CredEye: A Credibility Lens for Analyzing and Explaining Misinformation

Kashyap Popat
Max Planck Institute for Informatics
Saarbrücken, Germany
kpopat@mpi-inf.mpg.de

Subhabrata Mukherjee*
Amazon Inc.
Seattle, USA
subhomj@amazon.com

Jannik Strötgen
Max Planck Institute for Informatics
Saarbrücken, Germany
jstroetge@mpi-inf.mpg.de

Gerhard Weikum
Max Planck Institute for Informatics
Saarbrücken, Germany
weikum@mpi-inf.mpg.de

ABSTRACT

Rapid increase of misinformation online has emerged as one of the biggest challenges in this post-truth era. This has given rise to many fact-checking websites that manually assess doubtful claims. However, the speed and scale at which misinformation spreads in online media inherently limits manual verification. Hence, the problem of automatic credibility assessment has attracted great attention. In this work, we present CredEye, a system for automatic credibility assessment. It takes a natural language claim as input from the user and automatically analyzes its credibility by considering relevant articles from the Web. Our system captures joint interaction between language style of articles, their stance towards a claim and the trustworthiness of the sources. In addition, extraction of supporting evidence in the form of enriched snippets makes the verdicts of CredEye transparent and interpretable.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; *Machine learning*; • **Information systems** → World Wide Web;

KEYWORDS

Fact Checking; Credibility Analysis; Interpretable Learning

ACM Reference Format:
Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2018. CredEye: A Credibility Lens for Analyzing and Explaining Misinformation. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3184558.3186967

1 INTRODUCTION

The explosive growth of the Web, online news and social media has led to a proliferation of misinformation. These range from posting fake reviews in e-commerce portals, propagating rumors and hoaxes in social networks to erroneous quoting of celebrities and politicians. Misinformation can have disastrous consequences: for example, rumors during hurricane Irma forced the US government to start a rumor control website[1] to avoid panic.

State-of-the-art and its Limitations: Prior works in credibility analysis and truth-finding primarily focus on structured data, typically in the form of subject-predicate-object statements [2, 5, 7, 8, 12]. Works on detecting fake statements in social media [1, 4, 6, 11] leverage social network metadata like user-user interactions and social links as well as profiles, reputation features based on votes or likes, and demographic information. Most importantly, all these prior approaches provide black-box techniques and lack the ability to *explain* why a certain statement is classified as true or false.

In our own prior work [9, 10], we address these limitations by considering user-provided *natural language claims*, and develop a general framework which does not make any assumptions about the structure of the claim or characteristics of the community or website where the claim is reported. Our method is based on distantly supervised learning with joint inference over the language style of relevant articles, their stance towards the claim (support or refute), and the trustworthiness of the underlying Web sources. In addition to the automatic assessment, our method extracts interpretable evidence and identifies crucial features to explain its verdicts (see Figure 2, discussed later). Two recent systems along similar lines are ClaimBuster [3] and ClaimVerif [13]. However, neither of these consider the language style of the articles that serve as evidence or counter-evidence. Also, neither provides feature-level explanations of their assessment scores; rather they merely list online articles related to the claim.

Contributions: In this paper, we demonstrate CredEye, an automatic credibility analyzer based on our prior work. Its unique point is that it considers language style as a key component of its assessments, and also provides explanations in terms of automatically extracted snippets from supporting and refuting articles enriched with language features.

Given an input claim in arbitrary textual form on an arbitrary topic, CredEye automatically retrieves relevant articles from the Web, using a search engine. It analyzes the credibility of each text

*Work done at the Max Planck Institute for Informatics prior to joining Amazon.

