Credibility Assessment of Textual Claims on the Web

Kashyap Popat Subhabrata Mukherjee Jannik Strötgen Gerhard Weikum
Max Planck Institute for Informatics
Saarbrücken, Germany
{kpopat,smukherjee,jstroetge,weikum}@mpi-inf.mpg.de

ABSTRACT

There is an increasing amount of false claims in news, social media, and other web sources. While prior work on truth discovery has focused on the case of checking factual statements, this paper addresses the novel task of assessing the credibility of arbitrary claims made in natural-language text — in an open-domain setting without any assumptions about the structure of the claim, or the community where it is made. Our solution is based on automatically finding sources in news and social media, and feeding these into a distantly supervised classifier for assessing the credibility of a claim (i.e., true or fake). For inference, our method leverages the joint interaction between the language of articles about the claim and the reliability of the underlying web sources. Experiments with claims from the popular website *snopes.com* and from reported cases of Wikipedia hoaxes demonstrate the viability of our methods and their superior accuracy over various baselines.

Keywords

Credibility Analysis; Rumor and Hoax Detection; Text Mining

1. INTRODUCTION

Motivation: With the explosive growth of the Web, online news, and social media, there is also a large amount of false claims. This issue is present in many domains, ranging from fake reviews on product websites, erroneous stock prices, manipulative statements about companies, celebrities, and politicians, all the way to disseminating false news [4, 7]. Determining the credibility of a claim is a challenging task. As reported in [5], 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.

With the increasing number of hoaxes and rumors, fact-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 *fake*) with supporting evidence. The work in this paper aims to replace this manual verification/falsification with an automated system.

State of the Art and its Limitations: Prior work on credibility analysis (see [9] for a survey) has focused on factual claims (e.g., [7, 8, 10]) and/or online communities with specific characteristics like user metadata, who-replied-to-whom, who-edited-what, etc. (e.g., [5, 12]). Truth-finding methods of this kind, starting with the seminal work of [19], assume that claims follow a structured template with clear identification of the questionable values [7, 8], or correspond to subject-predicate-object triples obtained by information extraction [13]. 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.

Novel Problem: The work in this paper aims to overcome these limitations by addressing the case of arbitrary *textual claims* that are 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.

Example: Consider the following claim[1] from the fake news website *thenochill.com*: *"15 Year Old Killed Trespassing While Playing Pokemon Go"*. Our objective is to assess the credibility of this statement as *true* or *fake*. For instance, our model classifies this claim as *fake*. Another example of such a claim is the statement *"I want to share this shocking news: Obama care will require all Americans to be implanted with RFID chips. This chip serves no purpose but a sinister agenda."* which appeared in a social media site a few years ago.

Our Approach: We present a novel approach to identify fake *textual claims*, in an *open-domain* setting, where we do not assume any community-specific characteristics or structure in the input data. Given a claim in the form of a sentence or paragraph, we first use a search engine to identify documents from multiple web-sources, which refer to

[1] http://thenochill.com/teen-killed-while-playing-pokemon-go/

CIKM'16 , October 24-28, 2016, Indianapolis, IN, USA

DOI: http://dx.doi.org/10.1145/2983323.2983661

the claim. We refer to these documents as *reporting articles* in this paper. Then, we analyze the interplay between the *language* (e.g., bias, subjectivity, etc.) of the retrieved articles, and the *reliability* of the web-sources where the articles appeared. Finally, we propose a *Distant Supervision* based classifier which uses these factors to assess the credibility of the claim reported by multiple sources (*cf.* Section 2, 3).

We perform experiments with claims from the fact-checking website *snopes.com* and with data about hoaxes and fictitious persons in Wikipedia. The performance of our approach demonstrates major improvements in accuracy over various baselines (*cf.* Section 4, 5).

2. OVERVIEW OF OUR APPROACH

We capture the following factors that help in determining the credibility of a claim:

i) How is the claim reported? The *writing style* of the articles reporting the claim gives important clues about the credibility of the claim. For example, related work in detecting biased language [17] and credibility analysis in closed communities [12, 11] leverage linguistic features like discourse, subjectivity, and modality.

ii) Who is reporting the claim? The *provenance* of the claim coupled with the *reliability* of the source plays a key role in understanding its credibility. For instance, *theonion.com* is known to publish satirical articles, whereas *wikipedia.org* usually provides objective information according to its *Neutral Point of View* policy.

