Automating the Fact-checking task
Challenges and Directions

Dr. Diego Esteves

Farfetch, Portugal
SDA Research, Germany
diegoesteves@gmail.com

07.05.2020

me.

- Professional Experience
  - +/- 17 years professional experience (8 w/ data, 9 w/ eng)
    - Principal DS @Farfetch.com, Portugal
    - Research Scientist@SDA, Germany
    - Data Analyst@BTG Pactual, Brazil
    - …
- Academic
  - PhD in CS (Bonn Universitat)
  - MSc in Eng. (Instituto Militar de Engenharia - IME)
  - MBA (Universidade Federal do Rio de Janeiro - UFRJ)

https://www.linkedin.com/in/diegoestevesde/

Research Group

SMART DATA ANALYTICS FROM DATA TO KNOWLEDGE

DBpedia

UNIVERSITÄT LEIPZIG

AKSW

EIS ENTERPRISE INFORMATION SYSTEMS

TIB LEIBNIZ-INFORMATIONSZENTRUM TECHNIK UND NATURWISSENSCHAFTEN UNIVERSITÄTSBIBLIOTHEK

Prof. Dr. Soren Auer

SMART DATA ANALYTICS FROM DATA TO KNOWLEDGE

UNIVERSITÄT BONN

Prof. Dr. Jens Lehmann

DICE

UNIVERSITÄT PADERBORN

Prof. Dr. Axel Ngonga

HTWK Leipzig

Prof. Dr. Thomas Riechert

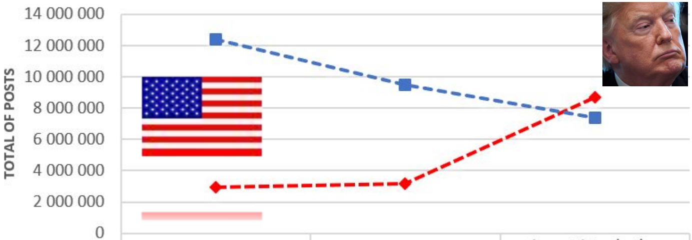
InfAI® Institut für Angewandte Informatik

Dr. Sebastian Hellmann

Motivation

The Rise of Fake News: A Global Threat

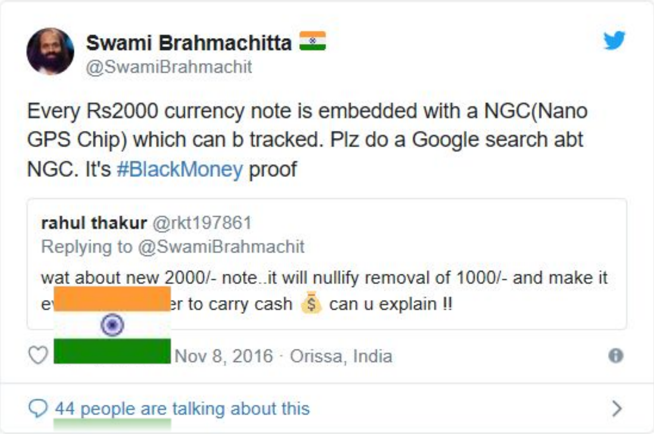
BUZZFEED NEWS: ELECTION CONTENT ENGAGEMENT



	February to April	May to July	August 1 to election day
Fake	2 970 000	3 183 000	8 711 000
Real	12 395 000	9 492 000	7 367 000

The **#top1 fake news** against Jair Bolsonaro, had 596k engagements, a number **bigger than all his #top10 real news**

A report found **3 out of the 5** most shared stories on Facebook **were false as the Dilma Rousseff impeachment process intensified**



Swami Brahmachitta
@SwamiBrahmachit

Every Rs2000 currency note is embedded with a NGC(Nano GPS Chip) which can b tracked. Plz do a Google search abt NGC. It's #BlackMoney proof

rahul thakur @rkt197861
Replying to @SwamiBrahmachit
wat about new 2000/- note..it will nullify removal of 1000/- and make it e er to carry cash 💰 can u explain !!

Nov 8, 2016 · Orissa, India

44 people are talking about this

"Today we have fake sites, bots, trolls – things that regenerate themselves, reinforcing opinions with certain algorithms, and **we have to learn to deal with them**."

The New York Times

Fake News Is Poisoning Brazilian Politics. WhatsApp Can Stop It.

By Cristina Tardáguila, Fabrício Benevenuto and Pablo Ortellado

Ms. Tardáguila, Mr. Benevenuto and Mr. Ortellado are the authors of a new report on misinformation in Brazil.

Fake News: How do they proliferate?

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

- ✓ Someone attempting to steal information/money
- ✓ Satirists who want to either make a point or entertain you
- ✓ Poor or untrained journalists (i.e., people who do not follow journalistic standards of ethics)
- ✓ Partisans who want to influence political beliefs and policy makers
- ✓ **According to *Vosoughi et al. (2018)*, falsehood is diffusing <u>faster</u> and in <u>larger scale</u> than the truth itself.**

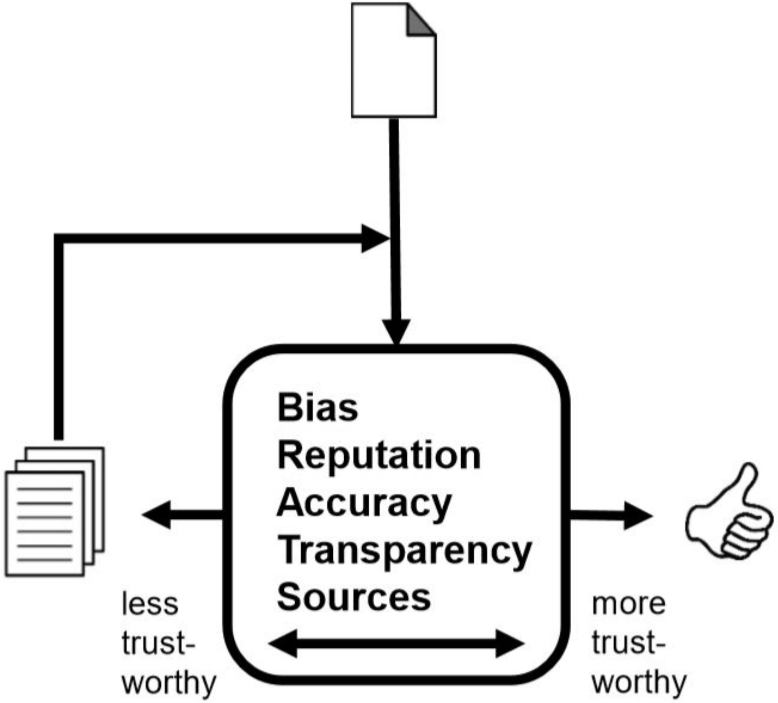
Fake News: How are they classified?

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

- ✓ **Disinformation**
  - intentionally false, spread deliberately
- ✓ **Misinformation**
  - unintentionally false
- ✓ **Clickbait**
  - exaggerating information and under-delivering it
- ✓ **Satire**
  - unintentional false for humorous purposes
- ✓ **Biased Reporting**
  - reporting only some of the facts to serve an agenda

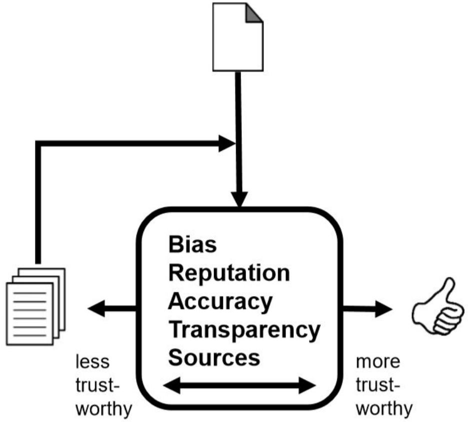
Fake News: Dealing with in real life
The BRATS Method

- ✓ **B**ias of author/publisher
- ✓ **R**eputation of author/publisher
- ✓ **A**ccuracy in reporting/use of sources
- ✓ **T**ransparency of sources/methods
- ✓ **S**ources used by author



Research Problem

Can the fact-checking task be automatized?



Computing Trustworthiness of Web Sources — Challenge 2

Detecting Entities on Noisy Data — Challenge 1

Claim Verification — Challenge 3

Automated Fact-checking Frameworks

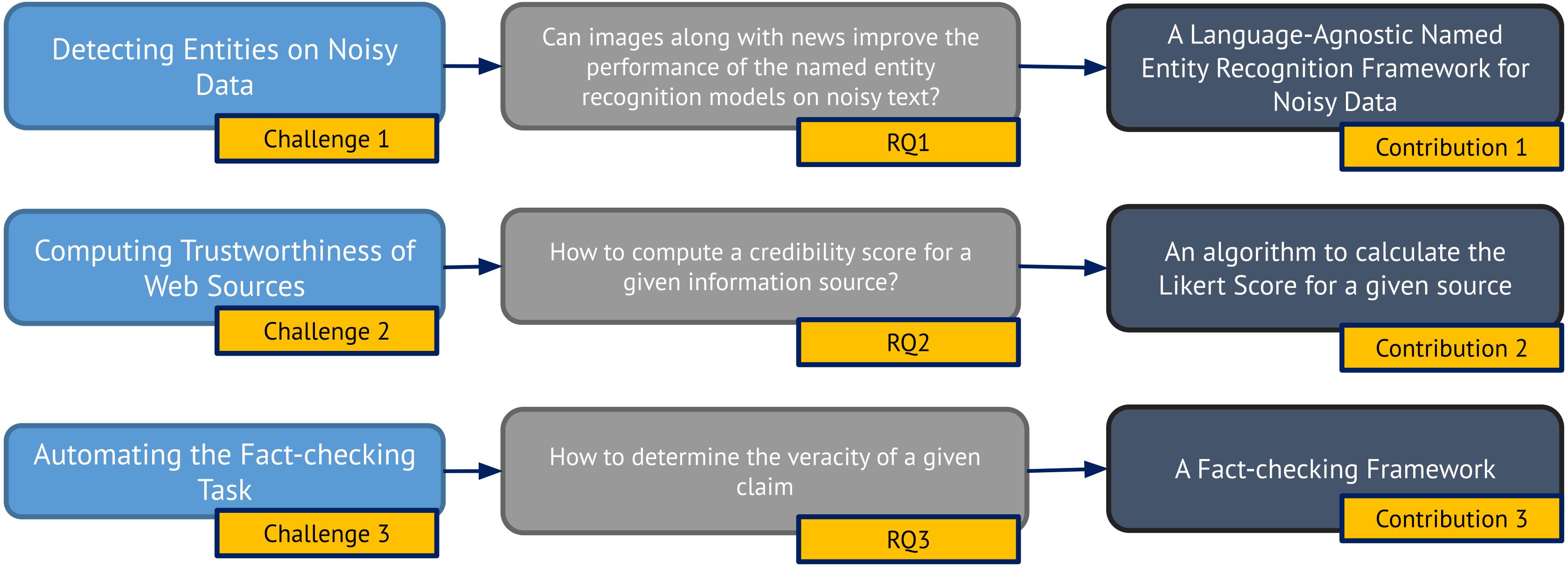
Evidence Extraction

- Document Selection
- Source Trustworthiness (*)
- Sentence Selection
- Claim Classification

FactBench

WSDM 2017 Triple Scoring

Research Questions & Contributions



Named Entity Recognition for Noisy Data

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

RQ1	Can images along with news improve the performance of the named entity recognition models on noisy text?

