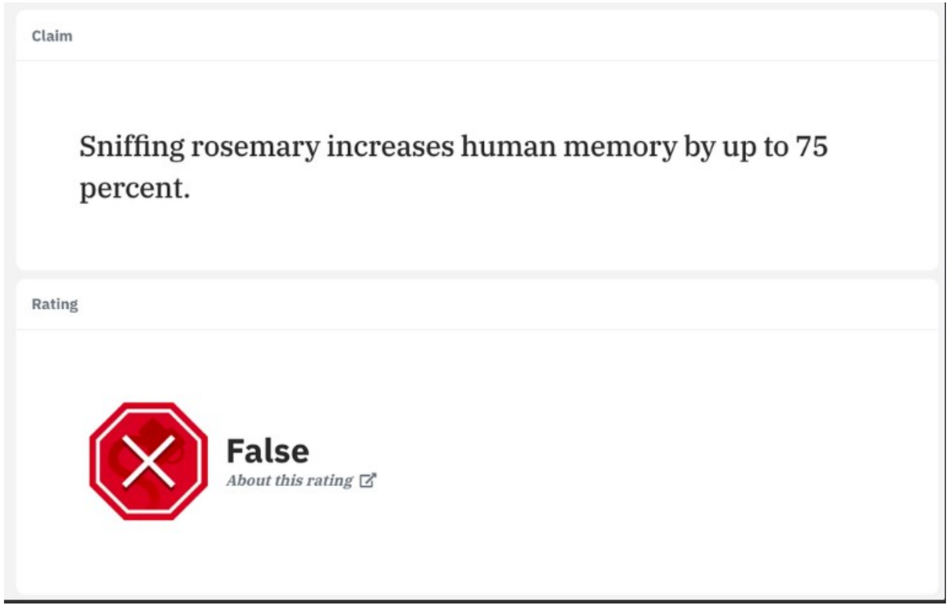
Fake News Detection

Cong wan

Example of fake news



Overview

- Motivation
- Credibility Assessment of Textual Claims on the web
- Where the Truth lies:Explaining the Credibility of Emerging Claims on the web and Social Media

Motivation

- **Different types of fake news:**
  - wrong news on websites,
  - erroneous stock prices,etc

- **Fact-checking websites have become popular**
  - snopes.com, politifact.com……
  - However, low efficiency. These websites are written by experts, who **manually** investigate claims and provide a verdict (true or false)

Limitation of prior work

- Prior works **follow a structured template** (who replied to whom, who edited what…) , for example:**obama is born in Kenya**. These works **cannot handle many kinds of claims**, which are in form of **long sentences** or **other structures.**

- this paper aims to overcome these limitations by **making no assumption** on the structure of the claim,and by addressing the case of **arbitrary textual claims** that are expressed freely in an open-domain setting.

Overview of Approach

- Given a claim in the form of a sentence or paragraph.
  - Firstly, using a **search engine** to identify **documents** from multiply web-source, which refer to the claim(**reporting article**)
  - Then, analyze the interplay between the **language**(bias, subjectivity,etc.) of the retrieved articles,and the **reliability** of the relevant web-sources(where the articles appeared).
  - Finally, propose a **Distant Supervision** based classifier to assess the credibility of the claim.

Credibility Assessment

1) Language Stylistic Features

- true claim ⟷ objective and unbiased language.

- less credible claim ⟷ highly subjective or sensationalized style

--From Amazon Mechanical Turk

- Language stylistic features

Type of features	description
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