Consider a set of textual claims $\langle C \rangle$ in the form of sentences or short paragraphs, and a set of web-sources $\langle WS \rangle$ containing articles $\langle A \rangle$ that report on the claims. Let $a_{ij} \in A$ denote an article of web-source $ws_j \in WS$ about claim $c_i \in C$. Each claim c_i is associated with a binary random variable y_i that depicts its credibility label, where $y_i \in \{T, F\}$ (T stands for *True*, whereas F stands for *Fake*). Each article a_{ij} is associated with a random variable y_{ij} that depicts the credibility opinion (*True* or *Fake*) of the article a_{ij} (from ws_j) regarding c_i – when considering only this article. Figure 1 illustrates this model. Given the labels of a subset of the claims (e.g., y_1 for c_1, and y_3 for c_3), our objective is to predict the credibility label of the remaining claims (e.g., y_2 for c_2).

To learn the parameters in our credibility assessment model, we use *Distant Supervision* to attach observed true/fake labels of claims to corresponding reporting articles, and learn a *Credibility Classifier*. In this process, we need to (a) understand the language of the article, and (b) consider the *reliability* of the underlying web sources reporting the articles. Thereafter, we (c) compute the credibility opinion scores of individual articles, and finally, (d) *aggregate* these scores from all articles to obtain the overall credibility label of target claims.

3. CREDIBILITY ASSESSMENT

The following sections describe the features used in our model and how we learn the parameters.

3.1 Language Stylistic Features

The style in which a claim is reported in an article plays a critical role in understanding its credibility. A true claim is assumed to be reported in an objective and unbiased language. On the other hand, if a claim is reported in a highly subjective or a sensationalized style, then it is likely to be less credible. This hypothesis is validated in [13] through an experiment using Amazon Mechanical Turk.

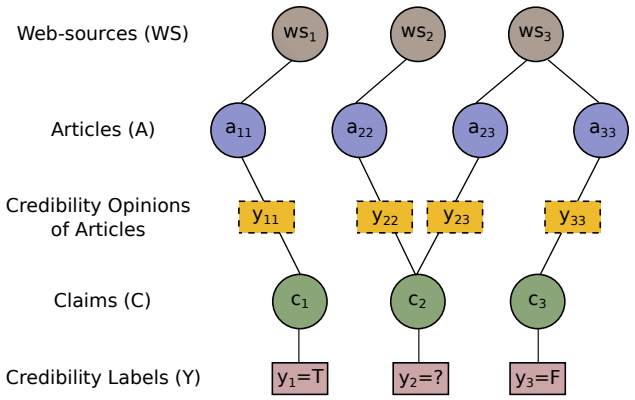

Figure 1: Factors for credibility analysis.

In order to capture the linguistic style of the reporting articles to model the above hypothesis, we use the set of lexicons from [11], 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,
Feature vector construction: For each article a_{ij}, we compute the normalized frequency of all the linguistic features $\langle f_k \rangle$. Given all the stylistic language features, we compute

$$F^L(a_{ij}) = \langle freq^{f_k}_{a_{ij}} = n^{f_k}_{a_{ij}} / length(a_{ij}) \rangle$$

where, $n^{f_k}_{a_{ij}} = number\ of\ times\ f_k\ occur\ in\ a_{ij}$.

3.2 Source Reliability

Apart from the reporting style of the article, 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[2], as opposed to a claim on the "World Health Organization" website.

To capture the reliability of the web-source for each web article, we determine the AlexaRank and PageRank of its source and use them as proxies for the source reliability. AlexaRank [3] is based on a combined measure of unique visitors and page views of the website. PageRank determines importance of the website by counting the number and qual-

[2]A satire, spoof, parody and humour portal: http://www.theunrealtimes.com/
[3]https://support.alexa.com/hc/en-us/articles/200449744-How-are-Alexa-s-traffic-rankings-determined-

Type of Feature	Number of Features
Linguistic	
Assertive Verbs	66
Factive Verbs	27
Hedges	100
Implicatives	32
Report Verbs	181
Discourse Markers	13
Subjectivity and Bias	8770
Reliability	
Source Identity	#web-sources
PageRank	1
AlexaRank	1

Table 1: Statistics of features used in our model.

ity of links to and from the website. To avoid modeling from sparse observations, we combine all the web-sources having less than 10 articles in the dataset to a single web-source.