Named-entity recognition (NER) is a subtask of information extraction aiming to locate **named entities** in natural language documents:

S = Diego Esteves lives in Porto, Portugal.

RQ1 Contribution

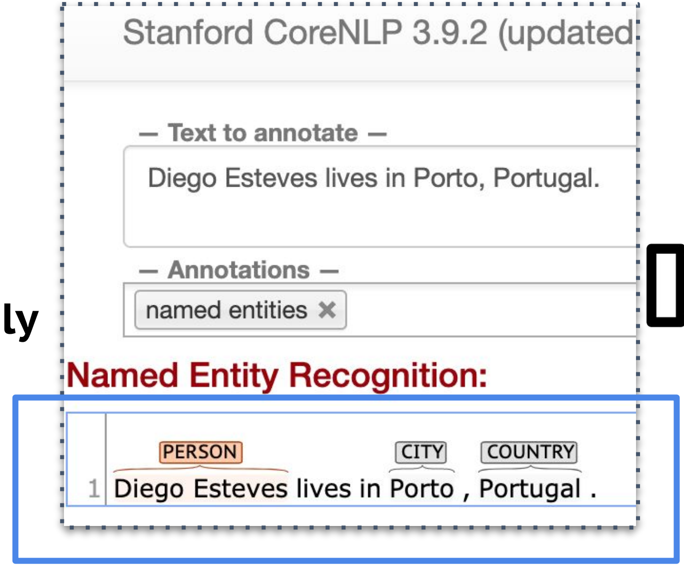
A Named Entity Recognition Framework for Noisy Data

Contribution 1

Named Entity Recognition for Noisy Data

RQ1	Can images along with news improve the performance of the named entity recognition models on noisy text?

- Lexical, Shape and Orthographic features
- Gazetteers
- **SOTA high performance in formal domains (easily 0.90 F1)**

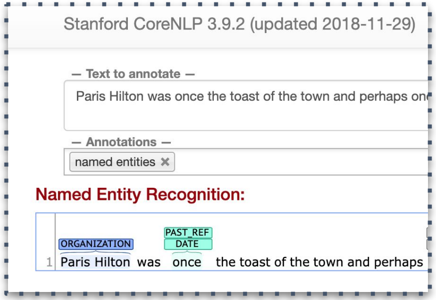


Named Entity Recognition for Noisy Data

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

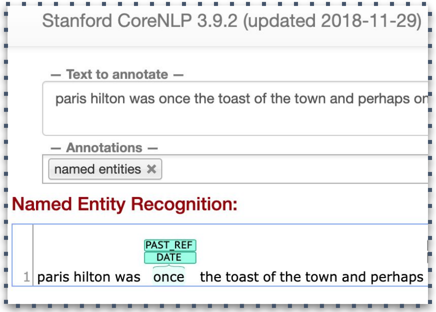
RQ1 Can images along with news improve the performance of the named entity recognition models on noisy text?

S = Paris Hilton was once the toast of the town and perhaps one of Hollywood's most famous socialites.



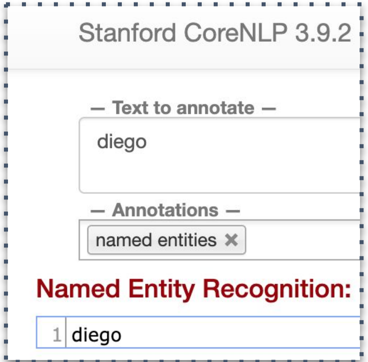
What if small variations are applied?

e.g., *S* = paris hilton ?



What about non-english names?

e.g., S = diego ?



Named Entity Recognition for Noisy Data

RQ1 — Can images along with news improve the performance of the named entity recognition models on noisy text?

- Look-up strategies and standard local features struggle on noisy data

- **F1 0.20 and 0.60**
  [Ritter et al. 2011]
  [Derczynski et al., 2015]
  [Esteves et al., 2017]
  [Qi Zhang et al., 2018]

'2m', '2ma', '2mar', '2mara', '2maro', '2marrow', '2mor', '2mora', '2moro', '2morow', '2morr', '2morro', '2morrow', '2moz', '2mr', '2mro', '2mrrw', '2mrw', '2mw', 'tmmrw', 'tmo', 'tmoro', 'tmorrow', 'tmoz', 'tmr', 'tmro', 'tmrow', 'tmrrow', 'tmrrw', 'tmrw', 'tmrww', 'tmw', 'tomaro', 'tomarow', 'tomarro', 'tomarrow', 'tomm', 'tommarow', 'tommarrow', 'tommoro', 'tommorow', 'tommorrow', 'tommorw', 'tommrow', 'tomo', 'tomolo', 'tomoro', 'tomorow', 'tomorro', 'tomorrw', 'tomoz', 'tomrw', 'tomz'

- Joint clustering to minimise the gap between world knowledge and KBs

- Basic idea:
  - Correlation between images and entities
  - Correlation between search textual results and entities

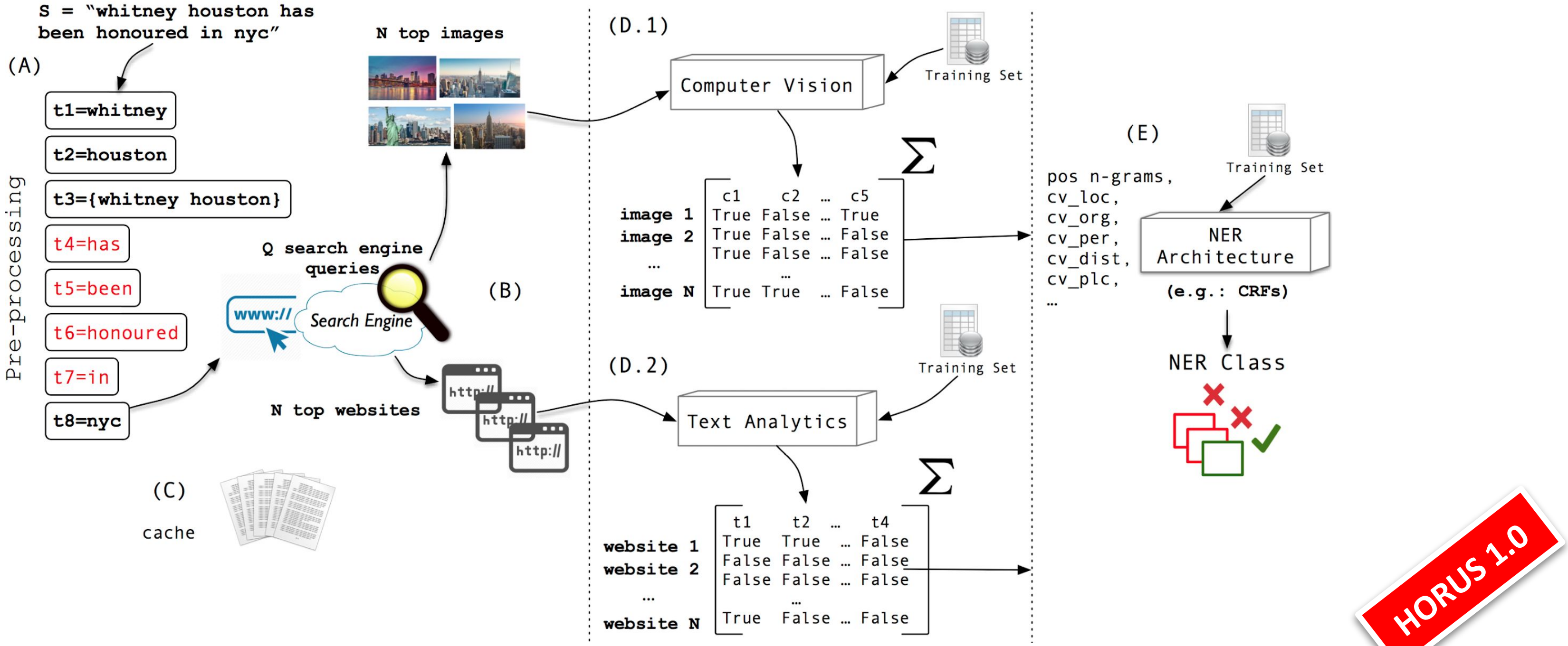
- Combination of text and image features with simple decision trees-based models

- Majority voting committee

Named Entity Recognition for Noisy Data

RQ1

Can images along with news improve
the performance of the named entity recognition models on noisy text?



Named Entity Recognition for Noisy Data

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

RQ1	Can images along with news improve the performance of the named entity recognition models on noisy text?

Object detection

SIFT (Scale Invariant Feature Transform): image descriptor extraction

BoF: clustering of feature histograms (k-means)

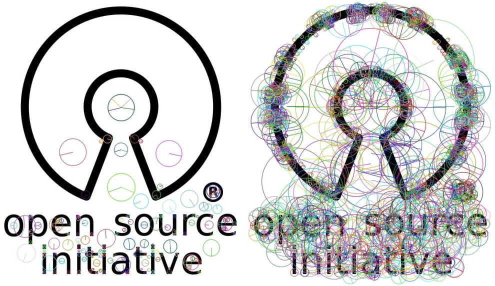
- Image ~ histogram of visual words frequencies
- Some image groups are related to certain named entities

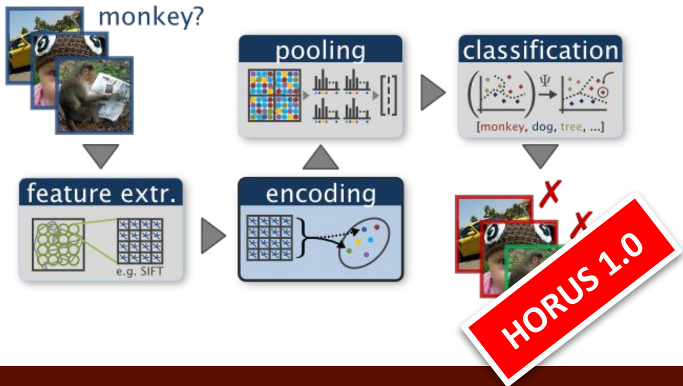
Classifiers: Unsupervised + Supervised learning

Training datasets

- LOC: *Scene 13*
- PER: *Caltech 101 Object Categories*
- ORG: *METU*

NER Images Candidates (number of trained models)	
LOC	Building, Suburb, Street, City, Country, Mountain, Highway, Forest, Coast and Map (10)
ORG	Company Logo (1)
PER	Human Face (1)





Named Entity Recognition for Noisy Data

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

RQ1	Can images along with news improve the performance of the named entity recognition models on noisy text?