[1] https://www.fema.gov/hurricane-irma-rumor-control

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3186967

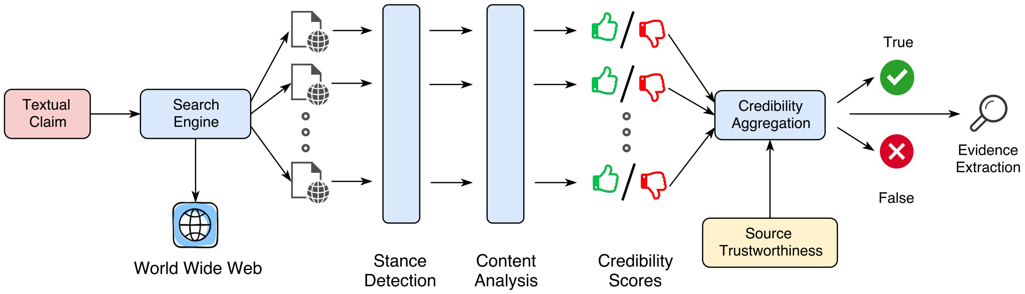


Figure 1: Credibility analysis pipeline of CredEye.

Method	True-Claims Accuracy (%)	False-Claims Accuracy (%)	Macro-Avg. Accuracy (%)
Pipeline	83.20	80.78	82.00
CRF	71.26	88.74	80.00
LSTM	77.90	78.27	78.09

Table 1: Different configurations of CredEye.

by language features, the stance of the text, and the trustworthiness of the source, aggregating all these into an overall verdict. The UI of CredEye (see Figure 2) enables users to dissect and drill down into the assessment by browsing through judiciously and automatically selected snippets with markup of indicative words. The latter capture linguistic features that express bias and subjectivity (decreasing credibility) or neutral and objective language (increasing credibility). Details of the analysis are shown in the form of per-article and per-source scores. CredEye is available at https://gate.d5.mpi-inf.mpg.de/credeye/.

2 CREDIBILITY ASSESSMENT PIPELINE

CredEye takes a *natural language claim* as input from the user, and computes its credibility assessment along with enriched evidence as output. Its core is the analysis of the credibility of the claim, based on the overall evidence or counter-evidence from a set of automatically retrieved Web articles. We have developed three methods to this end: a pipeline of classifiers and scoring models, a joint-inference model in the form of a Conditional Random Field, and a deep-learning neural network based on a bidirectional LSTM. In our experiments (see below) – with limited training data – the pipeline architecture performed best. Hence, we focus on this configuration. Note that the scarceness of training samples is typical in coping with misinformation, not just a limitation of our experiments.

Figure 1 gives an overview of the system architecture. The pipeline consists of the following stages: (i) *Retrieval* of articles from diverse Web sources by sending the claim text to a search engine, (ii) *Stance Detection* to understand the stance of each article, (iii) *Content Analysis* to understand the credibility of each article by utilizing the language style and stance-related features, (iv) *Credibility Aggregation* to merge these per-article assessments to compute the overall scoring of the claim being *true* or *false*, and (v) *Evidence Extraction* to extract supporting evidence in the form of informative snippets from the relevant web articles.

The classifiers are trained by distant supervision using data from snopes.com, a popular fact-checking website that *manually* validates Internet rumors, hoaxes, urban legends, and other stories of unknown or questionable origin. We used 5,000 claims from Snopes, each labeled true or false, and retrieved 30 relevant Web articles for each of them. By assuming that the unlabeled Web articles should predominantly inherit the claim's label (hence *distant* supervision), we could train logistic-regression classifiers for per-article stance and per-article credibility. Table 1 shows accuracy results for the Snopes data, using 10-fold cross-validation.[2]

