label of each claim as the credibility opinion of corresponding articles reporting the claim. Then, we train a logistic regression model on this labeled data per reporting article.

For any *test* claim whose credibility label is unknown, along with its corresponding reporting articles, we use this *Credibility Classifier* to obtain the corresponding credibility opinions of the reporting articles. Then, we determine the overall credibility of the claim by considering a weighted contribution of its *per-article* credibility opinions, using the corresponding source reliability values as weights.

The preliminary study of understanding the effects of language style on credibility has been published in [19]. The detailed study along with other contributing factors and evidence extraction is presently under the submission.

Total Claims	4856
True claims	1277 (26.3%)
False claims	3579 (73.7%)
Total Web articles	133272

Table 2: *Snopes* data statistics.

4.5 Case Study: Snopes

To validate our approach, we performed experiments with data from a typical fact checking website: *snopes.com*. *Snopes* covers Internet rumors, hoaxes, urban legends, e-mail forwards, and other stories of unknown or questionable origin [25]. It is a well-known resource for validating and debunking such stories, receiving around 300,000 visits a day [18]. They typically collect these rumors and claims from *Facebook*, *Twitter*, *Reddit*, news websites, e-mails by users, etc.

Each claim, e.g., *"The process of adding fluoride to public water reduces the IQ of the individuals in those areas."*, has a corresponding article verifying it. The credibility verdict (*True* or *False*) is assigned to each claim manually by the Snopes editors. Few of the claims have labels like *Mostly True* or *Mostly False*. We map *Mostly True* labels to *True*, and *Mostly False* labels to *False* — thereby considering only *binary* credibility labels for this work. The credibility verdict is accompanied by a description how the editor(s) came across the claim (e.g., it was collected from a Facebook post, or received by an email from a user etc.), an *Origin* section describing the origin of the claim, and an *Analysis* section justifying the verdict.

We collected these fact-checking articles from *Snopes* published until February 2016 and crawled all details about claims from the web site. For each claim, we fired the *claim text* as a *query* to the Google search engine[2] and extracted the first three result pages (i.e., 30 articles) as a set of articles reporting the claim. We then crawled all these articles from their corresponding web sources. We removed search results from the *snopes.com* domain to avoid any kind of bias. Statistics of the data crawled from *snopes.com* is given in Table 2.

4.5.1 Experimental Setup

We conducted preliminary experiments using the data from *Snopes* to test the performance of our method. In order to remove any training bias, we ignored all *Snopes*-specific references from the data and the search engine results. For addressing the data imbalance issue, we set the penalty for the true class to 2.8[3] — given by the ratio of the number of *false* claims to *true* claims in the *Snopes* data.

[1]A satire, spoof, parody and humour portal: `http://www.theunrealtimes.com/`

[2]Our system has no dependency on the Google. Other search engines or other means of evidence gathering could easily be used.

[3]We set the weight parameter in the LibLinear classifier to attribute a large penalty in the loss function for the positive class.

Configuration	*True* Claims Accuracy (%)	*False* Claims Accuracy (%)	Macro-averaged Accuracy (%)	AUC
LG + ST + SR	**83.21**	**80.78**	**82.00**	**0.88**
ST + SR	80.12	79.22	79.67	0.86
LG + ST	77.47	70.04	73.76	0.81
LG + SR	74.55	68.13	71.34	0.77
ST	72.77	65.17	68.97	0.76
LG	74.12	64.02	69.07	0.75

Table 3: Performance of credibility classification with different feature configurations. LG: language stylistic features, ST: stance features, SR: web-source reliability.

Claim	Verdict & Evidence
Scientific studies demonstrate that the process of adding fluoride to public water reduces the IQ of the individuals in those areas.	**[Verdict]:** False **[Evidence]:** Australia's chief health and medical research agency says fluoride in drinking water does not lower a person's IQ, cause cancer or cause any other negative health effects.
Wrestler 'Big Show' was killed in a car accident.	**[Verdict]:** False **[Evidence]:** A story posted by a blog called "WWE" that claimed the professional wrestling star "Big Show" had died in a car accident is false.
Pranksters briefly changed California's iconic "Hollywood" sign to read "Hollyweed."	**[Verdict]:** True **[Evidence]:** Authorities say someone managed to modify the famed Hollywood sign to read "Hollyweed" in an overnight act of trespass.
Amazon is taking part in a collection effort for Goodwill.	**[Verdict]:** True **[Evidence]:** Amazon and Goodwill are working together to make donating easier for you. Using the Give Back Box platform, a free shipping service, you can donate items you no longer need to Goodwill with ease and bring new life to your empty Amazon box.