Text Analytics

Features: term frequency-Inverse document frequency (TF-IDF)
Classifier: bag-of-words based
Training dataset: 15K DBpedia instances annotated with PER, ORG and LOC classes

```
SELECT ?location, ?abstract FROM <http://dbpedia.org> WHERE {?location
rdf:type dbo:Location .
       ?location dbo:abstract ?abstract .
FILTER (lang(?abstract) = ‘en’)} LIMIT 50000
```

DBpedia

HORUS 1.0

Named Entity Recognition for Noisy Data

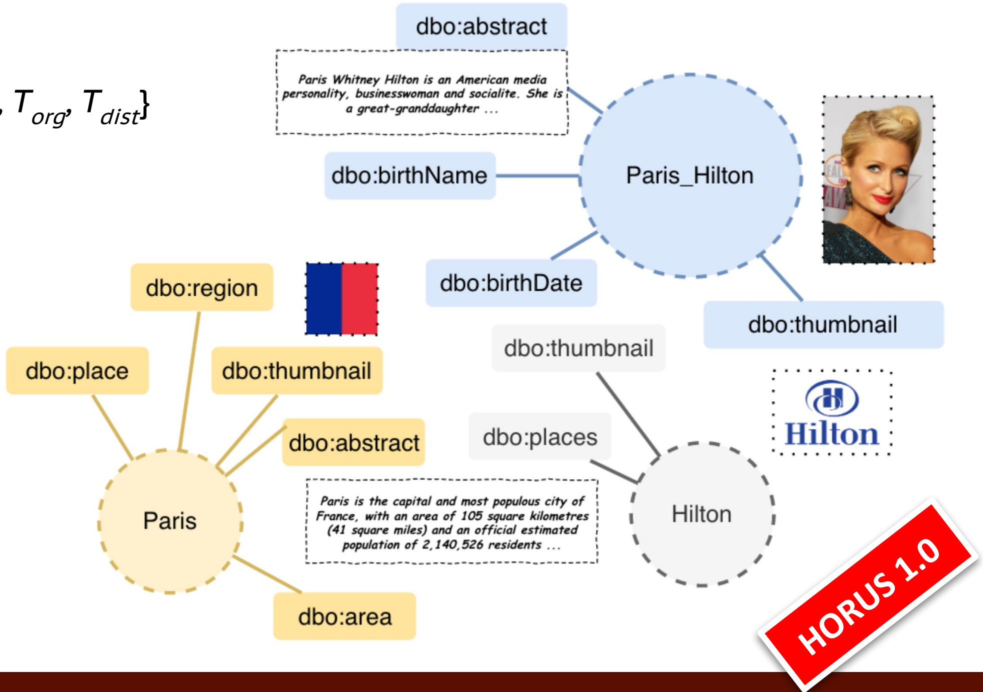
RQ1 Can images along with news improve the performance of the named entity recognition models on noisy text?

Heuristic-based DT

$$M_i = \{j, t, ng_{pos}, C_{loc}, C_{per}, C_{org}, C_{dist}, C_{plc}, T_{loc}, T_{per}, T_{org}, T_{dist}\}$$

for each sentence *i* and token *t* in position *j*

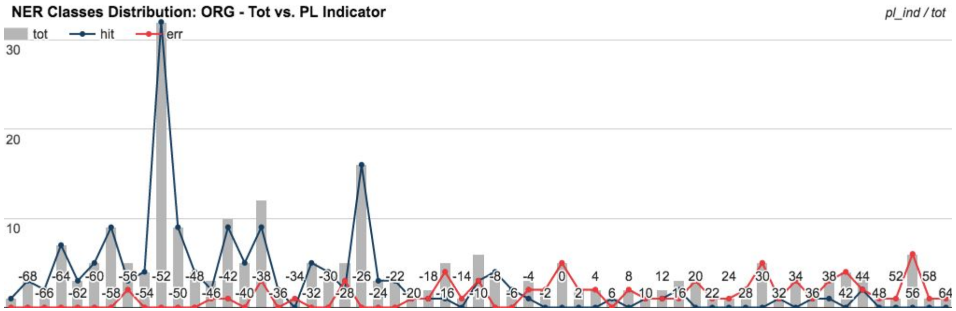
- ng_{pos} = n-gram of POS tag
- C_k, T_k = total objects found by classifier for class *k*
- C_{dist}, T_{dist} = distance b/w two top predictions
- C_{plc} = sum of all predictions by all LOC classifiers

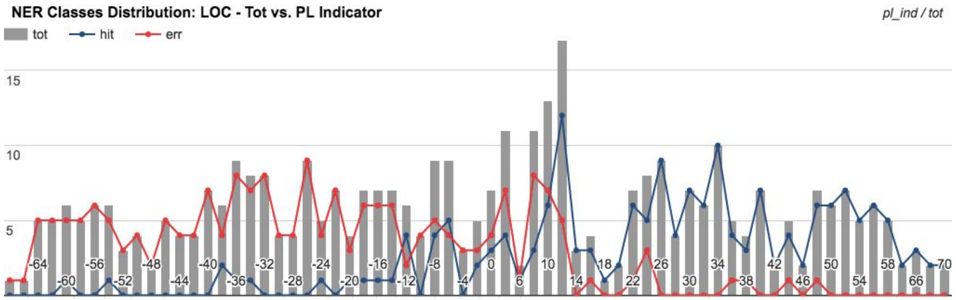


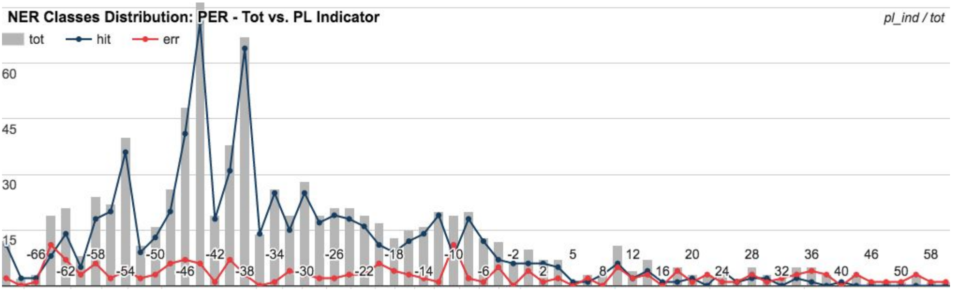
Named Entity Recognition for Noisy Data

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

RQ1 — Can images along with news improve the performance of the named entity recognition models on noisy text?







$$\Phi_l^{\mathrm{LOC}}(i) = \begin{cases} 1, & \text{if } \sum_{l=1}^{L} \Phi_l^{\mathrm{LOC}}(i) \geq \theta \\ -1, & \text{otherwise} \end{cases}$$

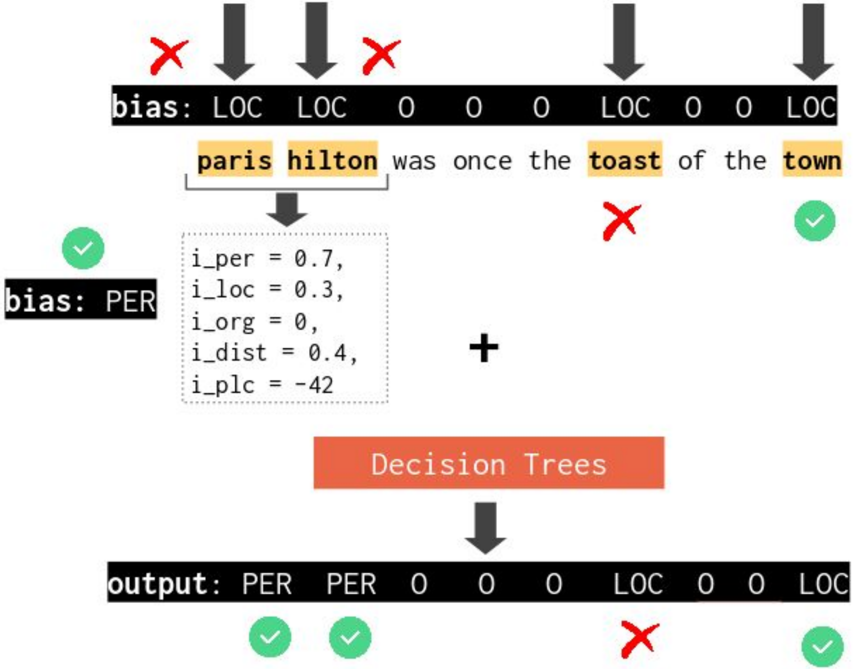
$$\text{if } c \neq \mathrm{LOC} \text{ we get } R^c_{\mathcal{I}_t} = \sum_{i \in \mathcal{I}_t} \sum_{l=1}^{L} \Phi_l^c(i)$$

HORUS 1.0

Named Entity Recognition for Noisy Data

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

RQ1 Can images along with news improve the performance of the named entity recognition models on noisy text?



NER Class	Precision	Recall	F-measure
Person (PER)	0.86	0.53	0.66
Location (LOC)	0.70	0.40	0.51
Organisation (ORG)	0.90	0.46	0.61
None	0.99	1.0	0.99
Average (PLO)	0.82	0.46	**0.59**

Table 2: Performance measure for our approach in Ritter dataset: 4-fold cross validation

NER System	Description	Precision	Recall	F-measure
Ritter et al., 2011 [19]	LabeledLDA-Freebase	0.73	0.49	0.59
Bontcheva et al., 2013 [3]	Gazetteer/JAPE	0.77	0.83	0.80
Bontcheva et al., 2013 [3]	Stanford-twitter	0.54	0.45	0.49
Etter et al., 2013 [6]	SVM-HMM	0.65	**0.49**	0.54
our approach	Cluster (images and texts) + DT	**0.82**	0.46	**0.59**

Table 3: Performance measures (PER, ORG and LOC classes) of state-of-the-art N for short texts (Ritter dataset). Approaches which do not rely on hand-crafted r *Gazetteers* are highlighted in gray. Etter et al., 2013 trained using 10 classe

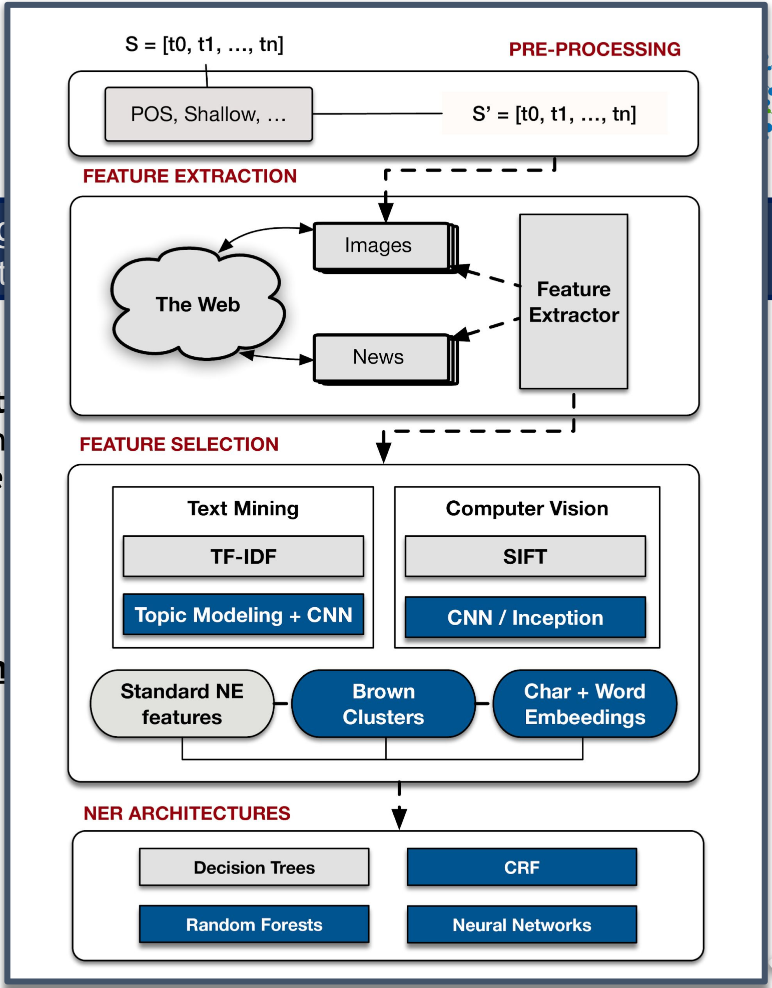
HORUS 1.0

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

Named Entity Recognition for

RQ1

Can images along
the performance of the named ent



HORUS 2.0

Advantages

- CV module makes the approach **language agnost**
- Each text snipped is **automatically translated** (en
- Very **simple algorithms (DT) performing really we** SOTA)
- **NO Gazetteers**!

