Where the Truth Lies: Explaining the Credibility of Emerging Claims on the Web and Social Media

Kashyap Popat, Subhabrata Mukherjee,
Jannik Strötgen, Gerhard Weikum

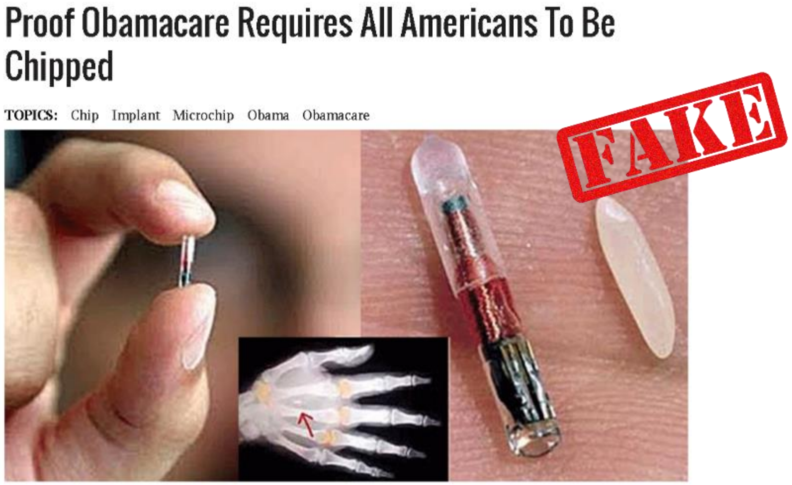
WWW 2017

MAX-PLANCK-GESELLSCHAFT

max planck institut
informatik

MOTIVATION

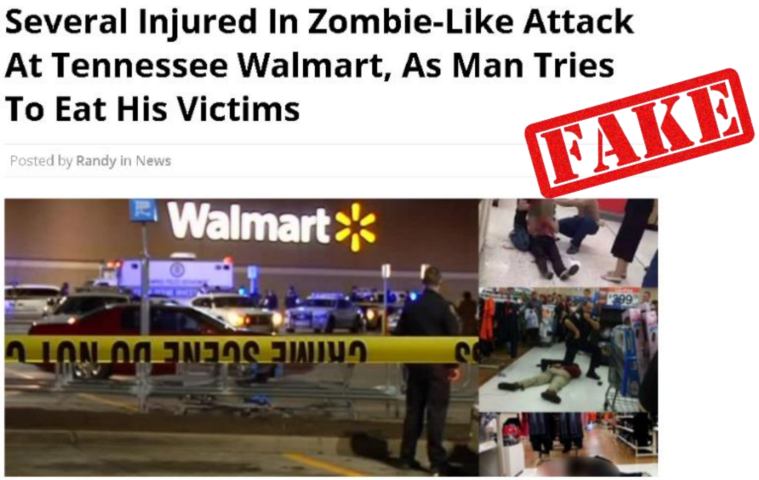
- “Rapid spread of misinformation online" – one of the top 10 challenges as per *The World Economic Forum*
- Many truth-checking websites manually verify/falsify claims



Proof Obamacare Requires All Americans To Be Chipped

TOPICS: Chip Implant Microchip Obama Obamacare

FAKE



Several Injured In Zombie-Like Attack At Tennessee Walmart, As Man Tries To Eat His Victims

Posted by Randy in News

FAKE

[1] http://www.washingtonstarnews.com/proof-obamacare-requires-all-americans-to-be-chipped/

[2] http://theracketreport.com/several-injured-in-zombie-like-attack-at-tennessee-walmart-as-man-tries-to-eat-his-victims/

Related Work & Limitations

- Truth Finding
  - Conflict resolution amongst multi-source data
  - Uses unsupervised methods to jointly infer source reliability and truth