Types of linguistic features used in model

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

Credibility Assessment

2)Source Reliability

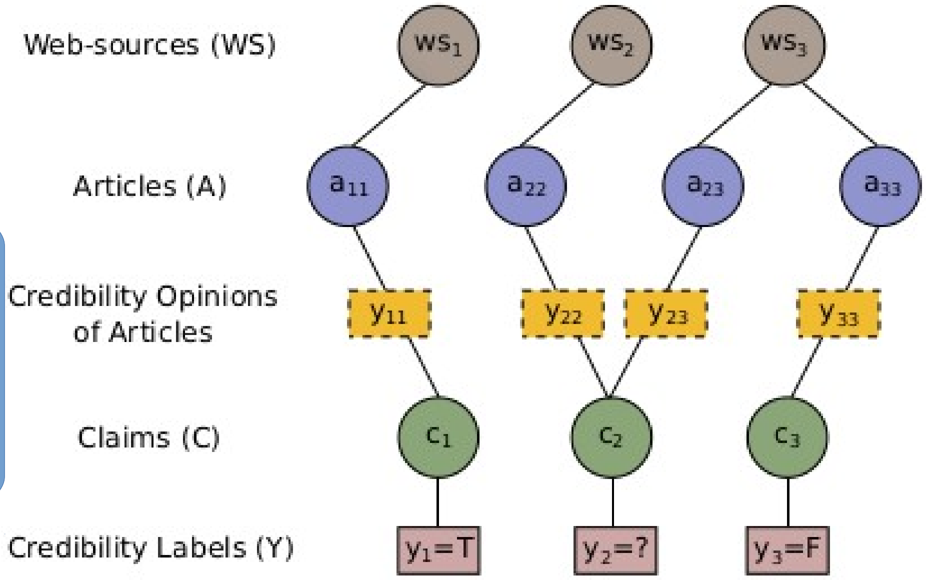
use **AlexaRank** and **PageRank** as proxies for the source reliability

Type of Feature	Number of Features
Reliability	
Source Identity	#web-sources
PageRank	1
AlexaRank	1

Table: Statistics of features used in model

Credibility classification

Alternative approach to generating training data



- Use **Distant Supervision** for training:
  - Attach the label y_i to article a_{ij}, for example: $y_1 = y_{11} = T$
  - For any claim c_i whose credibility label is unknown, determine the overall credibility label y_i of c_i by computing:

$$y_i = \arg\max_{l \in \{T,F\}} \sum_{a_{ij}} Prob(y_{ij} = l)$$

Data set

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.

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

Table 3: Wikipedia data statistics.

Data from snopes.com:
- Use only the claim and credibility verdict(true or false)
- Example:North Carolina no longer considers the $20 bill to be legal tender --false

Data from wikipedia.org:
- Collect a set of 100 proven hoaxes
- Collect a set of 57 fictitious people
- Ground-truth label for all of these claims is **False**

Experiment

$$F1 = \frac{2TP}{2TP + FN + FP} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

- Using the data from snopes.com to train the classifier,

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

Principle: Select a category with the highest probability as the classification result of the unknown sample

Result

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

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

Hoaxes: the authors collect a set of 100 proven hoaxes reported on Wikipedia,e.g.,”Alien autopsy film by Ray Santilli”,”disappearing blonde gene” etc, All these hoaxes can be mapped to claims of type “<ENTITY> exists” etc, the ground truth label for all of these claims is ***Fake***

Fictitious people: in addition, the authors also collect 57 fictitious people, e.g.,”Ern Malley,an Australian poet”, also in type of ”<ENTITY> exists”. The ground-truth label for all of these claims is also ***fake***.

conclusions

- Propose a general approach for credibility analysis of **unstructured** textual claims in an **open-domain.**
- Make use of **language style** and **source reliability** of articles reporting the claim to assess its credibility