Disadvantages

- Still **<u>NOT</u> achieving similar to SOTA in <u>formal dom</u>**
- Do **<u>NOT</u> scale well**!

Named Entity Recognition for Noisy Data

RQ1

Can images along with news improve
the performance of the named entity recognition models on noisy text?

Brown Clusters ($\mathcal{B}$)

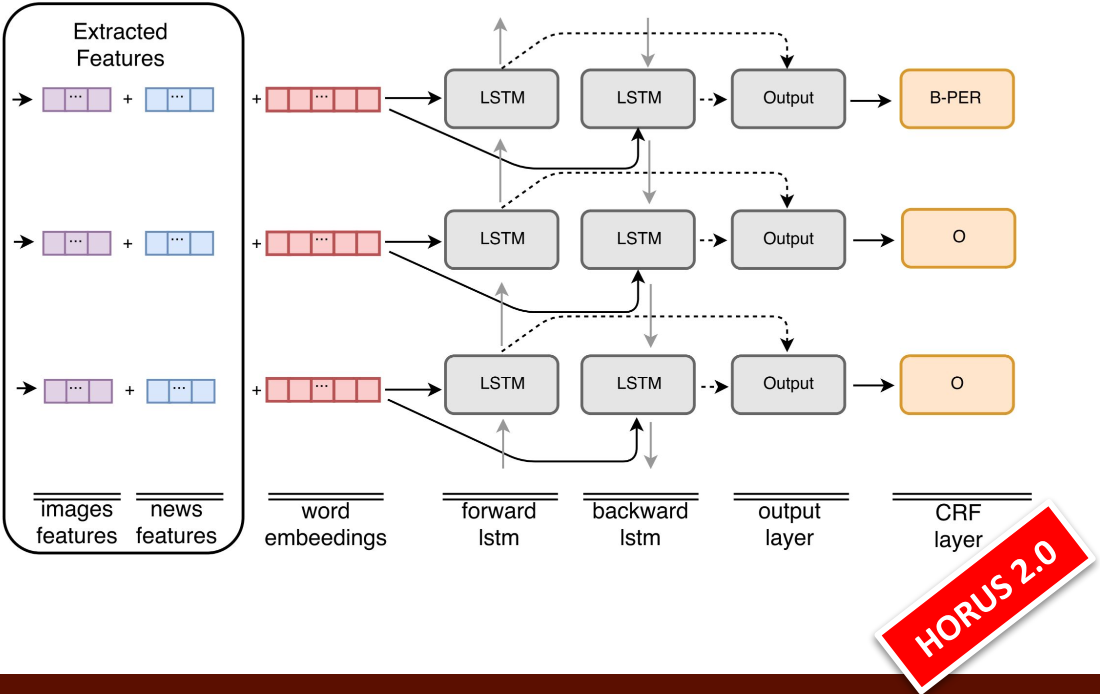
Standard Features: ($\mathcal{S}$)

Topic Modeling + CNNs ($\mathcal{TX}_{nn}$)

Seeds x Word2Vec ($\mathcal{TX}_{emb}$)

Text Correlation ($\mathcal{TX}_{stats}$)

Convolutional Neural Nets ($\mathcal{CV}_{nn}$)

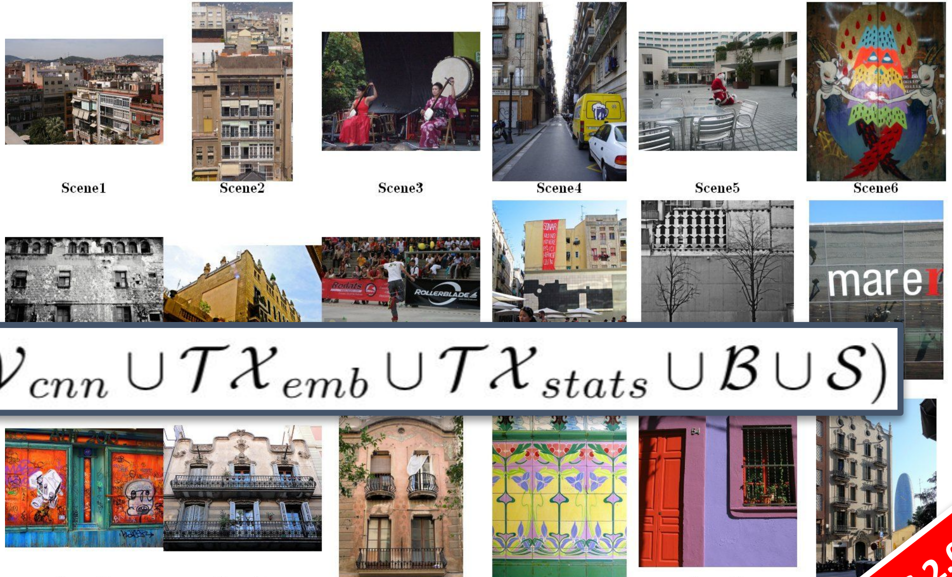


Named Entity Recognition for Noisy Data

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

RQ1	Can images along with news improve the performance of the named entity recognition models on noisy text?

- DT + HORUS
- CRF + HORUS
- B-LSTM + CRF + HORUS
- B-LSTM + CNN + CRF + HORUS
- Char + B-LSTM + CRF + HORUS



$$\mathcal{F} = (\mathcal{TX} \cup \mathcal{CV} \cup \mathcal{TX}_{cnn} \cup \mathcal{CV}_{cnn} \cup \mathcal{TX}_{emb} \cup \mathcal{TX}_{stats} \cup \mathcal{B} \cup \mathcal{S})$$

HORUS 2.0

Named Entity Recognition for Noisy Data

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

RQ1 Can images along with news improve the performance of the named entity recognition models on noisy text?

Cfg[4]	Features
cfg01	$\mathcal{S}$
cfg02	$\mathcal{S} + \mathcal{TX}$ (TF-IDF+SVM)
cfg03	$\mathcal{S} + \mathcal{CV}$ (SIFT+K-means+SVM)
cfg04	$\mathcal{S} + \mathcal{TX} + \mathcal{CV}$ [15]
cfg05	$\mathcal{S}$ + Lemma
cfg06	$\mathcal{S}$ + Lemma + $\mathcal{TX}$
cfg07	$\mathcal{S}$ + Lemma + $\mathcal{CV}$
cfg08	$\mathcal{S}$ + Lemma + $\mathcal{TX} + \mathcal{CV}$

HORUS 1.1

Cfg	Features
cfg09	$\mathcal{S}$ + Brown 64M c320
cfg10	$\mathcal{S}$ + Brown 64M c640 ($\mathcal{B}_{best}$)
cfg11	$\mathcal{S}$ + Brown 500M c1000
cfg12	$\mathcal{S}$ + Lemma + Brown 64M c320
cfg13	$\mathcal{S}$ + Lemma + Brown 64M c640
cfg14	$\mathcal{S}$ + Lemma + Brown 500M c1000
cfg15	$\mathcal{S} + \mathcal{B}_{best} + \mathcal{CV}$
cfg16	$\mathcal{S} + \mathcal{B}_{best} + \mathcal{TX}$
cfg17	$\mathcal{S} + \mathcal{B}_{best} + \mathcal{CV} + \mathcal{TX}$

Brown Clusters

Cfg	Features	Cfg	Features
cfg18	$\mathcal{S}+\mathcal{CV}_{cnn}$	cfg30	$=18+\mathcal{B}_{best}$
cfg19	$\mathcal{S}+\mathcal{TX}_{cnn}$	cfg31	$=19+\mathcal{B}_{best}$
cfg20	$\mathcal{S}+\mathcal{TX}_{emb}$	cfg32	$=20+\mathcal{B}_{best}$
cfg21	$\mathcal{S}+\mathcal{TX}_{stats}$	cfg33	$=21+\mathcal{B}_{best}$
cfg22	$\mathcal{S}+\mathcal{TX}_{cnn}+\mathcal{TX}$	cfg34	$=22+\mathcal{B}_{best}$
cfg23	$\mathcal{S}+\mathcal{TX}_{cnn}+\mathcal{TX}+\mathcal{TX}_{e}+\mathcal{TX}_{stats}$	cfg35	$=23+\mathcal{B}_{best}$
cfg24	$\mathcal{S}+\mathcal{TX}_{cnn}+\mathcal{CV}_{cnn}$	cfg36	$=24+\mathcal{B}_{best}$
cfg25	$\mathcal{S}+\mathcal{TX}_{cnn}+\mathcal{TX}+\mathcal{CV}$	cfg37	$=25+\mathcal{B}_{best}$
cfg26	$\mathcal{S}+\mathcal{CV}_{cnn}+\mathcal{CV}$	cfg38	$=26+\mathcal{B}_{best}$
cfg27	$\mathcal{S}+\mathcal{CV}_{cnn}+\mathcal{CV}+\mathcal{TX}$	cfg39	$=27+\mathcal{B}_{best}$
cfg28	$\mathcal{S}+\mathcal{CV}_{cnn}+\mathcal{CV}+\mathcal{TX}_{cnn}+\mathcal{TX}$	cfg40	$=28+\mathcal{B}_{be}$ [illegible]
cfg29	$\mathcal{S}+\mathcal{CV}_{cnn}+\mathcal{CV}+\mathcal{TX}_{cnn}+\mathcal{TX}+\mathcal{TX}_{emb}+\mathcal{TX}_{stats}$	cfg41	$=2$ [illegible]