Limited only to the structured data

No usage of linguistic cues

Related Work & Limitations

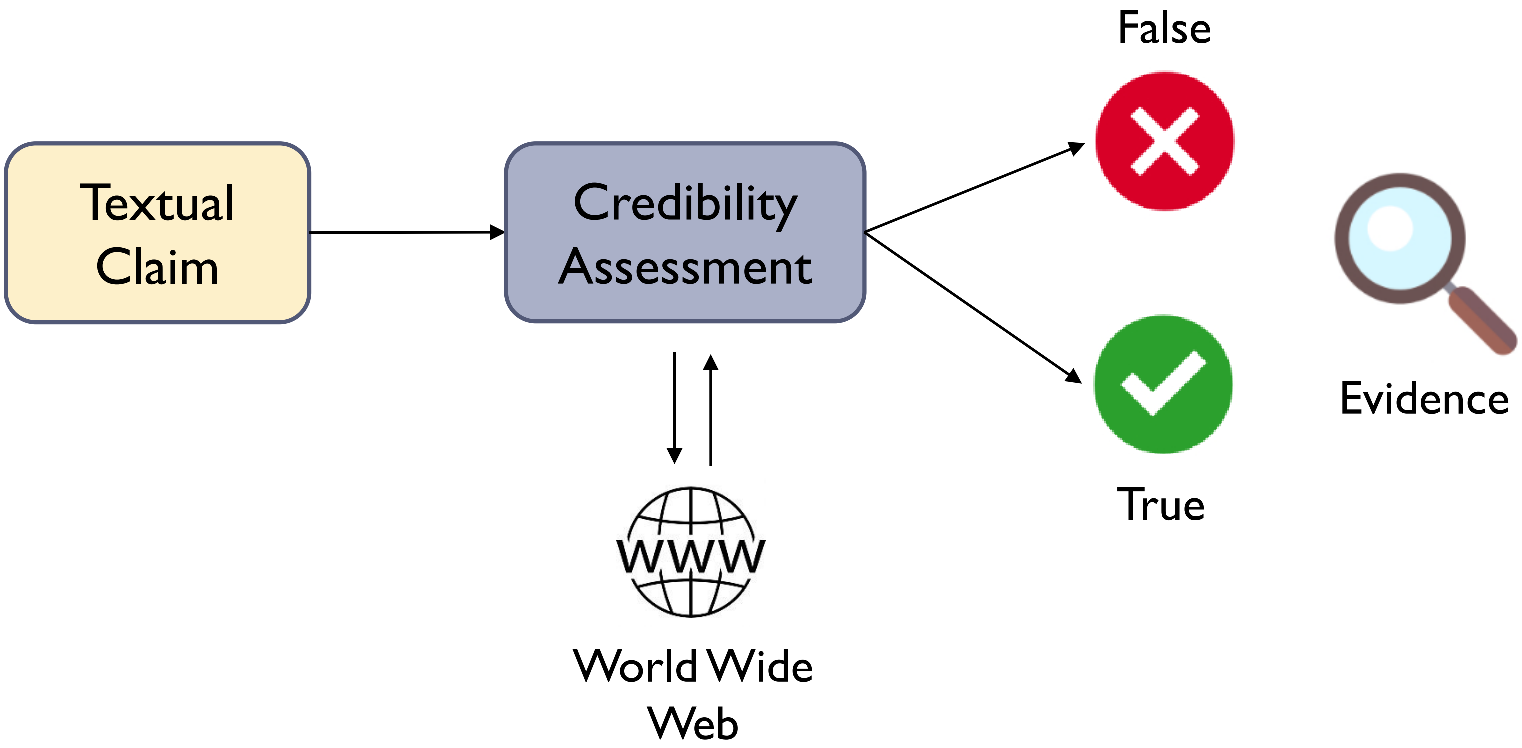
- Truth Finding
  - Conflict resolution amongst multi-source data
  - Uses unsupervised methods to jointly infer source reliability and truth
- Credibility Analysis within Communities and Social Media
  - Probabilistic graphical models
  - Social Network analysis

Focused only on closed communities

Community specific features

Problem Statement

- Given a textual claim, build an automatic system which assesses its credibility and tells whether it is *true* or *false*
  - Presents interpretable evidence supporting the assessment



Textual Claim

Credibility Assessment

World Wide Web

False

True

Evidence

OUTLINE

- Motivation
- Problem Statement
- Our Approaches
  - Key Contributors
  - Approach: Content-aware Approach
  - Approach:Trend-aware Approach
- Experiments & Results
- Conclusion