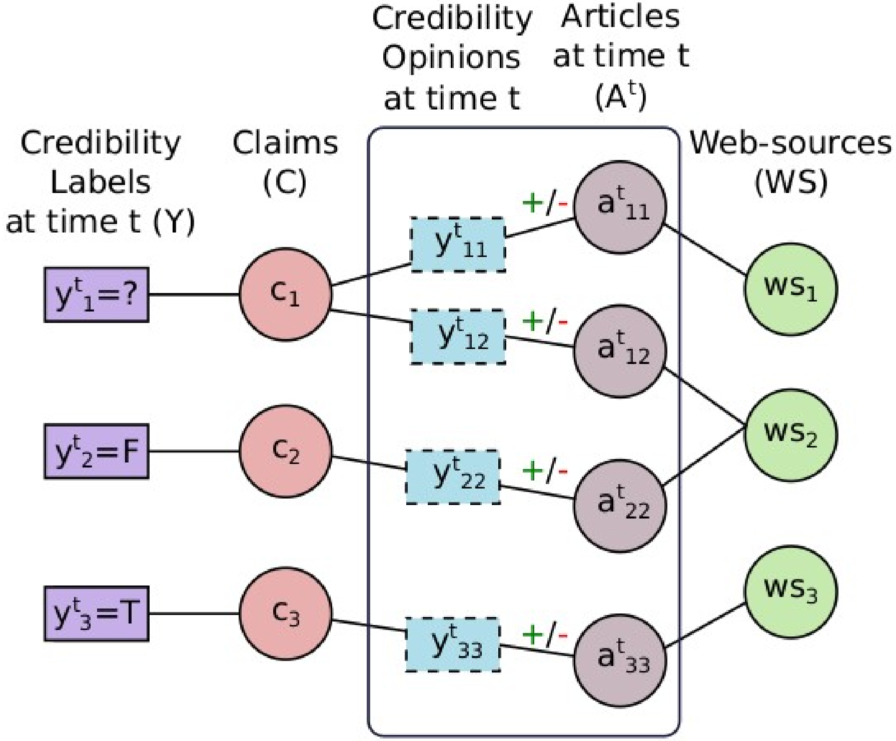
Overview

- Motivation
- Credibility Assessment of Textual Cliams on the web
- **Where the Truth lies:Explaining the Credibility of Emerging Claims on the web and Social Media**

Limitation of the 1st paper(prior work)

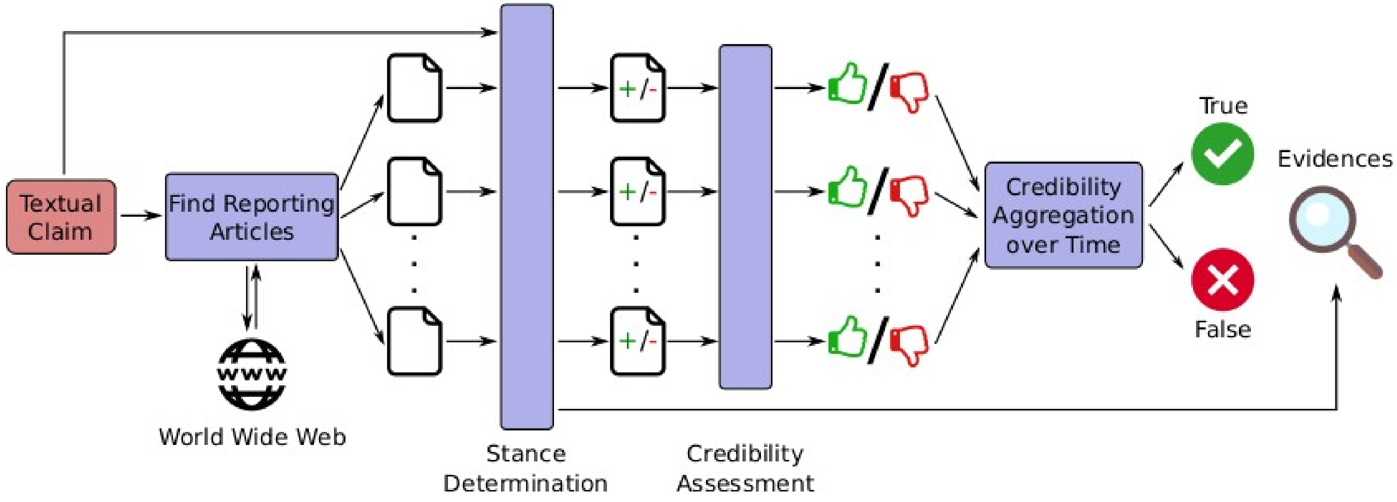
- 1st paper just Computing a verdict(true or false) without **providing the explanation,** this will be achieved in this paper by providing user interpretable explanations for the verdict.
  - 1st paper assume that they could easily retrieve enough evidence or counter-evidence from a snapshot of the web, disregarding the dynamics of how claims emerge, spread, and are supported or refuted, in this paper, the author will take these factors into account by determining the stance, reliability, and trend of retrieved sources of evidence or counter evidence.