Deep Learning

HORUS 2.0

Named Entity Recogniti… or Noisy Data

RQ1

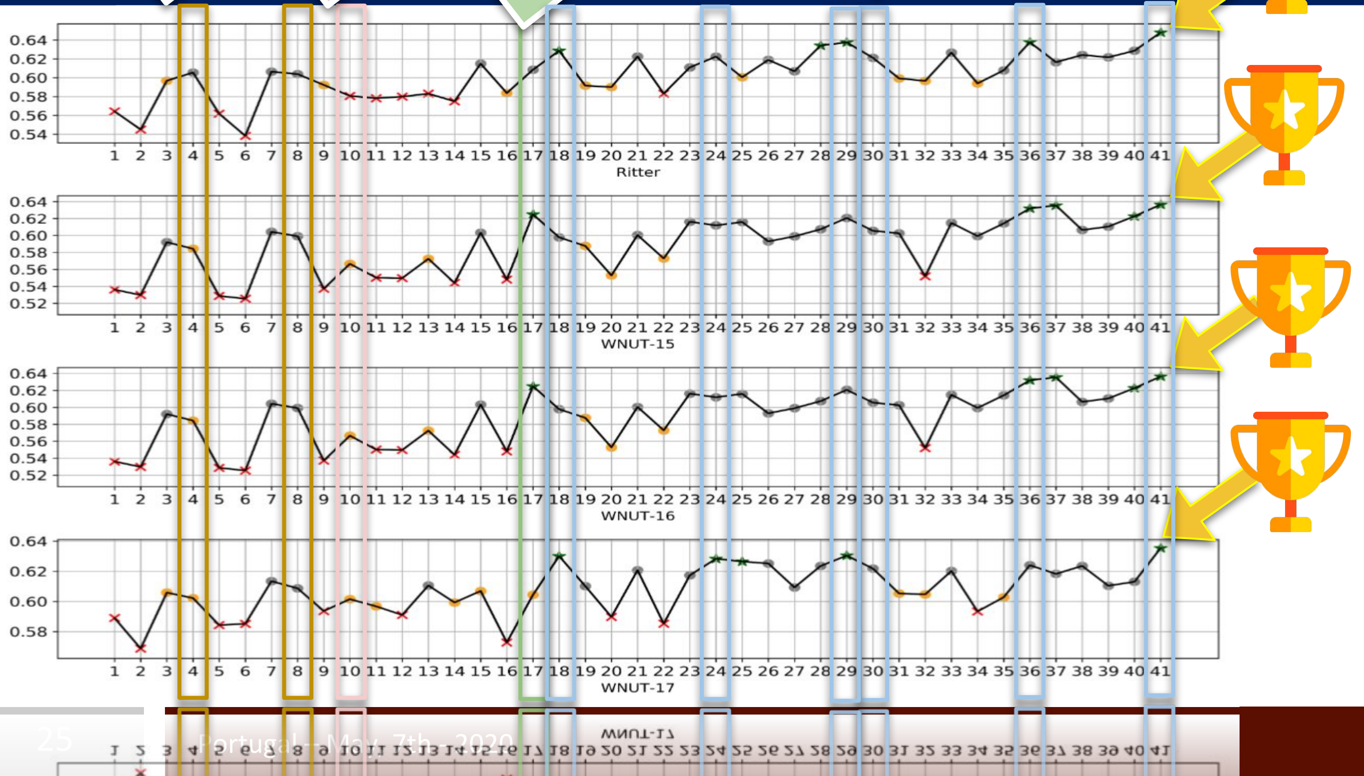
…an images al… …th news improve
…he perf… …ce of the named e… …ity recognition mode… o… noisy text?

HORUS 1.1

Brown C.

Brown C.+HORUS 1.1

HORUS 2.0



Below the baseline

HORUS 1.1

Above the baseline

TOP

HORUS 2.0

Named Entity Recognition for Nois... ata

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

10 = Baseline

04 = HORUS 1.0

41 = HORUS 2.0

RQ1

Can images al...
the performance of the named... ty re... tio... ls on noisy text?

Dataset		Decision Trees			Random Forest			CRF			B-LSTM CRF [20]			B-LSTM C+CRF [23]			B-LSTM C+CRF+CNN [30]		
cfg ⟶		10	04	41	10	04	41	10	04	41	10	04	41	10	04	41	10	04	41
Ritter	P	0.48	+2%	+4%	0.51	+1%	+24%	0.73	+5%	+7%	0.77	+1%	−3%	0.81	−5%	−1%	0.81	−5%	−5%
	R	0.49	+1%	+3%	0.48	−1%	−2%	0.58	−8%	−2%	0.63	+5%	+5%	0.59	+5%	+4%	0.62	+3%	+5%
	F	0.49	+1%	+3%	0.49	+4%	+7%	0.58	+2%	+7%	0.68	+1%	+1%	0.67	+1%	+1%	0.69	−1%	+1%
WNUT-15	P	0.49	+2%	+5%	0.52	+7%	+25%	0.72	+7%	+9%	0.72	−4%	−2%	0.77	−3%	−4%	0.78	−4%	−5%
	R	0.50	+0%	+5%	0.49	+0%	+1%	0.48	−1%	+6%	0.69	+1%	+1%	0.65	+2%	+2%	0.66	+2%	+2%
	F	0.50	+0%	+5%	0.50	+5%	+9%	0.56	+2%	+8%	0.68	+0%	+0%	0.69	+0%	−1%	0.71	−1%	−2%
WNUT-16	P	0.49	+1%	+6%	0.52	+14%	+23%	0.72	+7%	+9%	0.72	−4%	−2%	0.77	−3%	−3%	0.78	−4%	−6%
	R	0.50	+1%	+6%	0.48	+0%	+2%	0.48	−1%	+6%	0.69	+0%	+1%	0.65	+2%	+2%	0.66	+2%	+2%
	F	0.49	+1%	+6%	0.50	+5%	+10%	0.56	+2%	+8%	0.69	−1%	+0%	0.69	+0%	+0%	0.71	−1%	−2%
WNUT-17	P	0.44	+3%	+7%	0.47	+13%	+24%	0.76	+2%	+1%	0.76	−2%	−2%	0.76	+0%	−2%	0.77	−3%	−3%
	R	0.45	+4%	+6%	0.44	+3%	+4%	0.50	+0%	+5%	0.63	+1%	+1%	0.64	+0%	+1%	0.62	+1%	+1%
	F	0.44	+4%	+6%	0.45	+6%	+12%	0.60	+0%	+4%	0.67	+0%	+0%	0.69	+0%	−1%	0.67	+0%	−1%

HORUS 2.0

Named Entity Recognition for Noisy Data

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

RQ1	Can images along with news improve the performance of the named entity recognition models on noisy text?

Contributions made:

- Novel NER Architecture based on Images and News (**NO Gazetteer!**)
- **Language Agnostic** NER Framework for Noisy Data (English and Portuguese)
- **Improved Recall** for NNs, <u>but at cost of precision</u>
- **Great improvement for CRF-based** models; <u>results comparable to SOTA NNs</u>

Named Entity Recognition for Noisy Data

BONUS =)

RQ1

Can images along with news improve
the performance of the named entity recognition mod noisy te

04 = HORUS 1.0

41 = HORUS 2.0

But, what if more data is available...?

	+cfg10	+cfg04	+cfg41
B-LSTM+CRF	0.5217	0.5352 ↑	0.5376 ↑

Table 5: B-LSTM+CRF F1-measure with expanded training/dev/test data over different feature sets.

HORUS 3.0

Trustworthiness

RQ2 How to compute a credibility score for a given information source?

How credible a given website is?

"*A credible web page is one whose information one can accept as the truth without needing to look elsewhere*". [Olteanu et al., 2013; Waweret al., 2014]

RQ2 Contribution

An algorithm to calculate the Likert Score for a given source

Contribution 2

Alternatives

1. PageRank (shutdown) and Alexa (paid)
2. Existing data is too small/do not scale! Manual annotation is costly [Haas and Unkel, 2017].
3. Theoretical research or confidential data (e.g., mouse movement, time spent, etc..) in a restricted simulation environment (e.g., Google and Microsoft) [Liu et al. (2015)]
4. Open-source work = *f(q, S)* [Nakamura et al. 2007] or likert [Olteanu et al. (2013), Wawer et al. (2014)]

Trustworthiness

RQ2 How to compute a credibility score for a given information source?

Credibility: Likert Scale (experiment configurations)					
5-class		**3-class**		**2-class**	
1	very non-credible	**1**	low	**1**	low
2	non-credible	**2**	neutral	**2**	high
3	neutral	**3**	high		
4	credible				
5	very credible				

Features	**Type of features (e.g.)**	**Advantages**	**Drawbacks**
Content-based	Textual, Appearance and Meta-information	Mostly textual features, which are easy to extract	Experiments show that they are not effective enough to generalize
Social-based	Social and General Popularity, Page Rank and Alexa	Mostly based on (private) user-content information, which is - in its essence - more reliable	Data is not freely available to the community

Trustworthiness

RQ2 | How to compute a credibility score for a given information source?

- **Content-based** (25)
  - Text (20)
  - Appearance (4)
  - Meta-information (1)
- **Social-based** (12):
  - Social Popularity (9)
  - General Popularity (1)
  - Link structure (2)

Research is still very contradictory!

NO!

Appearance is not important! Most significant are **Social-based (12) and some of Text (20)!** [Olteanu et al., 2013, Dong et al., 2015]

YES!

Appearance is very important [Fogg et al.,2003; Shah et al., 2015; Haas and Unkel, 2017].

Trustworthiness

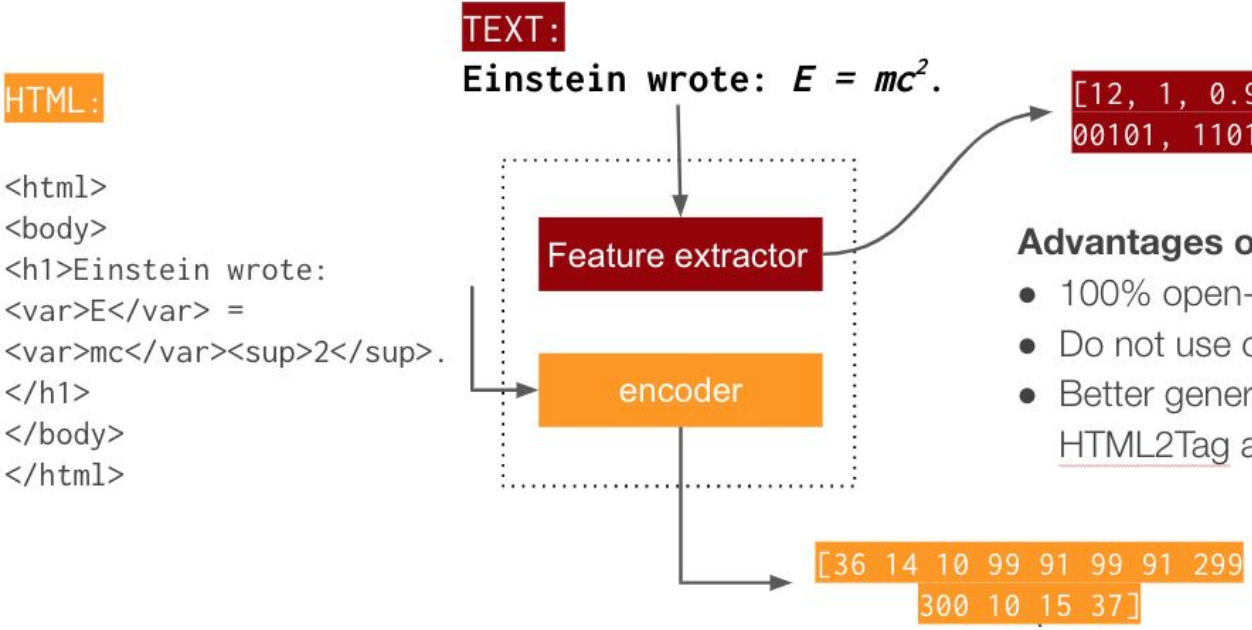
SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

RQ2	How to compute a credibility score for a given information source?

- Similarly to the concept "Bag-of-Words" we introduce a concept we named "**Bag-of-Tags**".
- We expect to capture not only **visual features**, but also **hidden patterns** in the source code.
- We **also explore different lexical** features (e.g., Vader Lexicon)

Trustworthiness

RQ2 How to compute a credibility score for a given information source?