Table 4: Example claims with verdict from Credibility Classifier and evidence (top-ranked snippets from articles) from Stance Classifier.

4.5.2 *Evaluation Measures*

Since we have the labeled data from *Snopes*, we can evaluate our model by reporting standard 10-fold cross-validation accuracy. However, *Snopes*, primarily being a hoax debunking website, is biased towards (refuting) the *False* claims — the data imbalance being $2.8 : 1$. Therefore, we report the *per-class accuracy*, and the *macro-averaged accuracy* which is the average of *per-class* accuracy — giving equal weight to both classes irrespective of the data imbalance. We also report the Area-under-Curve (AUC) values of the ROC (Receiver Operating Characteristic) curve.

4.5.3 *Model Configurations*

We compare the results of our model with different feature configurations for linguistic style, stance, and web-source reliability:

- Models using only *language* (LG) features, only *stance* (ST) features, and their combination (LG + ST). These configurations use simple averaging of *per-article* credibility scores to determine the overall credibility of the target claim.
- The aggregation over articles is refined by considering the reliability of the web source where the article was published, considering *language and source reliability* (LG + SR), and *stance and source reliability* (ST + SR).
- Finally, all the aspects *language, stance* and *source reliability* (LG + ST + SR) are considered together.

5. RESULTS

Table 3 shows the performance comparison of the different configurations. We can observe that using only language stylistic features (LG) is not sufficient; it is important to understand the stance (ST) of the article as well. Considering stance along with the language boosts the *Macro-averaged Accuracy* by $\sim 5\%$ points. The full model configuration, i.e., source reliability along with language style and stance features (LG + ST + SR), significantly boosts *Macro-averaged Accuracy* by $\sim 10\%$ points.

Given a claim, our *Stance Classifier* extracts top-ranked snippets from the reporting articles along with their stance (*support* or *refute* probabilities). Combined with the verdict (*true* or *false*) from the *Credibility Classifier*, this yields evidence for the verdict. Table 4 shows examples of our model's output for some claims, along with the verdict and evidence.

6. CONCLUSIONS AND FUTURE WORK

In this article, I presented an outline and current status of my research work which I plan to carry out for my PhD dissertation. This research aims to target the challenging task of assessing credibility of unstructured textual claims on the web which has become a serious problem worldwide. We proposed three factors which can help in solving the challenging task of automatic credibility assessment and presented encouraging results of the preliminary experiments.

Even though the performance of our current approach is very encouraging, the problem of automatic credibility assessment is far from being completely solved and there is a huge scope for improvement. To list a few major limitations of our method,

- The methods of capturing the language style and stance of the article are shallow and do not consider any deeper linguistic aspects.

- Retrieval of the related articles talking about the claim considers only lexical match with the claim text — missing out the related articles not using the same words as the claim text.
- It does not handle the negations very well. For example, for the claim *"Obama was not born in Kenya"*, evidence articles refuting *Kenya* as Obama's birthplace will misguide the system while obtaining their stance as the claim is in negated form. Tackling negations require special attention.
- Web-source reliability values are static, in the sense that after the training process is over, they remain the same. However, in real life scenario, reliability of the web-sources keep changing with time.
- Current approach works well when we have enough data for each claim. But in reality, this is not practical especially when the claims are new – having only few articles talking about it.

Along with addressing above limitations, in future, I would like to also consider following interesting challenges related to my dissertation:

- **Source of attribution:** Apart from the factors we explored so far, I would like to study the impact of attribution on credibility, e.g., text containing attribution as *"The spokesperson confirmed..."* is likely to be more credible than the text containing *"Sources suggest..."* as source of attribution.
- **Time aware analysis:** The idea is to explore if the behavior of change in belief about various claims and how they are discussed over the time has any relation with their credibility.
- **User feedback:** Considering the user feedback into the model learning process can help in making the system more robust and open for continuous improvements.

7. REFERENCES

[1] P. Bellot, A. Doucet, et al. Report on inex 2013. *SIGIR Forum*, 47(2):21–32, Jan. 2013.