2.1 Querying the Web

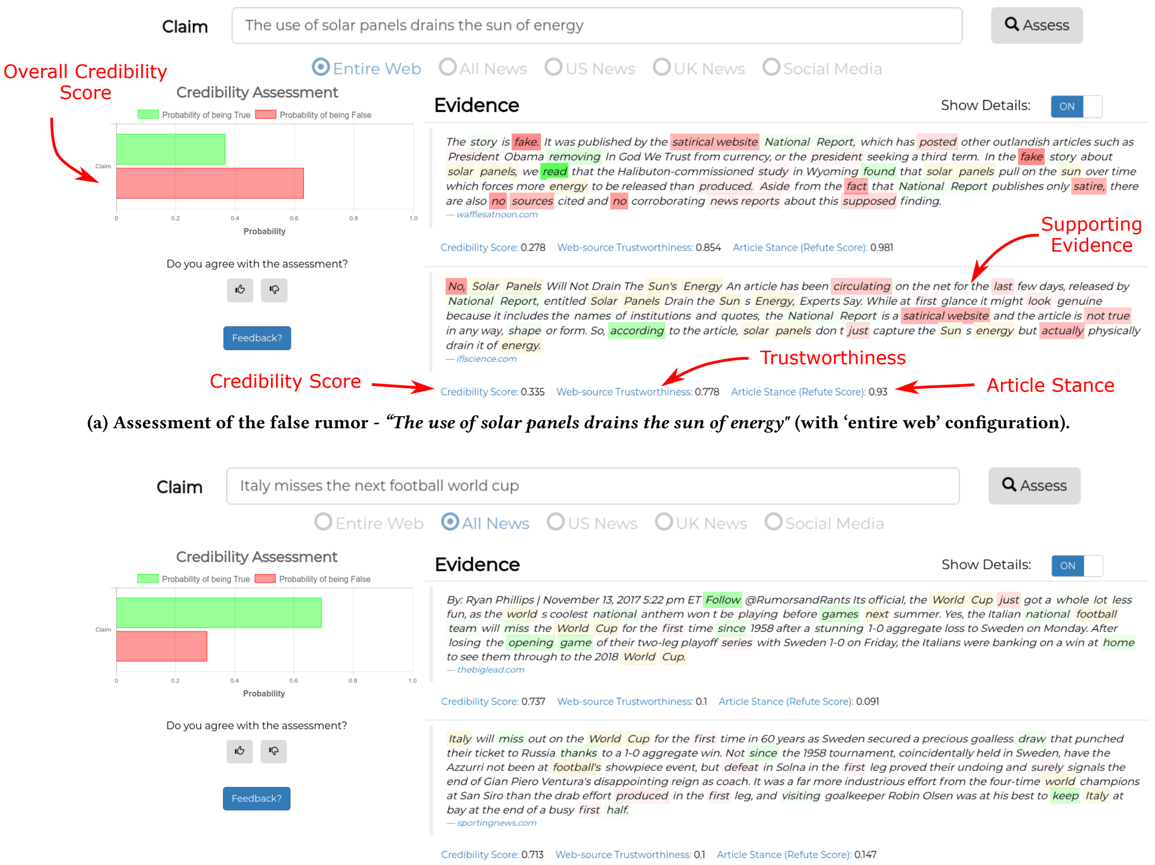
To extract Web articles relevant to the input claim, we use the Bing search API, which allows us to restrict results to specific types (e.g., entire Web, only news, only social media etc.) and geo locations. Our system supports five such configurations for selecting articles from: (i) the entire web (no restrictions), (ii) all news websites, (iii) popular US news websites, (iv) popular UK news websites, and (v) social media websites (like Quora, Twitter, Facebook, blogs etc.). For this demo, we focus on English language articles, without further restrictions.

Knowledge Base Lookup: Before moving to the next stage of the pipeline, we determine if the credibility of an input claim can be easily assessed by a Knowledge Base (KB) lookup. To this end, we first check if a representative <*subject, verb, object*> triplet could be extracted from the input claim. If yes, we query for the corresponding "subject+verb" and "object+verb", and check if the claim can be assessed from the retrieved instant answer. For instance, given the claim *"Obama was born in Kenya"*, the system queries for "obama+born" in Bing, and assesses the claim as false based on the retrieved instant answer. Instead of relying on Bing's internal KB, it is also possible to use any other KB for this lookup.