Key Contributors

- How is the claim reported? – Language style
  - Objective v/s subjective
  - Sensationalism
- Does the article support the claim? – Determining stance
  - Article can refer to the claim in negated form

    "… is a mere rumor…"
- Who is reporting the claim? – Web source reliability
  - Credible sources provide credible information
  - BBC v/s Trump Tweet
- Temporal footprint of the claim
  - Belief about various claims and how they are discussed keep changing over the time

Language Stylistic Features

Lexicon	Examples
Assertive Verbs	claim, point out…
Factive Verbs	realize, revealed…
Hedges	may have, possibly…
Implicatives	murdered, complicit…
Report Verbs	argue, denied…
Discourse Markers	could, therefore…
Subjectivity and Bias	fantastic, talented, hate…

- Normalized frequency as feature values

DETERMINING STANCE

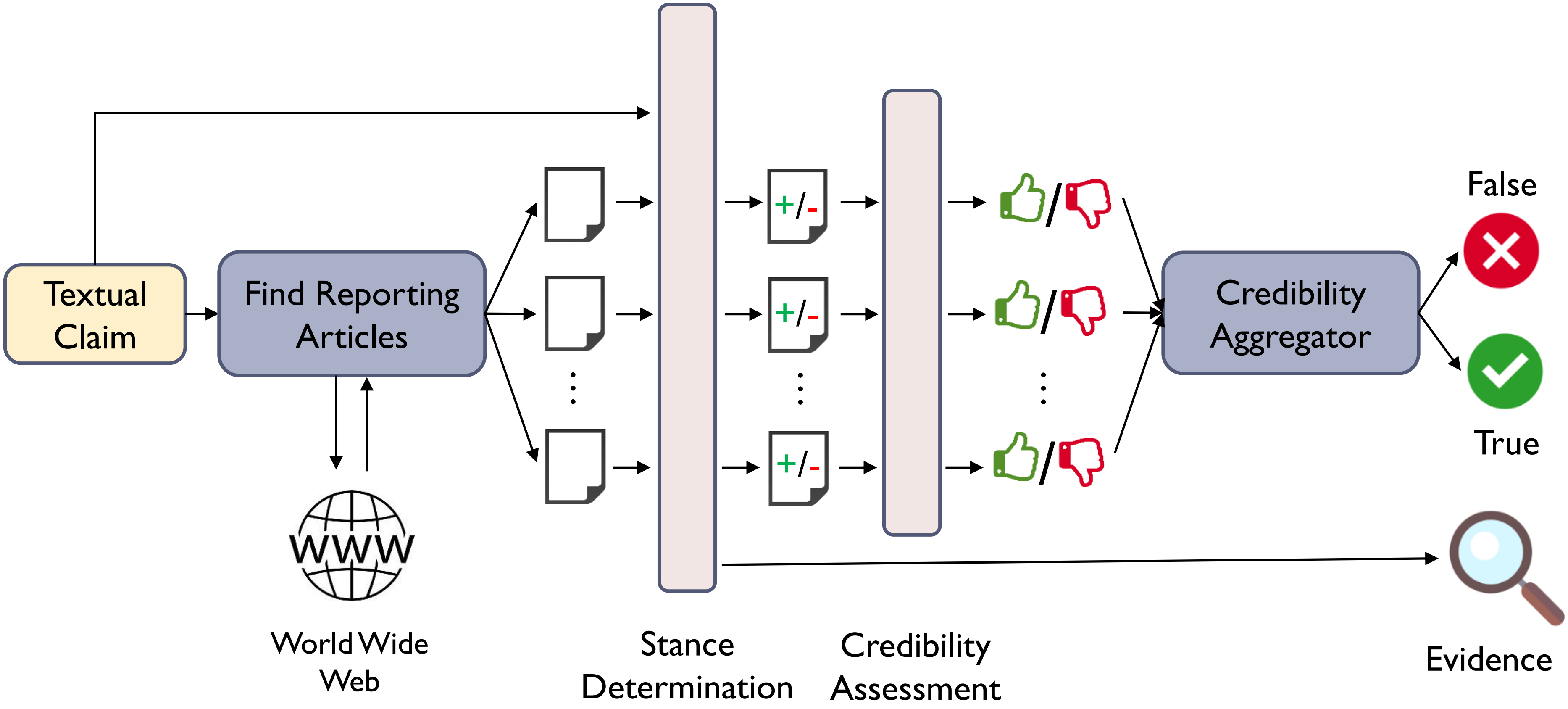
- To understand the stance of an article,
  - Divide the article into a set of overlapping snippets
  - Calculate *support* and *refute* probabilities of snippets using "*stance classifier*"
  - Get *top-k* snippets which are highly related to the claim and also have a strong refute or support probability
- Average *support* and *refute* scores of *top-k* snippets as two separate features in our model
- These *top-k* snippets are also used as supporting evidence
  - e.g., claim "X" is "*false*" because a credible website "*so-and-so*" mentions - "*… the information about X is false…*"