Feature vector construction: For each article a_{ij}, we capture the identity of its web-source ws_j using a one-hot vector of dimension $cardinality(< WS >)$ (after collapsing the "long-tail" sources to a single source) by setting the j^{th} element in the vector to 1, and the remaining ones to 0. We also use the AlexaRank and PageRank of the web-source as additional features capturing the source reliability.

$F^{SR}(a_{ij}) = \langle 0 \ldots, ws_j = 1, 0 \ldots, \log PR_{ws_j}, \log AR_{ws_j} \rangle$ where, PR and AR represent the PageRank, and the AlexaRank, respectively.

3.3 Credibility Classification using Distant Supervision

Credibility labels are available *per-claim*, and not per-reporting-article. Thus, in our approach for credibility aggregation from multiple sources, we use *Distant Supervision* for *training* — whereby we attach the (observed) label y_i of each claim c_i to each article a_{ij} reporting the claim (i.e., setting labels $y_{ij} = y_i$). For instance, in Figure 1, $y_{11} = y_1 = T, y_{33} = y_3 = F$. Using these $\langle y_{ij} \rangle$ as the corresponding training labels for $\langle a_{ij} \rangle$, with the corresponding feature vectors $\langle F^L(a_{ij}) \cup F^{SR}(a_{ij}) \rangle$, we train an L_1-regularized logistic regression model on the training data. Statistics of features used in our model are given in Table 1.

For any *test* claim c_i whose credibility label is unknown, and its corresponding reporting articles $\langle a_{ij} \rangle$, we use this *Credibility Classifier* to obtain the corresponding credibility opinions $\langle y_{ij} \rangle$ of the articles. We determine the overall credibility label y_i of c_i by considering a sum of *per-article* credibility probabilities:

$$y_i = \underset{l \in \{T,F\}}{\arg\max} \sum_{a_{ij}} Prob(y_{ij} = l) \quad (1)$$

4. CASE STUDIES

4.1 Snopes

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. It is a well-known resource for validating and debunking such stories, receiving around 300,000 visits a day [15]. They typically collect rumors and claims from *Facebook*, *Twitter*, *Reddit*, news websites, e-mails by users, etc.

Total claims	4856
True claims	1277 (26.3%)
Fake claims	3579 (73.7%)
Web articles	133272
Avg. articles per claim	27.44

Table 2: *Snopes* data statistics.

Each article verifies a single claim, e.g., *"North Carolina no longer considers the $20 bill to be legal tender"*. The Snopes editors assign a *manual* credibility verdict to each such claim: *True* or *False*. 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. Claims having labels like *Partially True* or *Partially False* are ignored. 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 etc.), an *Origin* section describing the origin of the claim, and an *Analysis* section justifying the verdict. Our model is agnostic of the structure of Snopes as we use only the claim and its credibility verdict, ignoring all other related information.

We collected data from *Snopes* published until February 2016. For each claim c_i, we fired the *claim text* as a *query* to the Google search engine and extracted the first three result pages (i.e., up to 30 articles) as a set of reporting articles $\langle a_{ij} \rangle$. We ignore the ranking information in the set of collected articles to have minimal dependency on the search engine. Other search engines, or other means of evidence gathering can easily be used. We then crawled all these articles from their corresponding web-sources $\langle ws_j \rangle$. 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.

4.2 Wikipedia

We collected a set of 100 proven hoaxes reported on Wikipedia[4], e.g., *"Alien autopsy film by Ray Santilli"*, *"Disappearing blonde gene"* etc. All these hoaxes can be mapped to claims of types: "*<ENTITY> exists*", "*<ENTITY> is genuine*" or "*<EVENT> occurred*". While collecting the data, hoaxes not falling under these categories were ignored. Words related to hoaxes, e.g., *fake*, *fictional*, *nonexistent*, etc., were removed from the claim description to avoid any kind of search bias while retrieving articles using a search engine. Since the dataset contains only hoaxes, the ground-truth label for all of these claims is *Fake*.