Problem statement



- Given the labels of a subset of the claim(e.g. y^t_2 for c_2,y^t_3 for c_3)
- The goal is to predict the credibility label of the newly emerging claim(e.g.,y^t_1 for c_1 at each time point t)

Approach



- Given a newly emerging claim in the form of a (long)sentence or a paragraph at time t

  1) Use a **search engine** to identify **documents** from a lot of web-sources **referring to the claim** and refer these documents as **reporting articles.**

  2) Stance Determination by analysing the interplay between following factors:
     - Linguistic features(same as the 1st paper)
     - The stance of the article towards the claim
     - The Reliability of the web source

  3) Credibility Assessment by using 2 methods :
     - Distant Supervision
     - Conditional Random Field(CRF)

Finding stance and Evidence

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 *truthorfiction.com* refuting the claim will make the claim less credible.

- **Stance determination method**:
  - Input: claim c_i and a corresponding reporting article a_{ij} at time t
  - Output: stance scores (support & refute) of a^t_{ij} about c_i
  - 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, and remove snippets having overlap less than a **threshold (η)**
  - use a **Stance Classifier** to determine whether a remaining snippet supports or refutes the claim.
  - Average the two stance probabilities(for support and for refute) over the top-k snippets

$$F^{St}(a^t_{ij}) = \langle avg(\langle p_s^+ \rangle), avg(\langle p_s^- \rangle) \rangle.$$

Stance classifier

- Goal: given a piece of text, the stance classifier should give the probability of how likely the text refutes or supports a claim based on the language stylistic features
- Data: from snopes.com and other website (these websites analyze the origin of the claim and its corresponding credibility label)
- Model: use L2 regularized Logistic Regression from the LibLinear package.

Source Reliability

- In the past only capture the authority an popularity of web-sources
- Now,takes the **authenticity** of articles in the web into account
- A web-source is considered reliable if it contains articles that **support true claim** and **refute false claim**.

Given a web-source ws_j with articles <a^t_{ij}> for claims <c_i> with corresponding credibility labels <y_i^t>, we compute its reliability as:

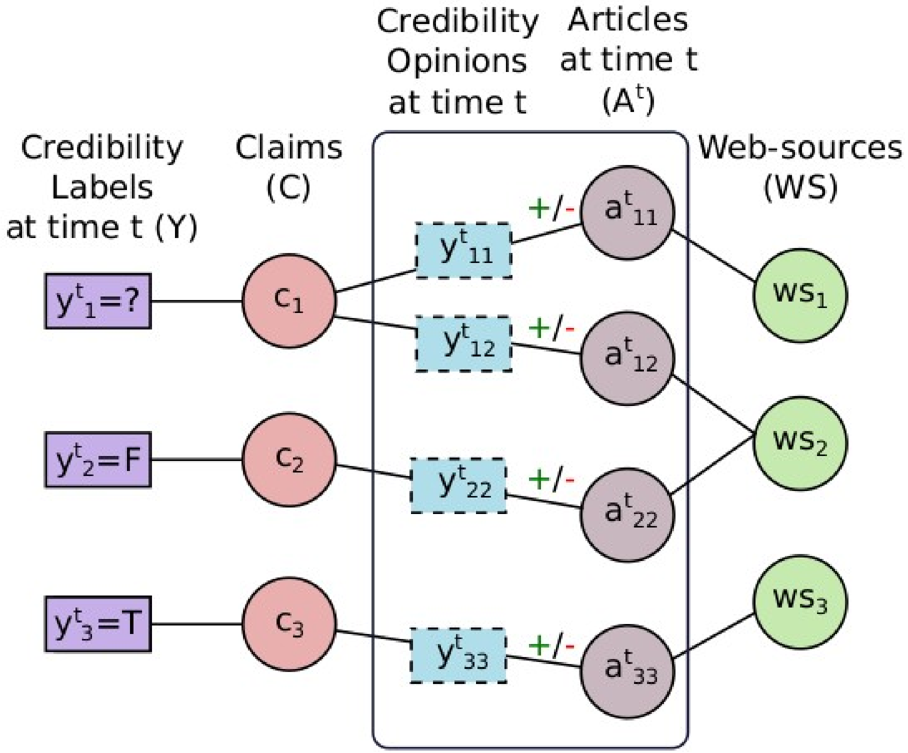
$$reliability(ws_j) = \frac{\sum_{a^t_{ij}} \mathbf{1}\{St_{a^t_{ij}} = \text{'+'}, y_i^t = T\} + \sum_{a^t_{ij}} \mathbf{1}\{St_{a^t_{ij}} = \text{'-'}, y_i^t = F\}}{\text{cardinality}(\langle a^t_{ij} \rangle)}$$