Web-Source Reliability

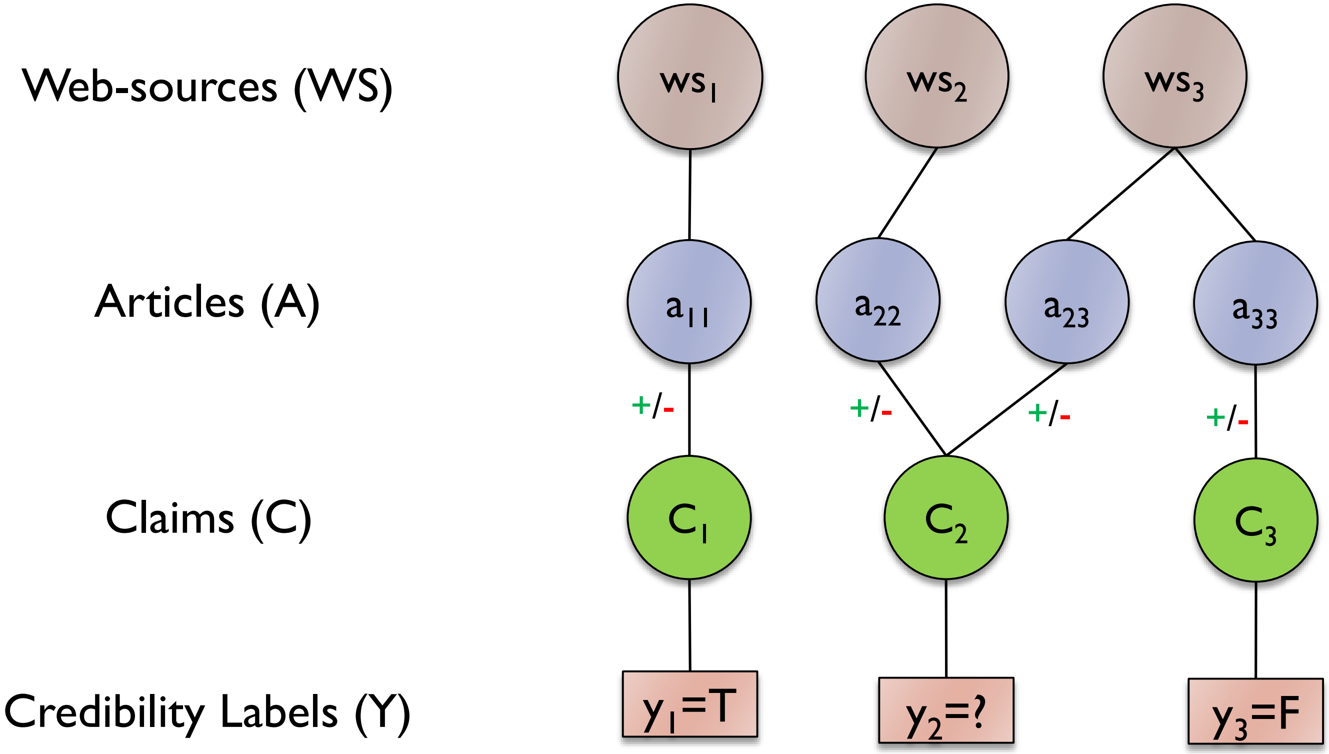
- A web-source is reliable if it publishes articles that support true claims and refute false claims
- Given a web-source ws with articles for claims with corresponding credibility labels

$$\text{reliability}(ws) = \frac{\#support\_true + \#refute\_false}{\#total\_articles}$$