2.2 Stance Detection

False claims are refuted by articles from trusted Web sources. Therefore, it is necessary to understand an article's stance towards the claim. To this end, we divide each retrieved article into a set of overlapping snippets, and extract snippets that are strongly related to the claim in terms of unigram and bigram overlap. We use the qualifying snippets to compute support and refute scores, using logistic regression classifiers trained on claims and evidence articles from Snopes. The scores are fed as features into the subsequent content analysis.

[2] Data available at http://bit.ly/web-credibility-analysis



Claim: The use of solar panels drains the sun of energy

● Entire Web ○ All News ○ US News ○ UK News ○ Social Media

Show Details: ON

The story is fake. It was published by the satirical website National Report, which has posted other outlandish articles such as President Obama removing In God We Trust from currency, or the president seeking a third term. In the fake story about solar panels, we read that the Halibuton-commissioned study in Wyoming found that solar panels pull on the sun over time which forces more energy to be released than produced. Aside from the fact that National Report publishes only satire, there are also no sources cited and no corroborating news reports about this supposed finding.
— wafflesatnoon.com

Credibility Score: 0.278 Web-source Trustworthiness: 0.854 Article Stance (Refute Score): 0.981

No, Solar Panels Will Not Drain The Sun's Energy An article has been circulating on the net for the last few days, released by National Report, entitled Solar Panels Drain the Sun s Energy, Experts Say. While at first glance it might look genuine because it includes the names of institutions and quotes, the National Report is a satirical website and the article is not true in any way, shape or form. So, according to the article, solar panels don t just capture the Sun s energy but actually physically drain it of energy.
— iflscience.com

Credibility Score: 0.335 Web-source Trustworthiness: 0.778 Article Stance (Refute Score): 0.93

Overall Credibility Score

Supporting Evidence

Trustworthiness

Credibility Score

Article Stance

Assess
Credibility Assessment
🟩 Probability of being True
🟥 Probability of being False
Claim
0
0.2
0.4
0.6
0.8
1.0
Probability
Evidence
Do you agree with the assessment?
Feedback?

(a) Assessment of the false rumor - *"The use of solar panels drains the sun of energy"* (with 'entire web' configuration).

Claim: Italy misses the next football world cup

○ Entire Web ● All News ○ US News ○ UK News ○ Social Media

Show Details: ON

By: Ryan Phillips | November 13, 2017 5:22 pm ET Follow @RumorsandRants Its official, the World Cup just got a whole lot less fun, as the world s coolest national anthem won t be playing before games next summer. Yes, the Italian national football team will miss the World Cup for the first time since 1958 after a stunning 1-0 aggregate loss to Sweden on Monday. After losing the opening game of their two-leg playoff series with Sweden 1-0 on Friday, the Italians were banking on a win at home to see them through to the 2018 World Cup.
— thebiglead.com

Credibility Score: 0.737 Web-source Trustworthiness: 0.1 Article Stance (Refute Score): 0.091

Italy will miss out on the World Cup for the first time in 60 years as Sweden secured a precious goalless draw that punched their ticket to Russia thanks to a 1-0 aggregate win. Not since the 1958 tournament, coincidentally held in Sweden, have the Azzurri not been at football's showpiece event, but defeat in Solna in the first leg proved their undoing and surely signals the end of Gian Piero Ventura's disappointing reign as coach. It was a far more industrious effort from the four-time world champions at San Siro than the drab effort produced in the first leg, and visiting goalkeeper Robin Olsen was at his best to keep Italy at bay at the end of a busy first half.
— sportingnews.com

Credibility Score: 0.713 Web-source Trustworthiness: 0.1 Article Stance (Refute Score): 0.147

Assess
Credibility Assessment
🟩 Probability of being True
🟥 Probability of being False
Claim
0
0.2
0.4
0.6
0.8
1.0
Probability
Evidence
Do you agree with the assessment?
Feedback?

(b) Assessment of the true statement - *"Italy misses the next football world cup"* (with 'all news' configuration).

Figure 2: CredEye interface.