The number of articles

Credibility Assessment Models

- Content-aware assessment
  - Model based on Distant Supervision(same as 1st paper)
  - Joint model based on CRF
- Trend-aware assessment
- Content- and trend-aware assessment

Content-aware assessment



- Joint model based on Conditional Random Field(CRF)
- Operate on the clique of the graph:
  - A clique is formed amongst a claim $c_i \in C$, a source $ws_j \in WS$ and an article $a_{ij} \in A$ about c_i found in ws_j, different clique are connected via the commen sources and claims.
  - Each clique has a set of associated feature functions with a weight vector
  - Estimate the **conditional distribution.**
  - Maximize the conditional **log-likelihood** of the data.

Trend-aware Assessment

- the credibility $Cr_{trend}(c_i, t)$ of a claim c_i at each day t is influenced by two components:

  a)the strength of support and refute till time t(denoted by $q^+_{i,t}$ and $q^-_{i,t}$,respectively)

- b)the slope of the trendline for the support and refute strength for the claim c_i till time t(denoted by $r^+_{i,t}$ and $r^-_{i,t}$,respectively)
- The score of claim c at time t:

$$Cr_{\text{trend}}(c_i, t) = [q^+_{i,t} \cdot (1 + r^+_{i,t})] - [q^-_{i,t} \cdot (1 + r^-_{i,t})]$$

Content and Trend-aware Asssessemts

- Combination of this two approachs:

$$Cr_{comb}(c_i, t) = \alpha \cdot Cr_{\text{content}}(c_i, t) + (1 - \alpha) \cdot Cr_{\text{trend}}(c_i, t)$$

Combination weight

Experiment

- Data set
  - Snopes.com and Wikipedia.org, just refer to the first paper.
  - Time-series dataset
    - It is quite difficult to get such time-series data for the open web,
    - to mimic the time-series behavior, use the Google search engine to search and retrieve relevant reporting articles on a claim on each day, starting from its day of origin to the next 30 days