SYSTEM FRAMEWORK



MODEL SETTING



Web-sources (WS)

Articles (A)

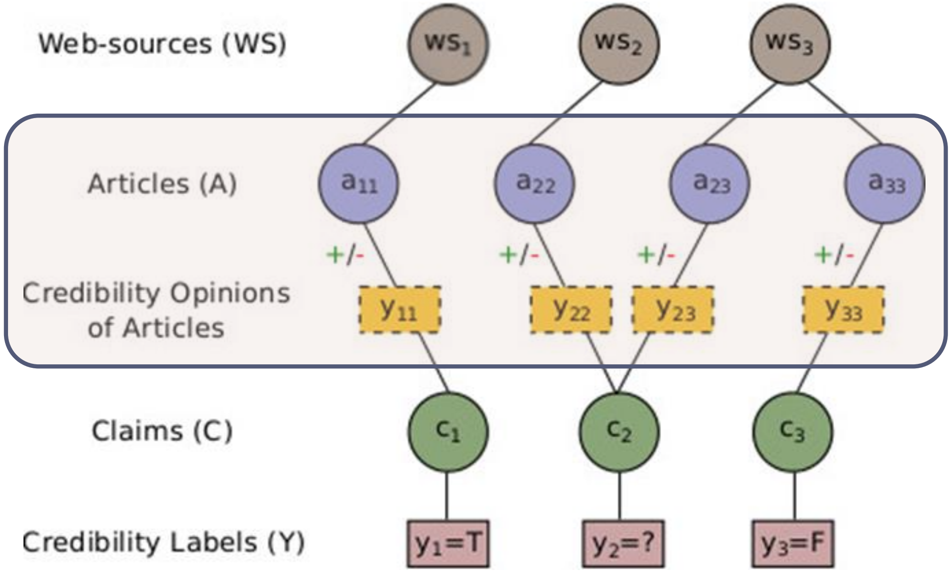
Claims (C)

Credibility Labels (Y)

- **Model:** Distant Supervision and CRF

APPROACH: CONTENT-AWARE APPROACH

- Train the logistic regression model using linguistic and stance related features – *Credibility Classifier*

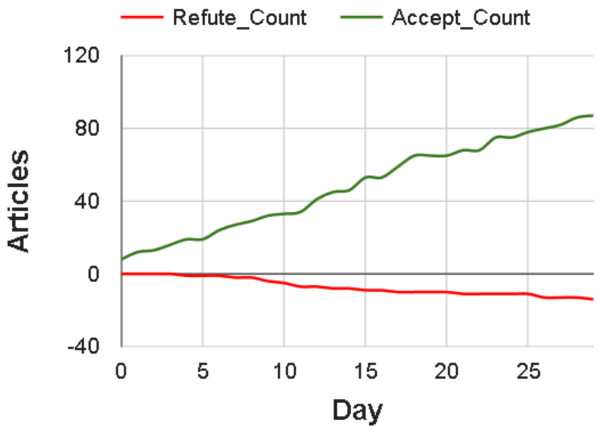


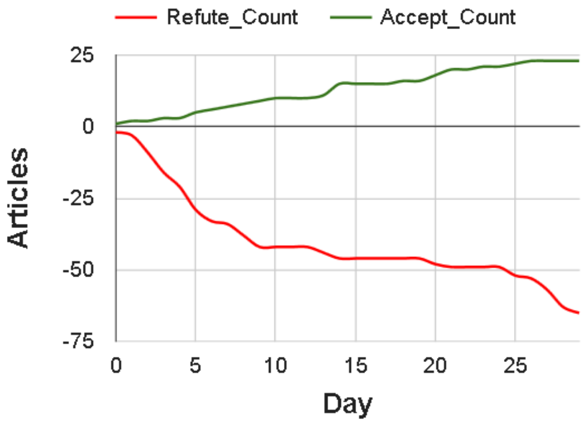
- Given a *test* claim c_i and its corresponding reporting articles, the credibility of claim is