Advantages of the framework:

- 100% open-source
- Do not use commercial data
- Better generalization level due to the HTML2Tag approach

Trustworthiness

RQ2 How to compute a credibility score for a given information source?

- **Microsoft Dataset** [Schwarz and Morris, 2011]
  - aprx. 1000 URLs
- **Content Credibility Corpus (C3)** [Kakol et al., 2017]
  - 15.750 evaluations of 5.543 URLs from
  - 2.041 participants

Trustworthiness

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

RQ2 — How to compute a credibility score for a given information source?

$$f_{arc}(w) = \left(\left[\frac{1}{\log(\Delta_b \times \Delta_e)} + \log(\Delta_a) + \frac{1}{\Delta_u}\right]\right) \times \gamma$$

$$w_b\colon \bigcup_{i=1}^{R} \varphi(i, w_b)$$

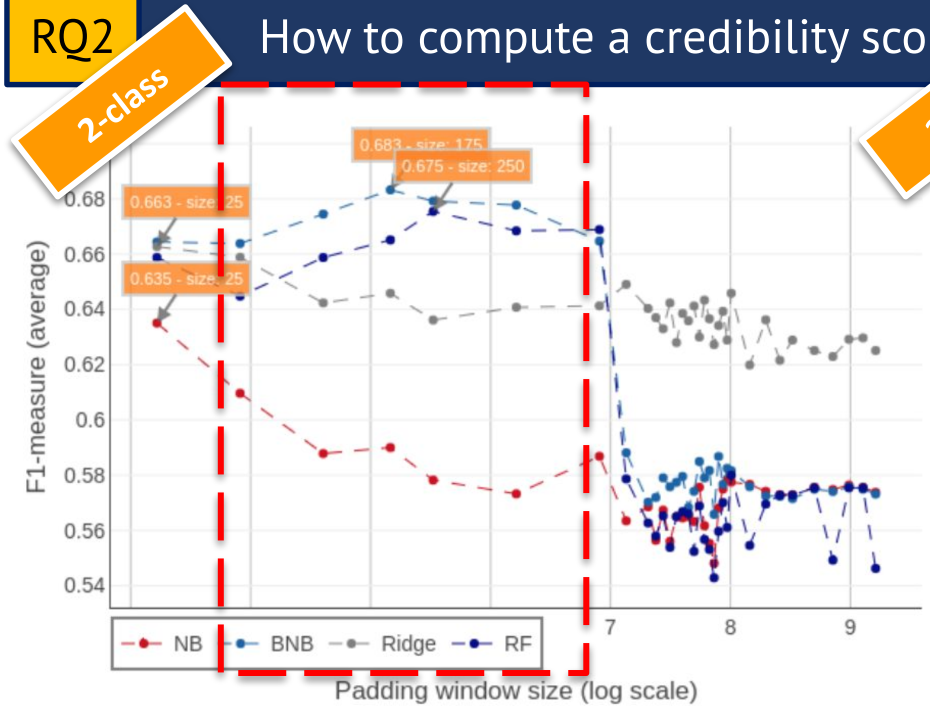
Features

1. Web Archive "Freshness"
2. Domain (enc)
3. Authority (enc)
4. Outbound Links $\sum_{n=1}^{P} \phi(w_c)$
5. Text Category $\sum_{s=1}^{w_s} \gamma(s) \frown \gamma(w_t)$
6. (5) - LexRank
7. (5) - Latent Semantic A.
8. Readability Metrics [Si and Callam, 2001]
9. SPAM $\psi(w_b) \frown \psi(w_t)$
10. Social Tags
11. OpenSources
12. PageRank CommonCrawl
13. General Inquirer
14. Vader Lexicon
15. **HTML2Seq (BoT)**

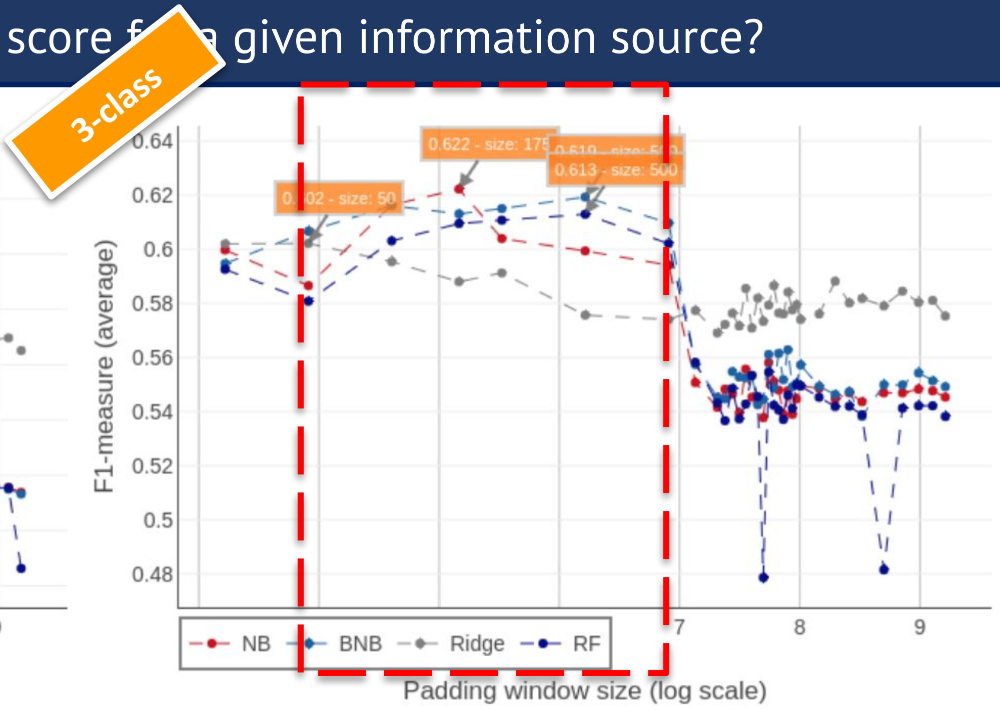
$$x = \begin{cases} 1, & \text{if } w \in \mathcal{O} \\ 0, & \text{if } w \notin \mathcal{O} \end{cases}$$

Trustworthiness

RQ2 How to compute a credibility score for a given information source?



(c) HTML2Seq (F1): C3 Corpus 2-classes



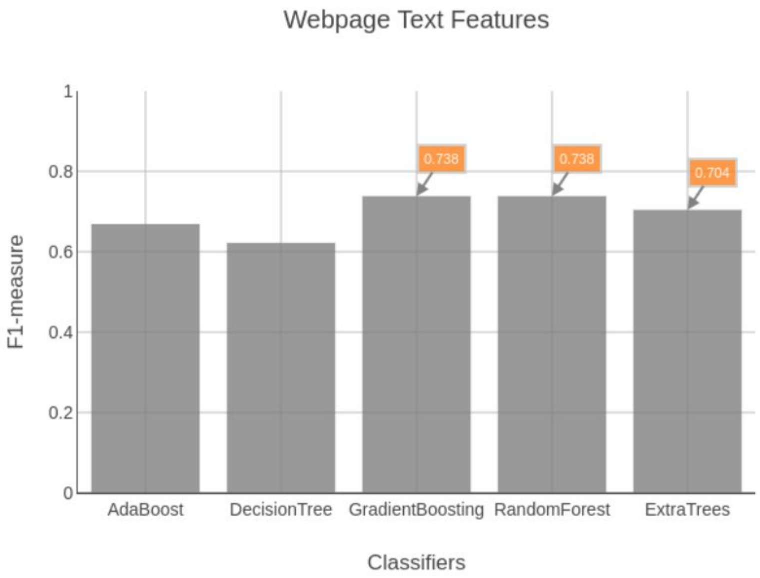
(d) HTML2Seq (F1): C3 Corpus 3-classes

Trustworthiness

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

RQ2 How to compute a credibility score for a given information sou

TOP 1



(a) Textual Features.

(b) Textual+HTML2Seq (best padding) Features.

Trustworthiness

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

RQ2 How to compute a credibility score for a given information source?

2-class — F1 SOTA = 0.745

Microsoft Dataset (Gradient Boosting, $K = 25$)			
Class	Precision	Recall	F1
low	0.851	0.588	0.695
high	0.752	0.924	0.829
weighted	0.794	0.781	0.772
micro	0.781	0.781	0.781
macro	0.801	0.756	0.762
C3 Corpus (AdaBoost, $K = 75$)			
Class	Precision	Recall	F1
low	0.558	0.355	0.434
high	0.732	0.862	0.792
weighted	0.675	0.695	0.674
micro	0.695	0.695	0.695
macro	0.645	0.609	0.613

Table 1: Text+HTML2Seq features (2-class): best classifier performance

3-class

...rosoft Dataset (... Boosting, $K = 75$)			
Class	Precision	Recall	F1
low	0.567	0.447	0.500
medium	0.467	0.237	0.315
high	0.714	0.916	0.803
weighted	0.626	0.662	0.626
micro	0.662	0.662	0.662
macro	0.583	0.534	0.539
C3 Corpus (AdaBoost, $K = 100$)			
Class	Precision	Recall	F1
low	0.143	0.031	0.051
medium	0.410	0.177	0.247
high	0.701	0.916	0.794
weighted	0.583	0.660	0.598
micro	0.660	0.660	0.660
macro	0.418	0.375	0.364

Table 2: Text+HTML2Seq features (3-class): best classifier performance

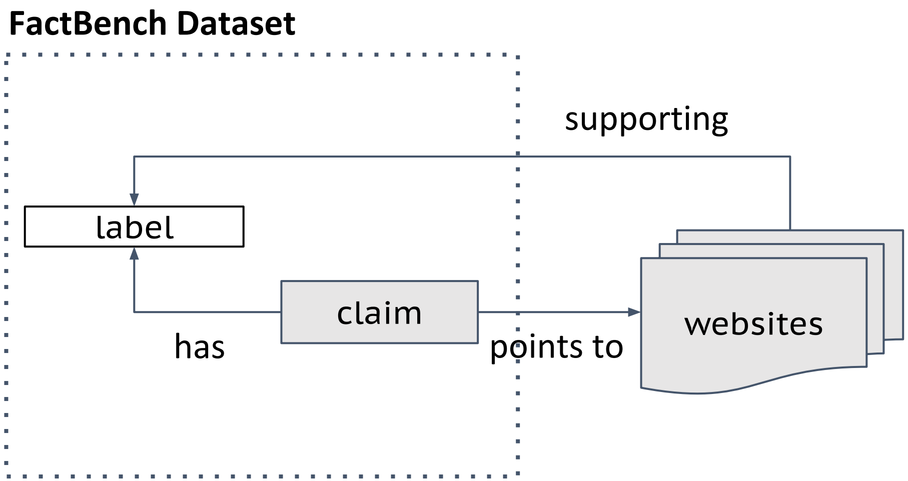
5-class — F1 SOTA = 0.763

Microsoft Dataset					
model	K	R^2	RMSE	MAE	
SVR	3	0.232	0.861	0.691	.238
Ridge	3	0.268	0.841	0.683	0.269
C3 Corpus					
model	K	R^2	RMSE	MAE	EVar
SVR	25	0.096	0.939	0.739	0.102
Ridge	25	0.133	0.920	0.750	0.134

Table 3: Text+HTML2Seq: regression measures (5-class). Selecting top K lexical features

Trustworthiness

RQ2 How to compute a credibility score for a given information source?