2.3 Content Analysis

The content analysis of the articles is the core part and distinguishing characteristic of CredEye. It assesses the credibility of each article based on a suite of linguistic features (see [10] for more details).

Features: Our hypothesis is that true and thus credible claims are reported in an objective and unbiased language. On the other hand, subjective or sensational style of reporting a claim decreases its credibility. To capture the language style of the article, we derive features from a predefined set of lexicons (e.g., assertive and factive verbs, hedges, report verbs, subjective and biased words etc.). In addition, the support and refute scores from the stance detection step are used as features.

Classifier: The credibility assessment model is a logistic regression classifier with L1-regularization, distantly trained on Snopes samples.

2.4 Credibility Aggregation

Not all Web sources are trustworthy. Hence, to aggregate per-article credibility scores, it is essential to determine the trustworthiness of each article's source.

Source Trustworthiness: Computing the trustworthiness of a source hinges on the following hypothesis: a Web source is trustworthy if it *refutes* non-credible claims and *supports* credible ones. We calculate the trustworthiness $tw(s)$ of source s as :

$$tw(s) = \frac{\#articles\_support\_true + \#articles\_refute\_false}{\#total\_articles} \quad (1)$$

where, $\#articles\_support\_true$ is the number of articles from s that support credible claims, $\#articles\_refute\_false$ represents the number of articles from s that refute non-credible claims, and $\#total\_articles$ is the total number of articles from s. We use the Snopes training data to pre-compute these trustworthiness scores for a wide variety of sources, including news sites, online communities, Wikipedia, and more. When we encounter a new source which is not present in our training data, we assign a default trustworthiness score of 0.1 (as used in our experiments).

Claim Credibility: Given a claim c and a set of relevant articles $\{a_i\}$ from sources $\{s_i\}$, we aggregate the per-article credibility scores as:

$$P(c = credible) = \frac{\sum_i tw(s_i) * p_{a_i}(c = credible)}{\sum_i tw(s_i)} \quad (2)$$

Here, $P(c = credible)$ denotes the aggregated score for the claim being credible, $p_{a_i}(c = credible)$ is the credibility score of a_i, and $tw(s_i)$ is the trustworthiness of s_i. This aggregation penalizes the credibility scores from non-trustworthy sources.

2.5 Evidence Extraction

To present users with comprehensible evidence for credibility verdicts, we utilize the snippets of articles extracted in the stance detection step. From each article, CredEye selects the snippet that is most related to the claim and has a support or refute score that is above a threshold and agrees with the overall verdict.

In addition, CredEye enriches the presented snippets by highlighting salient words and bigrams. Words that are also present in the claim are highlighted in *yellow*. Words which contribute most towards the aggregated credibility score are highlighted in different shades of *green* (signaling credibility) and *red* (signaling non-credibility). The intensity of colors reflects the words' importance for the assessment (based on feature weights from the classifier). The highlighted words and bigrams are judiciously selected from the features of the stance detection step, and also from various lexicons of subjective and emotional language (e.g., OpinionFinder MPQA).

3 DEMONSTRATION

CredEye can be accessed at https://gate.d5.mpi-inf.mpg.de/credeye/ (a recorded screencast available at https://youtu.be/t0SKDjovJiU). We encourage readers to try it out with their own inputs; we also offer some sample claims for illustration. Here, we consider two scenarios: (i) a false rumor *"The use of solar panels drains the sun of energy"* with 'entire web' configuration (see Figure 2a) and (ii) a true statement *"Italy misses the next football world cup"* with 'all news' configuration (see Figure 2b).

As shown in Figure 2, the *input area* of CredEye contains a text box where the user can enter any natural language text as an input claim for assessment along with a specific configuration to restrict the article sources. Upon submitting the claim, the back-end server of CredEye carries out its analysis and returns its verdict along with evidence snippets, displayed in the *output area*. The output includes the overall assessment, displayed in the form of green (true) and red (false) bars. There are also buttons for providing feedback.