[2] M.-A. Cartright, H. A. Feild, and J. Allan. Evidence finding using a collection of books. In *BooksOnline 2011*.

[3] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *WWW 2011*.

[4] X. L. Dong, E. Gabrilovich, et al. Knowledge-based trust: Estimating the trustworthiness of web sources. *Proc. VLDB Endow.*, 8(9):938–949, May 2015.

[5] W. E. Forum. Outlook on the global agenda 2014. `http://reports.weforum.org/outlook-14/`. [Online; accessed 2-Jan-2016].

[6] R. Gingras. Labeling fact-check articles in google news. `https://blog.google/topics/journalism-news/labeling-fact-check-articles-google-news/`. [Online; accessed 2-Jan-2016].

[7] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *WWW Companion 2013*.

[8] E. Kiely. Facebook's 'fake news' initiative. `https://www.factcheck.org/2016/12/facebooks-fake-news-initiative/`. [Online; accessed 2-Jan-2016].

[9] S. Kumar, R. West, and J. Leskovec. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *WWW 2016*.

[10] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang. Prominent features of rumor propagation in online social media. In *ICDM 2013*.

[11] X. Li, X. L. Dong, K. Lyons, W. Meng, and D. Srivastava. Truth finding on the deep web: Is the problem solved? *PVLDB*, 6(2):97–108, 2012.

[12] X. Li, W. Meng, and C. Yu. T-verifier: Verifying truthfulness of fact statements. In *ICDE 2011*.

[13] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han. A survey on truth discovery. *SIGKDD Explorations*, 17(2):1–16, 2015.

[14] Y. Li, Q. Li, J. Gao, L. Su, B. Zhao, W. Fan, and J. Han. On the discovery of evolving truth. In *KDD 2015*.

[15] S. Mukherjee and G. Weikum. Leveraging joint interactions for credibility analysis in news communities. In *CIKM 2015*.

[16] S. Mukherjee, G. Weikum, and C. Danescu-Niculescu-Mizil. People on drugs: Credibility of user statements in health communities. In *KDD 2014*.

[17] N. Nakashole and T. M. Mitchell. Language-aware truth assessment of fact candidates. In *ACL 2014*.

[18] D. Pogue. At snopes.com, rumors are held up to the light. `http://www.nytimes.com/2010/07/15/technology/personaltech/15pogue-email.html`, July 15, 2010. [Online; accessed 2-Jan-2016].

[19] K. Popat, S. Mukherjee, J. Strötgen, and G. Weikum. Credibility assessment of textual claims on the web. In *CIKM 2016*.

[20] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei. Rumor has it: Identifying misinformation in microblogs. In *EMNLP 2011*.

[21] S. Somasundaran and J. Wiebe. Recognizing stances in online debates. In *ACL 2009*.

[22] D. Sridhar, L. Getoor, and M. Walker. Collective stance classification of posts in online debate forums. In *ACL Joint Workshop on Social Dynamics and Personal Attributes in Social Media 2014*.

[23] V. V. Vydiswaran, C. Zhai, and D. Roth. Content-driven trust propagation framework. In *KDD 2011*.

[24] M. A. Walker, P. Anand, R. Abbott, and R. Grant. Stance classification using dialogic properties of persuasion. In *NAACL HLT 2012*.

[25] Wikipedia. Snopes.com. `https://en.wikipedia.org/wiki/Snopes.com`. [Online; accessed 2-Jan-2016].

[26] Q. Xu and H. Zhao. Using deep linguistic features for finding deceptive opinion spam. In *COLING 2012*.

[27] F. Yang, Y. Liu, X. Yu, and M. Yang. Automatic detection of rumor on sina weibo. In *MDS 2012*.

[28] A. Yates, N. Goharian, and O. Frieder. Extracting adverse drug reactions from social media. In *AAAI 2015*.

[29] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. *IEEE Trans. on Knowl. and Data Eng.*, 20(6):796–808, June 2008.

[30] B. Zhao, B. I. P. Rubinstein, J. Gemmell, and J. Han. A bayesian approach to discovering truth from conflicting sources for data integration. *Proc. VLDB Endow.*, 5(6):550–561, Feb. 2012.

[31] S. Zhi, B. Zhao, W. Tong, J. Gao, D. Yu, H. Ji, and J. Han. Modeling truth existence in truth discovery. In *KDD 2015*.