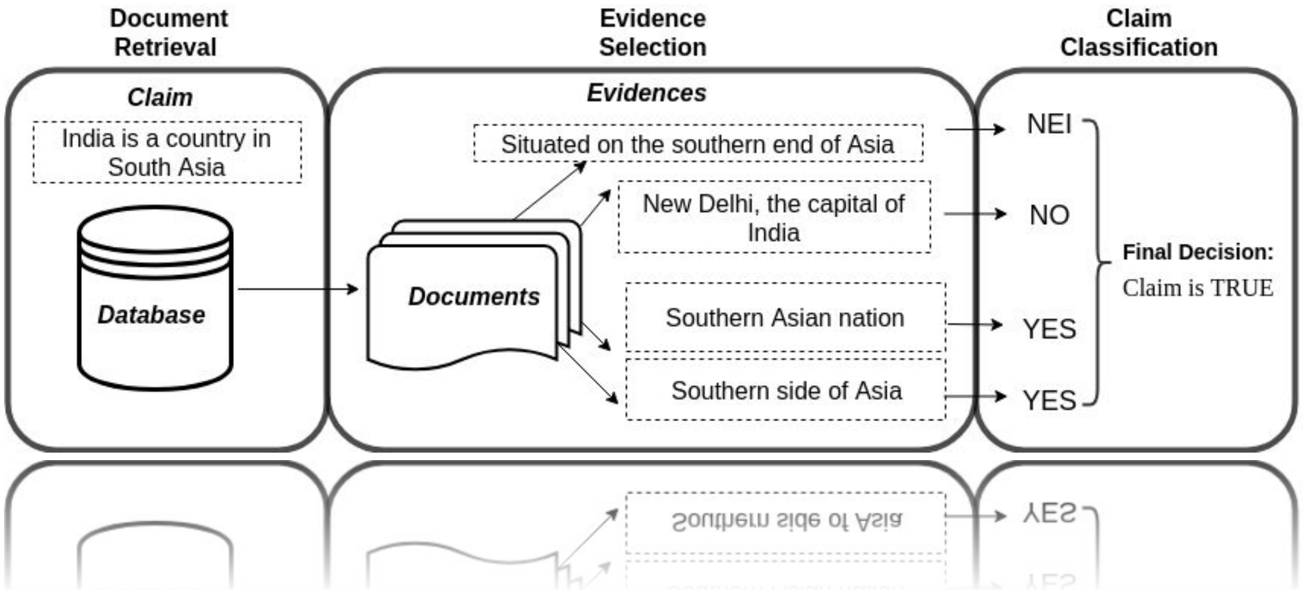


FactBench (Sample - Human Annotation)				
label	**claims**	**sites**	**non-cred**	**cred**
true	5	96	57	39
false	5	80	48	32
-	10	186	105	71
FactBench (Sample - Credibility Model)				
label	**non-cred**	**%**	**cred**	**%**
true	40	0.81	31	0.79
false	34	0.70	24	0.75

Table 5: FactBench Dataset: analyzing the performance of the credibility model in the fact-checking task.

Claim Verification

RQ3 How to determine the veracity of a given claim?



RQ3 Contribution

A Fact-checking Framework

Contribution 3

Claim Verification

RQ3 — How to determine the veracity of a given claim?

Types of claims

- Structured Claims: [dbr:Diego_Esteves; dbo:birthPlace; dbr:Brazil]
- Unstructured Claims: "*Diego is Brazilian.*"

Complexity

- Simple (1 sentence)
- Complex (1+ sentences)

Tasks

- Verification (true)
- Ranking (1+ claims)
- Plausibility (true)

Claim: Roman Atwood is a content creator. (**Supported**) **Evidence:** `[wiki/Roman_Atwood]` He is best known for his vlogs, where he posts updates about his life on a daily basis.
Claim: Furia is adapted from a short story by Anna Politkovskaya. (**Refuted**) **Evidence:** `[wiki/Furia_(film)]` Furia is a 1999 French romantic drama film directed by Alexandre Aja, ..., adapted from the science fiction short story Graffiti by Julio Cortázar.
Claim: Afghanistan is the source of the Kushan dynasty. (**NotEnoughInfo**)

Fig. 1. Three examples from the FEVER dataset [14].

[A. Soleimani et al. 2019]

Claim Verification

RQ3	How to determine the veracity of a given claim?

- Diego’s birthplace is Brazil
- Diego was born in Brazil
- Diego was born in Rio de Janeiro
- Diego is Brazilian

1. **Hard-coded** verbalisation and rules Represents facts about the world (scalability issues)
2. **Supervised** models (not optimal accuracy)
3. Use **distant supervision** methods (sub-optimal precision)
4. **External linguistic** corpora (e.g. lexical databases) to obtain synonym (**not good recall**)

Claim Verification

RQ3	How to determine the veracity of a given claim?

$\gamma(s, p, o, \mathcal{L}) = [\phi(s, l_1) \times \Gamma(p, l_1) \times \phi(o, l_1)] \cup [\phi(s, l_2) \times \Gamma(p, l_2) \times \phi(o, l_2)], \ldots, \cup[\phi(s, l_n) \times \Gamma(p, l_n) \times \phi(o, l_n)]$, where

(a) $\phi(x, l_i)$ returns a set of m labels $(x_1, x_2, \ldots, x_m)$ that are similar to the label of the resource x ($s \in \mathcal{S}$ and $o \in \mathcal{O}$), which is extracted from the `rdfs:labels` predicate for a given language $l_i \in \mathcal{L}$.

(b) $\Gamma(p, l_i)$ returns a set of verbalized patterns $\mathcal{P}$ for a given predicate p and a language $l_i \in \mathcal{L}$.

Feature	Definition
is sub	Checks if the document contains subject
is obj	Checks if the document contains object
is pred	Checks if the document contains predicate
dist sub obj Text follows	Distance between subject and object
pred between	Does predicate occur between subject and object
sub relax	Checks whether subject is present in partial form
obj relax	Checks whether object is present in partial form
pred relax	Checks whether predicate is present in partial form
Jaccard distance	Maximum Jaccard coefficient
Cosine similarity	Maximum cosine similarity
Semantic similarity	Similarity score of most semantically similar sentence

Claim Verification

RQ3 How to determine the veracity of a given claim?

Domain

	C	P	R	F_1	AUC	RMSE
J48	89.7%	0.898	0.897	0.897	0.904	0.295
SimpleLogistic	89.0%	0.890	0.890	0.890	0.949	0.298
NaiveBayes	81.2%	0.837	0.812	0.808	0.930	0.415
SMO	85.4%	0.861	0.854	0.853	0.854	0.382

Range

	C	P	R	F_1	AUC	RMSE
J48	90.9%	0.909	0.909	0.909	0.954	0.271
SimpleLogistic	88.0%	0.880	0.880	0.880	0.946	0.301
NaiveBayes	83.3%	0.852	0.833	0.830	0.933	0.387
SMO	83.3%	[illegible]52	0.833	0.830	0.833	0.409

DomainRange

	C	P	R	F_1	AUC	RMSE
J48	91.0%	0.910	0.910	0.910	0.953	0.270
SimpleLogistic	88.9%	0.889	0.889	0.889	0.950	0.2[illegible]
NaiveBayes	84.5%	0.861	0.845	0.843	0.935	0.380
SMO	83.6%	0.853	0.836	0.834	0.836	0.405

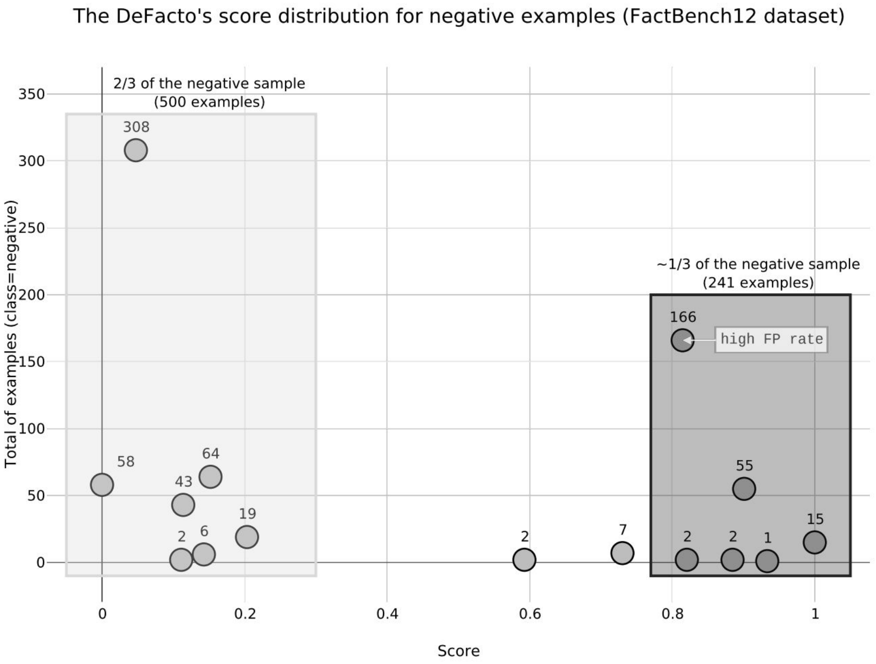
Property

	[illegible]	P	R	F_1	AUC	RMSE
J48	[illegible]0.8%	0.786	0.708	0.687	0.742	0.427
SimpleLogistic	64.9%	0.653	0.649	0.646	0.726	0.460
NaiveBayes	61.3%	0.620	0.613	0.608	0.698	0.488
SMO	64.6%	0.673	0.646	0.632	0.646	0.595

VERBALISATION!

Claim Verification

RQ3 How to determine the veracity of a given claim?