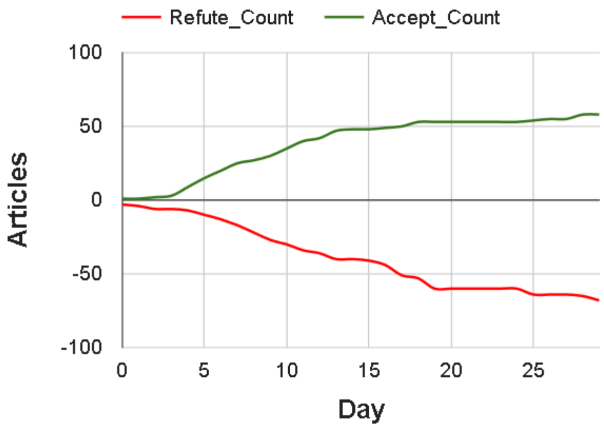
$$y_i = argmax_{\{True,False\}} \sum_{articles} [reliability(ws) * credibility\_opinion]$$

Temporal Footprint of Claims

- Belief about various claims and how they are discussed keep changing over the time
- The idea is to utilize these behavioral changes (gradient) for early detection







✓ The Centers For Disease Control confirmed that a patient in Dallas has tested positive for Ebola.

✗ Actor Macaulay Culkin has died.

✗ The iPhone 6 Plus will bend easily if placed in a pocket.

Replacing Absolute Count

- **Support/Refute Strength**: support/refute score weighted by the corresponding web source reliability instead of absolute count

$$strength^{+} = \sum_{articles} prob(support) \, * reliability\,(ws)$$

$$strength^{-} = \sum_{articles} prob(refute) \, * reliability\,(ws)$$

APPROACH: TREND AWARE APPROACH

- Calculate the slope of the trend line fitting the support/refute strength values over time
- Trend aware credibility score at time *t*,

$$Cr_{trend}(c,t) = strength_t^{+} * (1 + slope_t^{+}) - strength_t^{-} * (1 + slope_t^{-})$$

- Combining it with the content aware approach

$$Cr_{comb}(c,t) = \alpha * Cr_{content}(c,t) + (1-\alpha) * Cr_{trend}(c,t)$$

OUTLINE

- Motivation
- Problem Statement
- Our Approaches
- Experiments & Results
  - Assessment: Content-aware Approach
    - Case Study-1: Snopes
    - Case Study-2: Wikipedia
    - Handling "long-tail" claims
    - Social media as a source of evidence
  - Assessment: Trend-aware Approach
- Conclusion

ASSESSMENT: CONTENT-AWARE APPROACH

- Case Study-1: Snopes
  - Comparison with prior work baselines
  - Dissecting the performance
- Handling the "long-tail" claims
  - Does our approach handle claims with few articles?
- Social media as a source of evidence
  - How well does our approach utilize the social media?
- Case Study-2: Wikipedia
  - Evaluating the generality of our approach
- Evaluation Measures
  - Accuracy: overall, per-class, macro-averaged & AUC
  - Precision, Recall and F1-Score for *false* claims

Case Study-1: Snopes

- Used Snopes website (http://snopes.com/) to get the ground truth data for training
  - Verifies Internet rumors, hoaxes, and other claims
- Gathered ~4800 claims with their credibility (true/false)
- For each claim, fetched first 3 pages of Google search result

"Australia is the first country to begin microchipping its citizens"

"Entering your PIN in reverse at any ATM will automatically summon the police"

"President Obama ordered a life-sized bronze statue of himself to be permanently installed at the White House"

"Bernie Sanders purchased a $172,000 luxury car with presidential campaign donations"

Comparison with Baselines

Configuration	Macro-averaged Accuracy (%)
ZeroR	50.00
Generalized Investment (Pasternack et al., 2010)	54.33
Truth Assessment (Nakashole et al., 2014)	56.06
Truth Finder (Yin et al., 2008)	56.91
Generalized Sum (Pasternack et al., 2011)	62.82
Pooled Investment (Pasternack et al., 2010)	63.09
Average-Log (Pasternack et al., 2011)	65.89
Lang & Auth (Popat et al., 2016)	73.10
Our Approach: Distant Supervision	**82.00**