Experiment

- Content-aware Assessment on Snopes and wikipedia

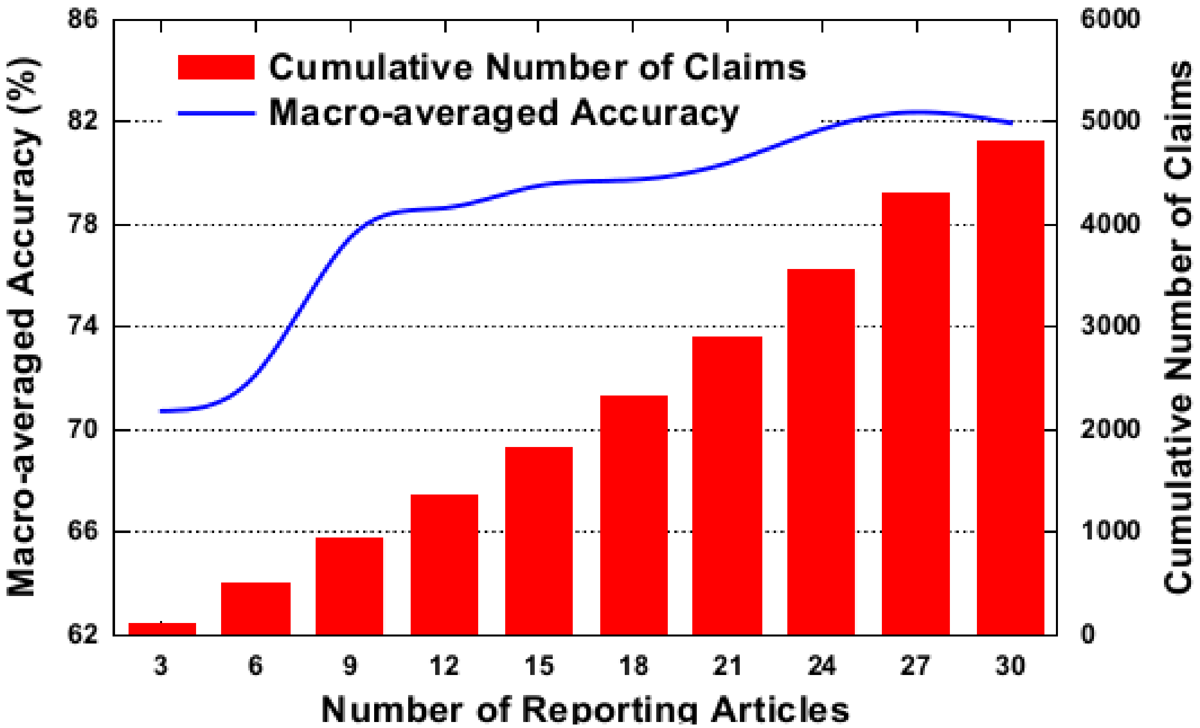
    perform 10-fold cross-validation on the claims by using 9-folds of the data for training, and 1 fold for testing

Configuration	Macro-averaged Accuracy (%)
ZeroR	50.00
Generalized Investment [25]	54.33
Truth Assessment [24]	56.06
TruthFinder [42]	56.91
Generalized Sum [27]	62.82
Pooled Investment [25]	63.09
Average-Log [27]	65.89
Lang. & Auth. [29]	73.10
Our Approach: CRF	**80.00**
Our Approach: Distant Supervision	**82.00**

Test Data	#Claims	Lang.+Auth. [29] Accuracy (%)	LG+ST+SR Accuracy (%)
WikiHoaxes	100	84	88
WikiFictitious People	57	66.07	82.14