In addition, we also collected a set of 57 fictitious people as reported on the Wikipedia page[5], e.g., *"Ern Malley, an Australian poet"*, *"P. D. Q. Bach, a composer"* etc. All these entities can be mapped to claims of type: "*<ENTITY> exists*". The ground-truth label for all of these claims is *Fake* as the dataset contains only fictitious people.

[4]https://en.wikipedia.org/wiki/List_of_hoaxes#Proven_hoaxes

[5]https://en.wikipedia.org/wiki/List_of_fictitious_people

	Hoaxes	Fictitious People
Total Claims	100	57
Web articles	2813	1552
Avg. articles per claim	28.13	27.22

Table 3: Wikipedia data statistics.

Table 3 reports the statistics of the dataset. As described earlier, we used a search engine to get a set of reporting articles for these claims. Similar to the previous case, we removed results from the *wikipedia.org* domain. Note that we trained our *Credibility Classifier* on *Snopes* data, and tested it on this data from *Wikipedia* — thereby demonstrating that our model generalizes and can be easily applied to data from other domains.

5. EXPERIMENTS

We conducted a set of experiments using data from *Snopes* and *Wikipedia* to test the performance of our methods.

Evaluation Measures: We train our models with *Snopes* data, and report standard 10-fold cross-validation accuracy on all datasets. *Snopes*, primarily being a hoax debunking website, is biased towards (refuting) the *Fake* claims. Therefore, we also 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. To highlight the effectiveness of our model in identifying fake claims (i.e., hoaxes, rumors etc.), we also report the precision, recall and F1 score for the *Fake* claim class.

5.1 Credibility Assessment: *Snopes*

While performing 10-fold cross-validation on the claims, we trained on any 9-folds of the data — where the algorithm learned the *Credibility Classifier* and web-source reliabilities from the reporting articles and their corresponding sources present in the training split. 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 adjust the classifier's loss function. We place a large penalty for misclassifying instances from the *true* class which boosts certain features from that class. The overall effect is that the classifier makes fewer mistakes for *true* instances, leading to balanced classification. We set the penalty for the *true* class to 2.8 — given by the ratio of the number of *fake* claims to *true* claims in the *Snopes* data.

We compare to the following baselines:

ZeroR[6]**:** This is a trivial baseline, designed for imbalanced data, that always labels a claim as the class with the largest proportion, i.e., *fake* in our case. The overall accuracy of this baseline is **73.69%**, and the macro-averaged accuracy is **50%**.

FactChecker: Recent work on fact checking [13] relies on the hypothesis that claims reported by objective articles are more likely to be true than those reported in subjective articles. The authors extracted *alternative* fact candidates for the given claim, and used the hypothesis to rank all candidates. This approach works well in their use case of knowledge base curation, as all the claims are factual and have the form of Subject-Predicate-Object (SPO) triples. On the other hand, the claims in our case are textual snippets without any explicit alternative candidates. Therefore, we could only implement this method as a baseline "in spirit". To this end, we used the code[7] of [11] to construct an "Objectivity Detector". Given a claim and a set of reporting articles, the target claim was labeled *true* if the sum of the objectivity scores of its reporting articles — as determined by the Objectivity Detector — was higher than the sum of the subjective scores, and *fake* otherwise. This approach resulted in **55.29%** overall accuracy and **56.27%** macro-averaged accuracy for credibility classification.

Along with the above baselines, we also report the results of our model with different feature configurations for linguistic style and web-source reliability:

- Model using only *language* (LG) features,
- Model using only *web-source reliability* (SR) features,
- Aggregated model with the combination of, *language* and *source reliability* (LG + SR) features.

Table 4 shows the 10-fold cross-validation accuracy of various baselines against different configurations of our model, with the ROC curves plotted in Figure 2. From the results, we observe that using only language stylistic features (LG) is not sufficient; it is important to understand the source reliability (SR) of the article as well. High precision score for the *Fake* claim class shows the strength of our model in detecting *Fake* claims.

5.2 Credibility Assessment: *Wikipedia*

To demonstrate the generality of our approach, the model trained on the *Snopes* dataset was tested on the *Wikipedia* dataset of hoaxes and fictitious persons. The results are shown in Table 5. Similar to the *Snopes* setting, we removed all references to Wikipedia from the data and the search engine results. As we can see from the results, our system is able to detect hoaxes and fictitious people with high accuracy, although the claim descriptions here are stylistically quite different from those of *Snopes*.