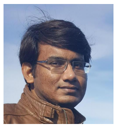
The most interesting part of the output is the explanation of the assessment, in the form of enriched text snippets from the Web articles that were retrieved during the analysis. As shown in Figure 2, salient words in the snippets are highlighted in different colors (see Section 2.5). Phrases present in the articles like *"fake"*, *"satirical website"*, *"supposed"* etc. in Figure 2a reduce the credibility of the claim which helps our credibility assessment pipeline to classify it as false. On the other hand, absence of biased and subjective words (decreasing credibility) in addition to objective words like *"follow"*, *"keep"*, *"games"*, etc. in Figure 2b increase the credibility of the claim. Hence, our pipeline assesses this factual statement as credible. In addition, CredEye shows the sub-scores from the various stages of its pipeline: the per-article credibility score, the refute score from the stance detection, and the trustworthiness of the source.

4 CONCLUSION

The CredEye system is a step towards coping with misinformation. One of its limitations is the lack of in-depth understanding of the exact scope and finer tone of claims. For instance, in a claim like "the US Civil War ended slavery world-wide" – it is challenging for the system to understand its finer scope 'world-wide'. Retrieving sufficient evidence or counter-evidence is another bottleneck where we hinge on search-engine results.

5 CREATORS

Kashyap Popat is a PhD candidate at the Max Planck Institute for Informatics. The focus of his research is on analyzing and explaining credibility of textual content. His research interests span text mining, natural language processing and deep learning.



Subhabrata Mukherjee is a scientist at Amazon building its product knowledge graph. He obtained his PhD from the Max Planck Institute for Informatics. His research interests span graphical models, deep learning, information extraction and recommender systems.

Jannik Strötgen is a senior researcher at the Max Planck Institute for Informatics. He is the lead researcher of the multilingual, domain-sensitive tool HeidelTime and his research interests are in natural language processing and information retrieval.

Gerhard Weikum is a scientific director at the Max Planck Institute for Informatics. His research spans transactional and distributed systems, self-tuning database systems, data and text integration, and the automatic construction of knowledge bases.

REFERENCES

[1] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information Credibility on Twitter. In *WWW 2011*.

[2] Xin Luna Dong, Evgeniy Gabrilovich, et al. 2015. Knowledge-based Trust: Estimating the Trustworthiness of Web Sources. *PVLDB 2015* (2015).

[3] Naeemul Hassan et al. 2017. ClaimBuster: The First-ever End-to-end Fact-checking System. *PVLDB 2017* (2017).

[4] Srijan Kumar, Robert West, and Jure Leskovec. 2016. Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes. In *WWW 2016*.

[5] Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiaweix Han. 2015. A Survey on Truth Discovery. *SIGKDD Explorations 17(2), 2015* (2015).

[6] Subhabrata Mukherjee, Gerhard Weikum, and Cristian Danescu-Niculescu-Mizil. 2014. People on Drugs: Credibility of User Statements in Health Communities. In *KDD 2014*.

[7] Ndapandula Nakashole and Tom M. Mitchell. 2014. Language-Aware Truth Assessment of Fact Candidates. In *ACL 2014*.

[8] Jeff Pasternack and Dan Roth. 2013. Latent Credibility Analysis. In *WWW 2013*.

[9] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. Credibility Assessment of Textual Claims on the Web. In *CIKM 2016*.

[10] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the Truth Lies: Explaining the Credibility of Emerging Claims on the Web and Social Media. In *WWW 2017*.

[11] Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying Misinformation in Microblogs. In *EMNLP 2011*.

[12] Xiaoxin Yin, Jiawei Han, and Philip S. Yu. 2008. Truth Discovery with Multiple Conflicting Information Providers on the Web. *TKDE 20(6), 2008* (2008).

[13] Shi Zhi et al. 2017. ClaimVerif: A Real-time Claim Verification System Using the Web and Fact Databases. In *CIKM 2017*.