10-fold cross-validation

Dissecting the Performance

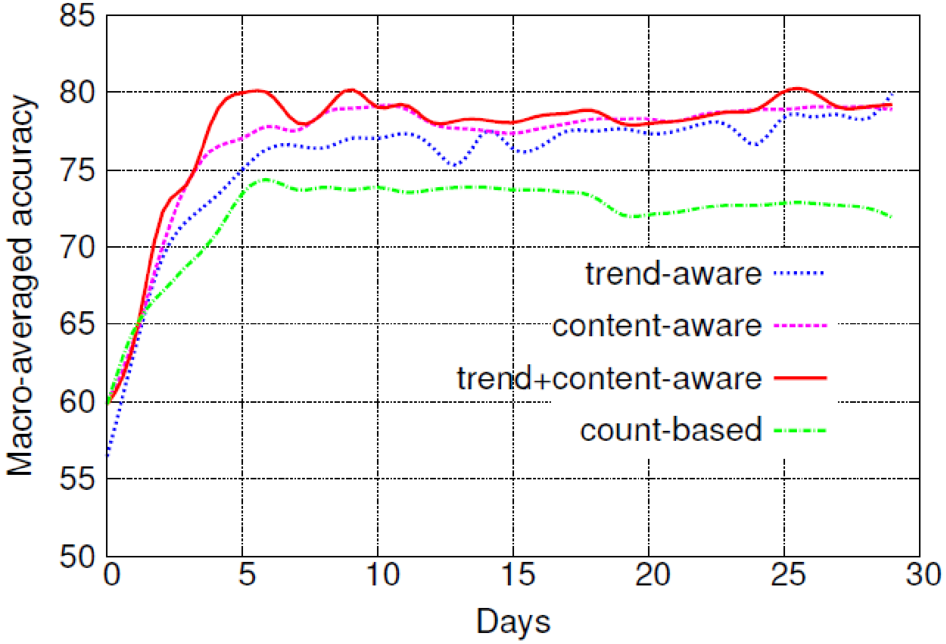
Configuration	Macro-averaged Accuracy (%)	AUC
Language + Stance + Reliability	**82.00**	**0.88**
Stance + Reliability	79.67	0.86
Language + Stance	73.76	0.81
Language + Reliability	71.34	0.77
Stance	68.97	0.76
Language	69.07	0.75

10-fold cross-validation

- Only language stylistic features not enough – crucial to understand the stance and web-source reliability

ASSESSMENT: TREND-AWARE APPROACH

- Compare performance on each day
- Combined approach performs the best
- Early detection of emerging claims in 4-5 days with high accuracy
- Absolute count of supporting/refuting articles is not sufficient



CONCLUSION

- Proposed a general approach for credibility analysis of unstructured textual claims in an open-domain setting
- Provide interpretable evidence
- Experiments on real-world claims demonstrate effectiveness of our approaches
- Early detection of emerging claims by capturing their temporal footprint
- Datasets available: bit.ly/web-credibility-analysis

THANK YOU!
KASHYAP – kpopat@mpi-inf.mpg.de

Claim	Verdict & Web Evidence
The use of solar panels drains the sun of energy.	False - Solar panels do not suck up the Sun's rays of photons. Just like wind farms do not deplete our planet of wind. These renewable sources of energy are not finite like fossil fuels. Wind turbines and solar panels are not vacuums, nor do they divert this energy from other systems.
A woman stabbed her boyfriend with a sharpened selfie stick because he didn't like her newest Instagram selfie quickly enough.	False - A weird kind of story in heavy circulation online states ... No, the claim is not a fact.
Between 1988 and 2006, a man lived at a Paris airport.	True - Mehran Karimi Nasseri (born 1942) is an Iranian refugee who lived in the departure lounge of Terminal One in Charles de Gaulle Airport from 26 August 1988 until July 2006 … His autobiography has been published as a book (The Terminal Man) and was the basis for the 2004 Tom Hanks movie The Terminal.
Soviet Premier Nikita Khrushchev was denied permission to visit Disneyland during a state visit to the U.S. in 1959.	True - Soviet Premier Nikita Khrushchev's good-will tour of the United States in September 1959. While some may have heard of Khrushchev's failed attempt to visit Disneyland, many do not realize that this was just one of a hundred things that went wrong on this trip.