6. ERROR ANALYSIS AND DISCUSSION

Poor performance on detecting fake claims: As we see from the results, the system accuracy for detecting fake claims is low compared to that for the true claims. While performing an error analysis of the results, we observed that many of the well written articles from reputed web-sources refer to the fake claims in *negated* form such as "... the company's spokesperson *denied* that ...". Our model does not capture these finer linguistic aspects like implicit or explicit negation, and, therefore, commits mistakes. In future, we would like to propose features which capture these finer semantics of the article text so that we can have a more accurate system.

[6]https://weka.wikispaces.com/ZeroR

[7]Code and data available from: http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/impact/credibilityanalysis/

Configuration	Overall Accuracy (%)	*True* Claims Accuracy (%)	*Fake* Claims Accuracy (%)	Macro-averaged Accuracy (%)	AUC	*Fake* Claims Precision	*Fake* Claims Recall	*Fake* Claims F1-Score
LG + SR	71.96	75.43	70.77	**73.10**	**0.80**	0.89	0.71	0.79
LG	69.43	66.47	70.55	68.51	0.75	0.85	0.71	0.77
SR	66.52	68.56	65.90	67.23	0.73	0.85	0.66	0.74
FactChecking	55.29	58.34	54.21	56.27	0.58	0.78	0.54	0.64
ZeroR	73.69	00.00	100	50.00	0.50	0.74	1.00	0.85

Table 4: Performance comparison of our model vs. related baselines with 10-fold cross-validation on Snopes data. LG: language stylistic features, SR: web-source reliability.

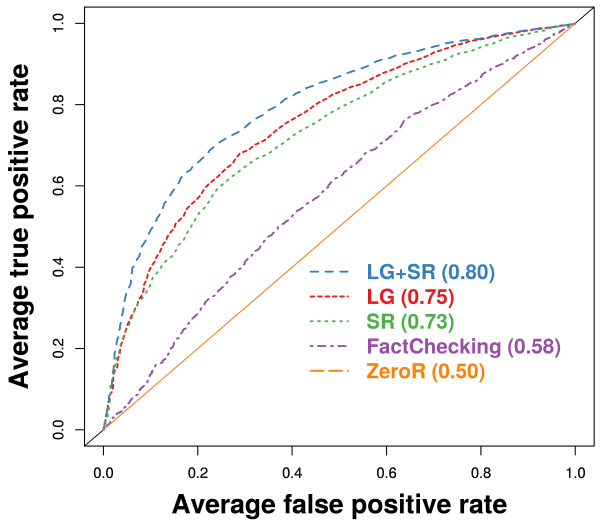

Figure 2: ROC curves for different model configurations.

Marginal contribution of web-source reliability: Results also indicate that the performance of the full model configuration (LG+SR) achieves only slight improvement over the configuration LG. This can be attributed to the fact that these rank measures (PageRank and AlexaRank) capture the authority and popularity of the web-sources, but not their reliability from the credibility point of view. For example, the PageRank of the satirical news website *The Onion* is very high (7 out of 10). However, this does not indicate anything about its reliability. Hence, as future work, it would be interesting to design an algorithm which automatically captures the ranking of web-sources based on their credibility.

Understanding the credibility assessment output: While performing error analysis, we observed that the probability scores do not help in understanding the output. This is also true for related truth finding approaches. It would thus be nice to have interpretable evidence as an additional output of the system which can explain the credibility assessment. Table 6 gives a snapshot of claims with the credibility assessment given by our system, along with manual annotation of snippets that can be used as evidence. As future work, we want to automate this process of generating evidence.

Test Data	#Claims	Accuracy (%)
Wiki Hoaxes	100	84.00
Wiki Fictitious People	57	66.07

Table 5: Accuracy of credibility classification on *Wikipedia* data.

7. RELATED WORK

Our work draws motivation from the following areas:

Truth discovery: In approaches to truth discovery [2, 3, 7, 10, 14, 19], the goal is to resolve conflicts in multi-source data. Input data is assumed to have a structured representation: an entity of interest along with its potential values provided by different sources. It is assumed that the conflicting values are already available.