Experiment

- Handling “Long-tail” claims



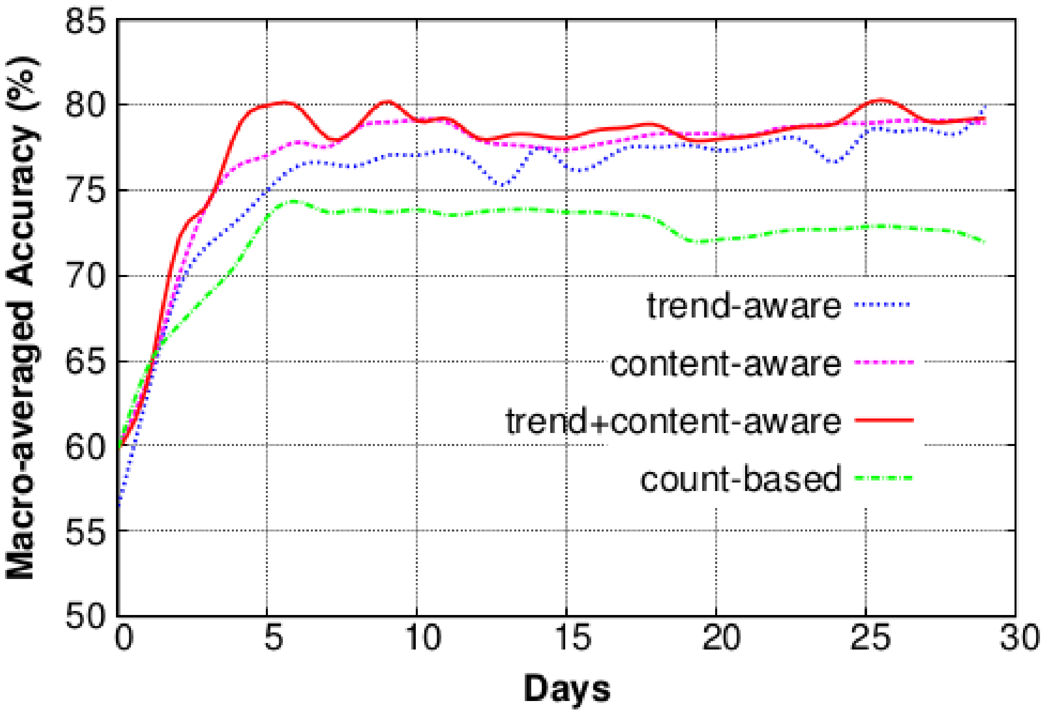
Long-tail claim: have a long sentence and have only few reporting articles. In this experiment ,only consider those claims ≤ k reporting articles, k∈{3,6,9,...30}

Figure shows the change in accuracy of claims with different k

Right side Y-axis :cumulative number of selected claims.

Experiment

- Credibility Assessment of newly Emerging Claims
- Compare with different approaches



Worst performance: count-based
Reason:simply counting the number of supporting / refuting articles is not enough.

best performance:content & trend-aware approach.
Reason: consider many different factors that may influence the result into account

30 days: the accuracy between them are similar

Experiment

- social media as source of evidence

***social media* as source of evidence**: e.g. Facebook, Twitter

***Web* as source of evidence**: consider reporting articles from all sources on the web, including the social media sources.

Configuration	Overall Acc. (%)	*True* Claims Acc. (%)	*False* Claims Acc. (%)	Macro-averaged Acc. (%)
Social Media	76.12	77.34	75.66	76.50
Web	84.23	86.01	83.56	84.78

Experiment

Evidence for Credibility Classification

Claim	Verdict & Evidence
Titanium rings can be removed from swollen fingers only through amputation.	**[Verdict]: False [Evidence]:** A rumor regarding titanium rings maintains that ... This is completely untrue. In fact, you can use a variety of removal techniques to safely and effectively remove a titanium ring.
The use of solar panels drains the sun of energy.	**[Verdict]: False [Evidence]:** 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.
Facebook soon plans to charge monthly subscription fees to users of the social network.	**[Verdict]:False [Evidence]:** 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.
Soviet Premier Nikita Khrushchev was denied permission to visit Disneyland during a state visit to the U.S. in 1959.	**[Verdict]: True [Evidence]:** 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.
Between 1988 and 2006, a man lived at a Paris airport.	**[Verdict]: True [Evidence]:** 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, when he was hospitalized for an unspecified ailment. His autobiography has been published as a book (The Terminal Man) and was the basis for the 2004 Tom Hanks movie The Terminal.

Conclusions

- propose approaches to leverage the stance, reliability and trend of sources of evidence and **counter-evidence** for credibility assessment of textual claims
- performs well on assessing the credibility of **newly emerging claims** within 4 to 5 days of its day of origin on the web with 80% accuracy,as well as for **“long-tail” claims** having as few as 3 reporting articles
- can effectively harness evidence from **noisy source**(social media) to validate or falsify a claim.
- provide explanations for the credibility verdict in the form of informative snippets from articles published by reliable sources that can be **easily interpreted** by the users

Thanks for your attention !