{'FAKE', 'REAL'}
6335
4244
2091
accuracy: 0.893
accuracy: 0.898
accuracy: 0.936
accuracy: 0.936

	agg_rank	count	label
said	**9.8**	**5**	**REAL**
friday	**2.66667**	**3**	**REAL**
monday	**3**	**3**	**REAL**
says	**8.33333**	**3**	**REAL**
gop	4	3	REAL
tuesday	**8.66667**	**3**	**REAL**
cruz	2.33333	3	REAL
conservative	6.66667	3	REAL
islamic	5.33333	3	REAL

	agg_rank	count	label
share	5.33333	3	FAKE
print	7.66667	3	FAKE
october	**2.66667**	**3**	**FAKE**
november	**5.33333**	**3**	**FAKE**
hillary	2	3	FAKE
article	4.33333	3	FAKE
2016	1.33333	3	FAKE

Basic AFC Architecture

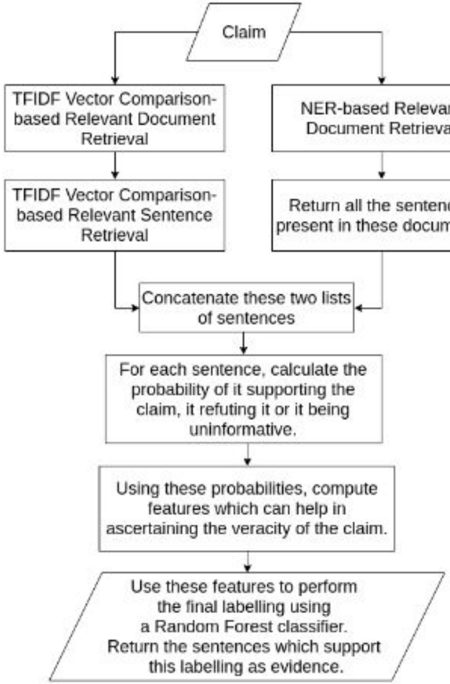


Figure 1: The main steps of our approach

DeFactoNLP 1.0

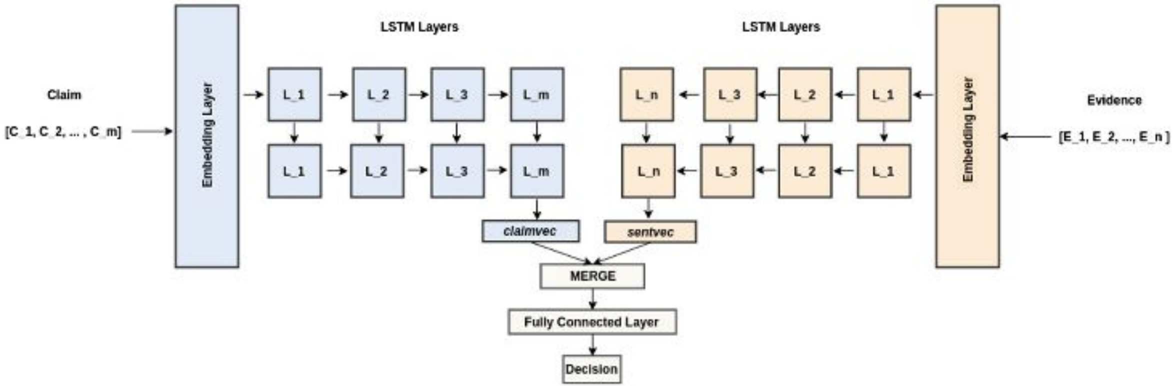


Fig. 2. SimpleLSTM model. The inputs are claim and evidence. Both, the evidence and the claim are fed to an embedding layer (common for both) that outputs embedding representation for each word. These embeddings are then passed through LSTM layers. The final output of LSTM, `sentvec` and `claimvec`, are merged and fed to the fully connected layer.

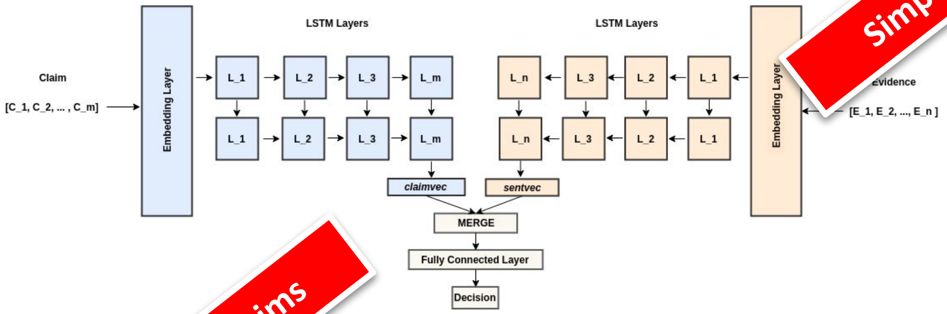
DeFactoNLP 2.0

Claim Verification

RQ3

How to determine the ver … of a given claim?

Simple Claims



Dataset	Classifier	Accuracy	Precision	Recall	f1 Score
FEVER Support	XGBoost [32]	0.766	0.766	0.766	0.762
	TE [9]	0.691	0.835	0.655	0.734
	XI-FEATURE RF	0.79	0.76	0.83	0.79
	XI-FEATURE SVM	0.79	0.71	0.85	0.77
	XI-FEATURE MLP	0.79	0.76	0.81	0.78
	SimpleLSTM	0.850	0.834	0.856	0.845
FEVER Reject	XGBoost [32]	0.74	0.738	0.736	0.73
	TE [9]	0.548	0.759	0.533	0.626
	XI-FEATURE RF	0.73	0.73	0.81	0.76
	XI-FEATURE SVM	0.642	0.73	0.78	0.75
	XI-FEATURE MLP	0.74	0.69	0.78	0.73
	SimpleLSTM	0.816	0.836	0.811	0.824
FEVER 3-class	XGBoost [32]	0.535	0.54	0.534	0.539
	TE [9]	0.418	0.372	0.622	0.465
	XI-FEATURE RF	0.55	0.60	0.61	0.60
	XI-FEATURE SVM	0.55	0.54	0.56	0.53
	XI-FEATURE MLP	0.59	0.61	0.62	0.61
	SimpleLSTM	0.635	0.643	0.620	0.642

Complex Claims

Metric	DeFactoNLP	Baseline
Label Accuracy	0.5136	0.4884
Evidence F1	0.4277	0.1826
FEVER Score	0.3833	0.2745

DeFactoNLP 3.0

- BERT + Re-ranking (similar to Soleimani et al.)

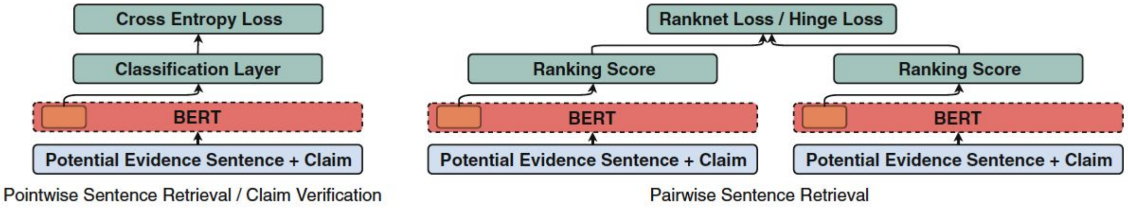


Fig. 2. Pointwise sentence retrieval and claim verification (left), Pairwise sentence retrieval (right). Orange boxes indicate the last hidden state of the [CLS] token. (Color figure online)

How to design AFC to debunk (real-life) fake news?

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

Claim Verification

Structured Claims

RDF Validation, Ranking and Plausibility
Esteves et al., 2018
Gerber et al., 2015

Unstructured Claims

Complex Claims Verification
Aniketh et al., 2018

Evidence Selection

Noisy Data

NER
Esteves et. al 2018
Peres et. al 2017
POS
Gui, Tao, et al., 2017

NLU

Argumentative Relations
Gil Rocha et al., 2018

Temporal Analysis

Temporal Scoping
Rula et al., 2019

Multilingual

Multilingual Evidence
Gerber et al., 2015

Semantics

LSTMs for Evidence Extraction
Piyush C. et al., 2015

Source Trustworthiness

Credibility

Graph-based
Nakamura et al., 2007

Likert Scale
Esteves et al., 2018

Bias

Text Network Analysis
Paranyushkin, D., 2019

Discourse Analysis
Fairclough, Norman, 1992

Bots

Troll detection
Fornacciari, Paolo, et al., 2018

Chatbot detection
Radziwill and Morgan, 2017

DeepFakes
Zakharov, Egor, et al., 2019

Source Retrieval

Document Retrieval

DrQA
Chen et al, 2017

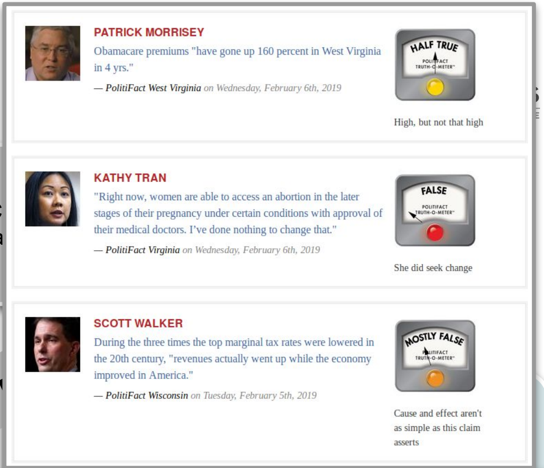
Cross-genre
Jurczk and Choi, 2018

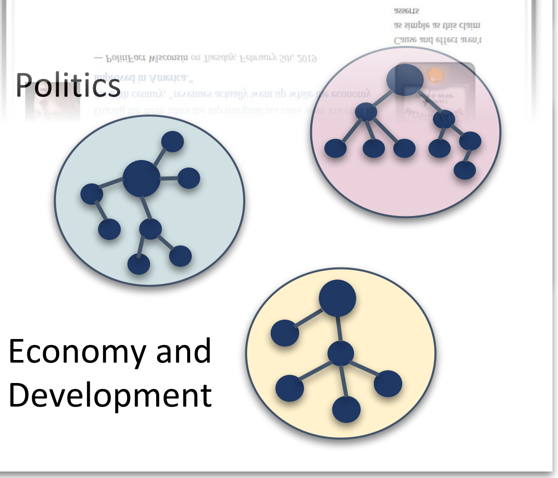
Data Fusion

KG Generation
Dong, Xin Luna, et al. 2014

Limitations and Future Work

Limitations	Future Work
Domain-specific and fine-grained AFCs	Ontology Engineering and KB population
Context-based information	Commonsense reasoning + Argumentation Bias and Bots detection
No Data, No Answer!	Data Acquisition and Fusion
Video and Photo manipulation	Deepfake AI





Projects

DeFacto: Deep Fact Validation

https://github.com/DeFacto

- University of Bonn
- FEUP

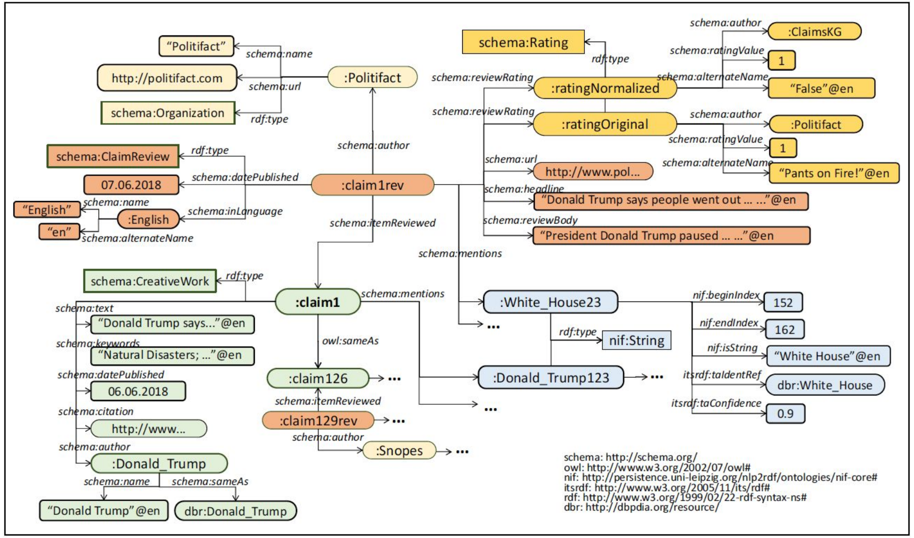


Fig. 3: Instantiation of the *Claims* model for a claim sourced from Politifact made by Donald Trump on June 6, 2018.

ClaimsKG: A KG of Fact-checked claims

https://data.gesis.org/claimskg/

Thanks

diegoesteves@gmail.com