Work in [13] goes a step further by proposing a method to generate conflicting *values* or *fact candidates* from Web contents. 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. The method proposed in [8] supports credibility assessment of statements but it relies on the *user* providing the *doubtful* portion of the input statement.

All the above approaches are limited to resolving conflicts amongst alternative fact candidates (or, multi-source data) in structured datasets. In our work, we address these limitations and propose a general approach to process unstructured natural-language claims without requiring any alternative claims.

Credibility analysis within communities and social media: An approach for credibility analysis within online health communities is proposed in [12], based on 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 [11]. Wikipedia hoaxes are studied in [5].

Prior research on credibility assessment of social media posts exploits *community-specific* features for detecting rumors, fake, and deceptive content [1, 16, 18]. Temporal, structural, and linguistic features were used to detect rumors on Twitter in [6]. Detecting fake images in Twitter based on influence patterns and social reputation is addressed in [4].

Claim	Verdict & Evidence
A woman stabbed her boyfriend with a sharpened selfie stick because he didn't like her newest Instagram selfie quickly enough.	**[Verdict]:** False **[Evidence]:** A weird kind of story in heavy circulation online states ... No, the claim is not a fact.
90% of people in the U.S. marry their high school sweethearts.	**[Verdict]:** False **[Evidence]:** The school category resulted in only 14% of total respondent base. In analyzing these surveys, one must realize that potential biases in survey methods exist, such as ... It seems absolutely clear that these and other surveys conducted in early 1990s represent nowhere nearly close to 90% ...
A Facebook coupon offering 50% off at Target stores is real.	**[Verdict]:** False **[Evidence]:** The newest questionable offer to take hold of Facebook newsfeeds involves the false promise of a coupon ... A rep for Target HQ confirms to Consumerist that there is no such coupon and this is a fake.
Two Maryland sheriff's deputies were fatally shot and a suspect killed on Wednesday in a shootout at a Baltimore-area Panera restaurant.	**[Verdict]:** True **[Evidence]:** Two Maryland sheriff's deputies were fatally shot and a suspect killed Wednesday in a shootout at a Baltimore-area Panera restaurant filled with lunchtime customers. (Reuters) Authorities found a semiautomatic handgun in Evans's vehicle, which he might have been living in.
A dying child was made an honorary fireman by the Phoenix Fire Department.	**[Verdict]:** True **[Evidence]:** We'll make him an honorary Fireman for the day. He can come down to the fire station, eat with us, go out on all the fire calls, the whole nine yards! The Fire Chief decided that the Phoenix Fire Department should make sure the dying boy had an experience truly befitting a fireman.
A declared-dead jockey returned to the track and shocked the grandstand crowd.	**[Verdict]:** True **[Evidence]:** When the crowd realized that the shirtless, bloodied, toe-tagged man who was staggering across the grandstand area was the jockey who had been declared dead about a half hour earlier, the crowd and the race officials rushed towards Neves, as shock turned to celebration.

Table 6: Snapshot of claims with assessment from Credibility Classifier, and manually annotated snippets as evidence.

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.

8. CONCLUSIONS

In this paper, we proposed a general approach for credibility analysis of unstructured textual claims in an open-domain setting. We make use of the language style and source reliability of articles reporting the claim to assess its credibility. Experiments on analyzing the credibility of real-world claims, from the fact-checking website *Snopes*, and on hoaxes and fictitious persons listed on *Wikipedia*, demonstrate the effectiveness of our approach. As future work, we want to investigate the role of attribution or speaker information, refined linguistic aspects like negation, and understanding the article's perspective about the claim.

9. REFERENCES

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

[2] X. L. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: The role of source dependence. *Proc. VLDB Endow.*, 2(1):550–561, 2009.

[3] A. Galland, S. Abiteboul, A. Marian, and P. Senellart. Corroborating information from disagreeing views. In *WSDM 2010*.

[4] 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*.

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

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

[7] 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.

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

[9] 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.

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

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

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

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

[14] J. Pasternack and D. Roth. Knowing what to believe (when you already know something). In *COLING 2010*.

[15] 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 26-Apr-2016].

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

[17] M. Recasens, C. Danescu-Niculescu-Mizil, and D. Jurafsky. Linguistic models for analyzing and detecting biased language. In *ACL 2013*.

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

[19] 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.